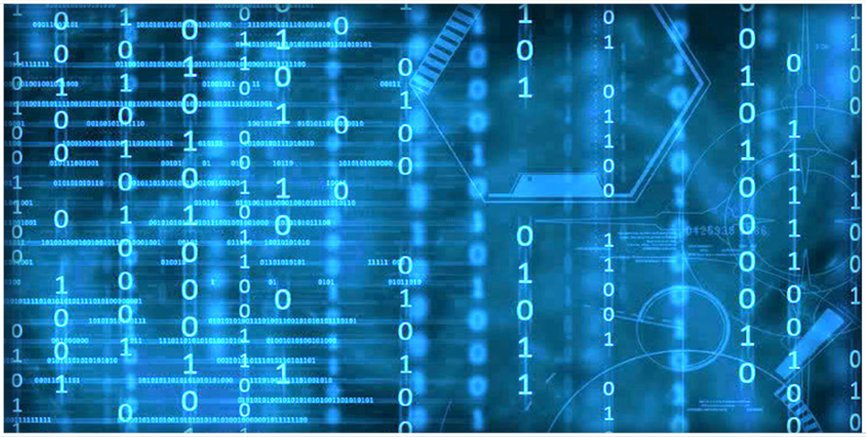
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 11 Issue 2

February 2020



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

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 11 Issue 2 February 2020
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Editor-in-Chief

Associate Editors

CONTENTS

A Review of Asset-Centric Threat Modelling Approaches

Livinus Obiora Nweke[1]
Information Security and Communication Technology
Norwegian University of Science and Technology (NTNU)
Gjøvik, Norway

Stephen D. Wolthusen[2]
School of Mathematics and Information Security
Royal Holloway, University of London
Egham, United Kingdom
Information Security and Communication Technology
Norwegian University of Science and Technology (NTNU)
Gjøvik, Norway

***Abstract*—The threat landscape is constantly evolving. As attackers continue to evolve and seek better methods of compromising a system; in the same way, defenders continue to evolve and seek better methods of protecting a system. Threats are events that could cause harm to the confidentiality, integrity, or availability of information systems, through unauthorized disclosure, misuse, alteration, or destruction of information or information system. The process of developing and applying a representation of those threats, to understand the possibility of the threats being realized is referred to as threat modelling. Threat modelling approaches provide defenders with a tool to characterize potential threats systematically. They include the prioritization of threats and mitigation based on probabilities of the threats being realized, the business impacts and the cost of countermeasures. In this paper, we provide a review of asset-centric threat modelling approaches. These are threat modelling techniques that focus on the assets of the system being threat modelled. First, we discuss the most widely used asset-centric threat modelling approaches. Then, we present a gap analysis of these methods. Finally, we examine the features of asset-centric threat modelling approaches with a discussion on their similarities and differences.**

***Keywords*—Threat modelling; asset-centric; asset-centric threat modelling approaches**

I. Introduction

Threats are events that could cause harm to the confidentiality, integrity, or availability (CIA model [1]) of information systems, through unauthorized disclosure, misuse, alteration, or destruction of information or information system [2]. The process of developing and applying a representation of those threats, to understand the possibility of the threats being realized is referred to as threat modelling. It includes selecting a threat modelling framework and populating that framework with specific values (e.g. adversary expertise, attack patterns and attack events) as relevant to the intended scope (e.g. architectural layers or stakeholder concerns). The populated framework can then be used to construct threat scenarios; characterize controls, technologies, or research efforts; and/to share threat information and responses [3].

Threat modelling methodologies are very few; although there are several frameworks or threat classification models that are usually combined and leveraged by threat modelling methodologies [4]. The choice of threat modelling approach to adopt for a particular situation is dependent on the business objectives. It follows that the first step towards choosing the threat modelling technique to use for a system is to have a clear understanding of what the system being threat modelled is supposed to do. Basically, there are no good or bad threat modelling methods but rather, there are good and bad threat modelling approaches for a particular system.

There are three main approaches that are usually deployed for threat modelling activities and they include: the approaches that focus on the assets of the system being threat modelled, which are referred to as asset-centric threat modelling approaches; the approaches that focus on the attackers, also called the attack-centric threat modelling approaches; and the approaches that focus on the software or the system, which are referred to as software-centric or system-centric threat modelling approaches [5]. We are mainly concern with the asset-centric threat modelling approaches in this paper.

In this paper, we provide a review of asset-centric threat modelling approaches. First, we examine the general objectives and benefits of threat modelling. Also, we present a discussion on the existing threat modelling approaches and justification for reviewing asset-centric threat modelling approaches. We observe that DREAD (damage, reproducibility, exploitability, affected users, discoverability), Trike, OCTAVE (operationally threat asset, and vulnerability evaluation) and PASTA (process for attack simulation and threat analysis) are the most widely used asset-centric threat modelling approaches. The limitation of these approaches is presented. We also examine the features of the asset-centric threat modelling approaches. And using these features, we present a discussion on their similarities and differences. The overall goal of this review is to serve as a foundation for selecting asset-centric threat modelling approaches and to further advance the use of asset-centric methodologies in threat modelling activities.

The rest of this paper is organised as follows. Section II examines the general objectives and benefits of threat modelling. Also, a discussion on the existing threat modelling approaches is presented with the justification for reviewing asset-centric threat modelling approaches. Section III presents state-of-the-art of the most widely used asset-centric threat modelling approaches. Section IV presents gap analysis of the asset-centric threat modelling approaches reviewed. Section V discusses the similarities and differences of the asset-centric threat modelling approaches based on their features. Section

VI concludes the paper and present future work.

II. Background

In this section, we examine the general objectives and benefits of threat modelling. We also present a discussion on the existing threat modelling approaches and justification for reviewing the state-of-the art in the asset-centric threat modelling approaches in this paper.

A. Threat Modelling

Threat modelling is a systematic approach for characterizing potential threats to a system. It ensures completeness by including the prioritization of threats and mitigation based on probabilities, business impacts and cost of countermeasures. Threat modelling provides a means of evaluating all possible risks throughout the system and not just concentrating on where flaws are expected to be discovered [6]. It is also useful in ranking the likelihood of a threat being realized. An essential step for threat modelling is having an understanding of assets and threats [4].

Assets are usually discrete data entities, but they can be physical objects, which feature in the business rules of a system [6]. Assets are artefacts which are important to a specific problem domain of a system, and not just in the actual implementation of a system. Identifying assets can be a very challenging endeavour, but it is the initial step that needs to be carried out in order to understand the amount of resource which can be allocated for threat modelling activities. Also, the amount of threats increases geometrically as the number of assets increases [6].

UcedaVelez and Morana [4] observe that most organizations, businesses, and governments depend on sources such as threat intelligence for the acquisition of threat knowledge. It is obvious that threats would mean different things to different types of organizations. For instance, in the case of private organization, potential threats are those targeting their business assets. For government organizations, potential threats are those relating to national security. Analysing the potential threat scenarios that target an organization's assets is important in determining the likelihood of the threats being realized.

Once the analysis of the potential threat scenarios has been concluded and it shows that the system being threat modelled is at risk, the next step of the risk mitigation strategy is to determine if similar assets are also exposed and can be affected [4]. Also, it is important to consider whether the mitigation measures suggested are able to eliminate the risk to the system without creating additional security threats. This ensures a wholistic mitigation measures are adopted to reduce the business impact of the threat being realized.

Another important factor to consider during threat modelling is the business impact of a threat being realized. A business impact is different from information security risk in that it measures the economic impact caused by either the loss or the compromise of an asset while information security risk affects the confidentiality, integrity and availability of data [4]. Determining the business impact requires a consideration for the business context in which the system operates. This can be achieved by examining at a high level, the assets of the system and the functionality the system provides based on these assets.

In general, threat modelling involves a great amount of effort and resources of so many individuals beyond those of information security [4]. It encourages collaboration and as such, the threat modelling methodology that should be deployed for a particular system may have to consider how collaboration can be fostered. The next subsection presents the different threat modelling approaches. We agree with the authors in [4] that none of these approaches are flawed but rather the way in which they are selected may be flawed.

B. Threat Modelling Approaches

Threat modelling approaches can be categorized according to the focus of the approaches. These approaches include those that focus on the assets of the system being threat modelled, which are referred to as asset-centric threat modelling approaches; the approaches that focus on the attackers, also called attack-centric threat modelling approaches; and the approaches that focus on the software or the system, which are referred to as software-centric or system-centric threat modelling approaches [5]. Deciding which of the method to deploy depends on the system being threat modelled and the tools available.

Asset-centric threat modelling approaches focus on the assets of the system being threat modelled. It involves analysing the information loss or business impact of targeted assets. Asset-centric threat modelling can be extended beyond identifying the motives and intentions of the attacker to incorporating the discovery of security gaps for the system environment [4]. Although, asset-centric threat modelling is not concerned about flaws or insecure coding/design practices, it could be used to uncover possible threats scenarios.

Attack-centric threat modelling approaches include those approaches that focus on the attacker. The idea here is to examine the threats against a system from the perspective of an attacker. Attack-centric threat modelling approach aims to identify which threats can be successfully executed against a system given a number of identified misuse cases, vulnerabilities, and more [4]. Also, the approach attempts to examine the motive, sources and relative identity of the attacker or group associated with the attacker as these can help to uncover the approach and resources of the attacker [4].

System-centric threat modelling approaches focus on the system being threat modelled. They first consider the design model of the system under consideration. The objective of these approaches is to ensure that the complexity of the system being threat modelled is well understood before considering threats the system may be exposed to. System-centric threat modelling approaches expects those involved in threat modelling of a system, to have a good grasp of the system they are developing [5].

In this paper, we interested in understanding the state-of-the-art in asset-centric threat modelling approaches. It is usually the case that most businesses have a clear understanding of their business objectives and assets to be protected. Also, the system to be threat modelled and the business impacts of threats being realized are likely to be known. Thus, the obvious threat modelling approaches that can be employed for the protection of assets, understanding and managing business risks for most businesses are the asset-centric threat modelling approaches. Therefore, we present this review to serve as a basis for selecting or combining the appropriate asset-centric threat modelling approaches and to further advance the use of asset-centric threat modelling techniques.

III. The State-of-the-Art in Asset-Centric Threat Modelling Approaches

In this section, we present a review of asset-centric threat modelling approaches. We observe that the most widely used asset-centric threat modelling approaches are DREAD, Trike, OCTAVE, and PASTA. We use this understanding to present a discussion on the state-of-the-art in asset-centric threat modelling approaches.

A. DREAD

DREAD is an acronym for Damage potential, Reproducibility, Exploitability, Affected users, and Discoverability. It is an asset-centric threat modelling approach developed by Microsoft. DREAD uses the traditional qualitative risk rating (HIGH, MEDIUM, LOW) with a qualitative risk rating 3,2,1 applied respectively. In general, DREAD threat modelling approach uses a scoring system to calculate the probability of occurrence for each of the identified areas of the asset being threat modelled. By combining the risk rating values obtained, DREAD threat modelling approach is able to predict the probability of occurrence of each threat identified during the threat modelling process [4].

The Damage potential refers to the level of havoc that could be done to users and the organization if an attack were to succeed. Damage could be concrete, such as financial liability or abstract, such as damage to organization's reputation. Also, it depends on the nature of the attack and the assets being targeted. Reproducibility measures the easy with which the attack can be replicated. The goal is to measure the effort that would be expended by an attacker for the realization of an attack and use such measure, in the scoring system. If an attack can be reproduced with much ease, the attack would be rated high in the scoring system as against an attack that cannot be reproduced with much ease.

The remaining letters of DREAD are described as follows. Exploitability describes the possibility of an attacker taking advantage of a vulnerability. Several exploits exist and they can be classified as those that are easily understood and could be accomplished by anyone and those that are difficult that required specialized skills to achieve. This understanding is used to rate threat that have high level of exploitability as high risk in the scoring system and those with low level of exploitability as low risk. Affected users refers to the number of users that will be affected by the realization of a particular threat. A threat that is likely to affect a great number of users when realized would have a higher risk factor rating compared to a threat that is likely to affect a limited number of users. Discoverability describes the ease with which the vulnerability is uncovered. There are threat that are very difficult to learn and those that can be learn with ease. Hence, a threat that is very difficult to learn would be rated lower than those that has been released in the public domain. The DREAD approach is summarised in Fig. 1.

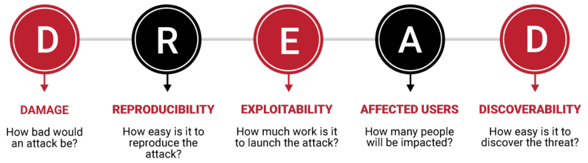

D
R
E
A
D
DAMAGE
How bad would an attack be?
REPRODUCIBILITY
How easy is it to reproduce the attack?
EXPLOITABILITY
How much work is it to launch the attack?
AFFECTED USERS
How many people will be impacted?
DISCOVERABILITY
How easy is it to discover the threat?

Fig. 1. DREAD Summary [7]

Although DREAD is an asset-centric threat modelling approach, several of its application in the literature is in combination with STRIDE (spoofing, tampering, repudiation, information disclosure, denial of service, and elevation of privilege) model [8], [9], [10], [11], [12]. In this type of approach, the DREAD scoring scheme is used to identify the likelihood that an attack is able to exploit a particular threat.

B. Trike

Trike offers a threat modelling approach which is asset-centric and it achieves that through the generation of threat models in a reliable, and repeatable manner [6]. It facilitates constructive interaction among relevant stakeholders by providing standardized framework for reasoning about threats that a system would have to overcome. The achievement of Trike objectives entail the following [6]:

- With assistance from the system stakeholders, ensure that the risk the system presents to each asset is acceptable to all stakeholders.
- Be able to tell whether that have been done.
- Communicate what have been done and it effects to the stakeholders.
- Empower stakeholders to understand and reduce the risks to themselves and other stakeholders implied by their actions within their domains.

Another important observation about Trike is that it follows a defensive approach. Understanding the system itself and the environment in which the system is going to be used is more important when using Trike threat modelling approach than understanding the capability of an attacker. This is because without a complete knowledge of the system, it is difficult to appropriately characterize the threat that a system would have to face [6].

C. OCTAVE

OCTAVE is another asset-centric threat modelling approach. It is an acronym for Operationally Threat Asset, and Vulnerability Evaluation. OCTAVE methodology takes the advantage of people's understanding of their organization's security-related practices and process to model the state-of-the-art of security practice within the organization. Threat to the most critical assets are used to prioritize areas of improvement and to set security strategy for the organization [13].

The two aspects that are the foundation of OCTAVE approach include: operational risk and security practices. The security practices encompasses efforts by an organization to refine its existing security practices. Technologies deployed by an organization in meeting its business objectives are evaluated in relation to security practices. For the operational risks, an organization considers all aspects of risk (asset, threats, vulnerabilities, and organization impact) in its decision making enabling the organization to match a practice-based protection strategy to its security risks [13]. The OCTAVE process is depicted in Fig. 2.

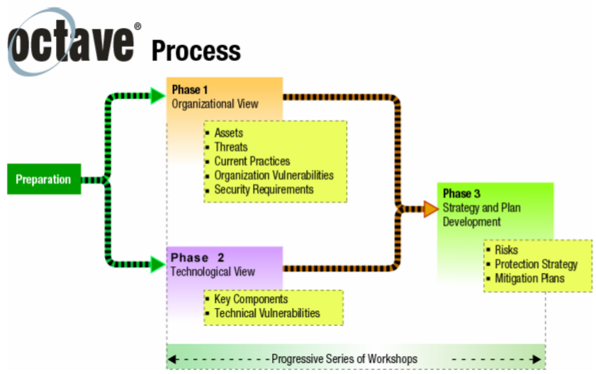


Fig. 2. OCTAVE Process [13]

The evaluation process of OCTAVE approach involves the following [13]:

- Identify information-related assets that are important to the organization;
- focus risk analysis activities on those assets judged to be most critical to the organization;
- consider the relationships among critical assets, the threat to those assets, and vulnerabilities (both organization and technological) that can expose assets to threats;
- evaluate risks in an operational context, i.e. how they are used to conduct an organization's business and how those assets are at risk due to security threats;
- create a practice-base protection strategy for organizational improvement as well as risk mitigation plans to reduce the risk to the organization's critical assets.

In addition, the evaluation process for the organizational, technological, and analysis aspects are complemented by a three-phased approach, namely build asset-based threat profiles; identify infrastructure vulnerabilities, and develop security strategy and plans [13].

It is also imperative to note that the essential elements or requirements of the OCTAVE approach are captures in a set of criteria [13]. As of now, there are three methods consistent with the criteria and they are: the OCTAVE Method, that is designed for large organization; the OCTAVE-S, which is well-suited for small organizations; and the most recent version called the OCTAVE Allegro. The OCTAVE Allegro has been applied in [14] to evaluate the security risks of IoT (Internet of things) based smart homes. The authors in [15] developed a university information security risk management framework using OCTAVE Method based on ISO/EIC 27001:2013. Also, the OCTAVE-S has been combined with ISO 27001:2005 in [16] for risk management.

D. PASTA

PASTA is an acronym for Process for Attack Simulation and Threat Analysis which is an asset-centric threat modelling approach. It combines topicality, substantiation, and probabilistic analysis as the key three attributes as part of its methodology [4]. According to UcedaVelez and Morana [4], PASTA approach can be deployed in almost any scenario except for those scenarios where executive sponsorship of its process and produced artefacts is not available. This is because the deliverables produced by the PASTA approach are supposed to be familiarized with the organization's executives too.

When adopting and executing PASTA threat modelling approach, it is essential to review the following: sponsorship and support (without executive support the process will not succeed); maturity, as the maturity of the processes and controls employed will affect the outcome of PASTA; awareness, efficient and effective communication is required for the entire activities; input and outputs, people are the main input to consider for the threat modelling activity and outputs are to be defined for each process involved in the threat modelling; and lastly participants are recruited and retrain [4].

For the actual deployment, the PASTA threat modelling methodology include the following stages. The first stage involves defining objectives, where the business objectives of the system to be threat modelled is clearly defined. The technological scope is defined in the second stage and it involves identifying all the assets of the system. Next, the system is decomposed to facilitate an understanding of the system's operations. In the fourth stage, threat analysis is carried out to identify threats to the system. Then, weakness and vulnerability analysis which allows vulnerable areas across the system to be identified and mapped to the attack tree introduced in the threat analysis stage. Attack modelling and simulation is followed and the focus is to study the possibility that the identified vulnerabilities can be exploited. Lastly, residual risk analysis and management is done to mitigate threat that are major concerns to the system. All these stages are shown in Fig. 3.

IV. Limitation of the Asset-Centric Threat Modelling Approaches

In this section, we present a gap analysis of the asset-centric threat modelling approaches discussed in Section III.

1) DREAD: has been shown to be fairly subjective and leads to inconsistent results [3]. In fact, as of 2010, Microsoft discontinued the use of DREAD for their software development life-cycle [3]. This further underscores the limitation of DREAD as a threat modelling approach. However, DREAD is still widely used and recommended for threat and risk modelling endeavours. Hence, useful suggestions have been made in [17] on modifications to the scoring scheme in order to improve its reproducibility.

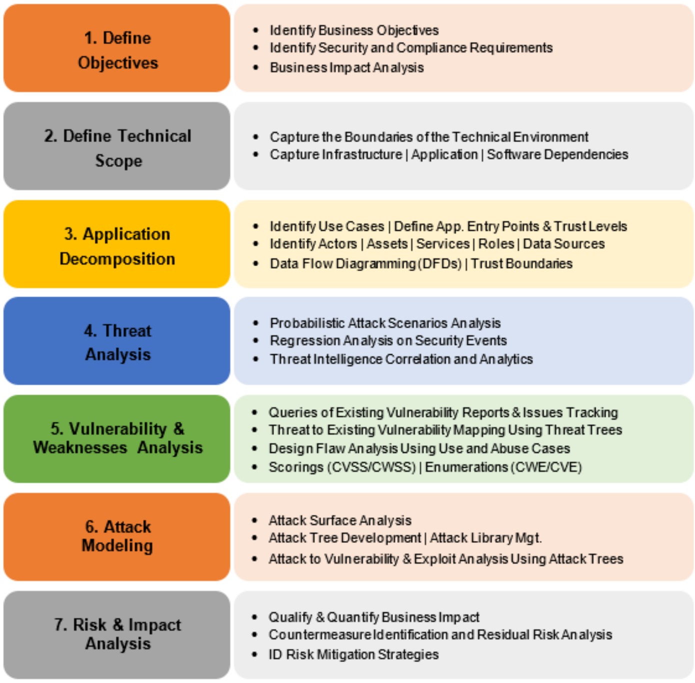

1. Define Objectives
Identify Business Objectives
Identify Security and Compliance Requirements
Business Impact Analysis
2. Define Technical Scope
Capture the Boundaries of the Technical Environment
Capture Infrastructure | Application | Software Dependencies
3. Application Decomposition
Identify Use Cases | Define App. Entry Points & Trust Levels
Identify Actors | Assets | Services | Roles | Data Sources
Data Flow Diagramming (DFDs) | Trust Boundaries
4. Threat Analysis
Probabilistic Attack Scenarios Analysis
Regression Analysis on Security Events
Threat Intelligence Correlation and Analytics
5. Vulnerability & Weaknesses Analysis
Queries of Existing Vulnerability Reports & Issues Tracking
Threat to Existing Vulnerability Mapping Using Threat Trees
Design Flaw Analysis Using Use and Abuse Cases
Scorings (CVSS/CWSS) | Enumerations (CWE/CVE)
6. Attack Modeling
Attack Surface Analysis
Attack Tree Development | Attack Library Mgt.
Attack to Vulnerability & Exploit Analysis Using Attack Trees
7. Risk & Impact Analysis
Qualify & Quantify Business Impact
Countermeasure Identification and Residual Risk Analysis
ID Risk Mitigation Strategies

Fig. 3. PASTA Stages [4]

2) Trike: requires an analyst undertaking a threat modelling exercise to have full a grasp of the whole system while assessing the risk of attacks. This can be very challenging if the system to be threat modelled is very large. Also, the authors in [18] observed that the Trike scoring system is too vague to represent a formal. In addition, Trike does not have sufficient documentation even though its website is still available.

3) OCTAVE: is a robust, asset-centric threat modelling approach but it is highly complex. It takes considerable time to learn and the processes involved can be time consuming. Also, OCTAVE documentation can become voluminous, which is likely to discourage policy makers from adopting it as a threat modelling approach for their organization.

Another limitation of OCTAVE threat modelling approach is the way in which the identification and classification of threat is achieved. The capturing of risks and threats using the threat tree when OCTAVE is employed can become undesirable for complex environment. As the number of paths increases in the case of a very large computing environment, it may become unclear which of the paths represent the threats being modelled.

4) PASTA: is design for organizations that desire to position threat modelling with their strategic objectives. This is because PASTA incorporates business impact analysis as an important part of the PASTA process, which extends security responsibilities to the entire organization. This positioning can become a drawback for using PASTA because it may require several hours of training and education of the key stakeholders.

V. Discussion

This section presents a discussion on the similarities and differences of the asset-centric threat modelling approaches we have presented so far. First, the features of the asset-centric threat modelling approaches are given in Table I. We then provide a discussion on their similarities and differences.

A feature that is common to all the asset-centric threat modelling approaches as can be observed from Table I, is the fact that they all contribute to risk management process. In fact, asset-centric threat modelling approaches are sometimes referred to as risk-based threat modelling approaches [4]. They employ a risk-based approach in analysing the business impact of possible threat scenarios. This can then be used to prioritize threat mitigation strategies, which is also a feature that all the asset-centric threat modelling approaches we have presented in this paper possesses.

Apart from DREAD, the remaining asset-centric threat modelling approaches encourage collaboration among the stakeholders and can be used to identify relevant mitigation techniques. Collaboration is an essential part of any threat modelling activities. Considering that majority of the asset-centric threat modelling approaches presented in this review encourage collaboration among relevant stakeholders further buttress the importance of collaboration during threat modelling process. Mitigation techniques ensures that actionable steps which can help to avoid the threats identified during the threat modelling process are recommended.

Another important desirable characteristics of any threat modelling approach are reproducibility and automation. Reproducibility refers to the ability of the threat modelling approach to have consistent results when repeated. Unfortunately, the only asset-centric threat modelling approach that seems to have such property is OCTAVE. Other approaches are usually subjective and depend on those carrying out the threat modelling activities. Automation ensures that the threat modelling process can be undertaken without human intervention. As of now, only Trike has automated components and given the insufficient documentation there is still a lot of work to be done in automating asset-centric threat modelling approaches.

VI. Conclusion

Asset-centric threat modelling approaches have shown to be effective for the protection of assets, understanding and managing business risks. In this paper, we have reviewed the state-of-the-art in asset-centric threat modelling approaches. We have observed that DREAD, Trike, OCTAVE, and PASTA are the most widely used asset-centric threat modelling approaches. Then, we present a discussion on the state-of-the-art of these approaches. Also, a gap analysis of these approaches is discussed. Finally, we describe the features of the asset-centric threat modelling approaches we have reviewed, with a discussion on their similarities and differences.

In the future, we hope to explore formal methods that can exploit asset-centric threat modelling approach to reason about the potential threats to a cyber-physical system. This is because the asset-centric threat modelling approaches we have reviewed in this paper are not suitable for capturing the potential threats to a cyber-physical system due to the timing, uncertainty, and dependencies that exist between its entities. Although, several attempts have been made in the literature to threat model cyber-physical systems [19], [20], [21], we intend to use the formal method for expressing the requirements that are unique to a

TABLE I. FEATURES OF ASSET-CENTRIC THREAT MODELLING APPROACHES

Asset-centric Threat Modelling Approach	Features
DREAD	● Helps to assess risk associated with a threat exploit ● Can predict the probability of an exploit being realized ● Contributes to risk management ● Has built-in prioritization of threat mitigation ● Offers flexibility and can be applied and adopted to any situation
Trike	● Encourages collaboration among stakeholders ● Has built-in prioritization of threat mitigation ● Has automated components ● Contributes to risk management ● Can identify mitigation techniques
OCTAVE	● Encourages collaboration among stakeholders ● Has built-in prioritization of threat mitigation ● Has consistent results when repeated ● It is designed to be scalable ● Contributes to risk management ● Can identify mitigation techniques
PASTA	● Encourages collaboration among stakeholders ● Has built-in prioritization of threat mitigation ● Contributes to risk management ● Can identify mitigation techniques.

cyber-physical system in order to facilitate the identification of potential threats to the system.

REFERENCES

[1] L. O. Nweke, "Using the cia and aaa models to explain cybersecurity activies," *PM World Journal*, vol. 6, 2017.

[2] NIST, "Information security: Guide for conducting risk assessments," Sep. 2012. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf

[3] D. J. Bodeau, C. D. McCollum, and D. B. Fox, "Cyber threat modeling: Survey, assessment, and representative framework," The Homeland Security Systems Engineering and Development Institute, Tech. Rep., 2018.

[4] T. UcedaVelez and M. M. Morana, *Risk Centric Threat Modeling: Process for Attack Simulation and Threat Analysis*. John Wiley & Sons, May 2015.

[5] A. Shostack, *Threat Modeling: Designing for Security*. John Wiley & Sons, Feb. 2014.

[6] M. Eddington, B. Larcom, and E. Saitta, "Trike v1 methodology document," 2005.

[7] Wildcard, "Threat modeling." [Online]. Available: https://wildcardcorp.com/security/threat-modeling

[8] M. Cagnazzo, M. Hertlein, T. Holz, and N. Pohlmann, "Threat modeling for mobile health systems," in *Proc. IEEE Wireless Communications and Networking Conf. Workshops (WCNCW)*, Apr. 2018, pp. 314–319.

[9] A. Omotosho, B. A. Haruna, and O. M. Olaniyi, "Threat modeling of internet of things health devices," *Journal of Applied Security Research*, vol. 14, pp. 106–121, 2019.

[10] A. Amini, N. Jamil, A. R. Ahmad, and M. R. Z'aba, "Threat modeling approaches for securing cloud computin," *Journal of Applied Sciences*, vol. 15, pp. 953–967, 2015.

[11] M. Abomhara, M. Gerdes, and G. M. Køien, "A stride-based threat model for telehealth systems," *Norsk informasjonssikkerhetskonferanse (NISK)*, vol. 8, no. 1, pp. 82–96, 2015.

[12] M. Hagan, F. Siddiqui, and S. Sezer, "Policy-based security modelling and enforcement approach for emerging embedded architectures," in *Proc. 31st IEEE Int. System-on-Chip Conf. (SOCC)*, Sep. 2018, pp. 84–89.

[13] C. Alberts, A. Dorofee, J. Stevens, and C. Woody, "Introduction to the octave approach," 2003.

[14] B. Ali and A. I. Awad, "Cyber and physical security vulnerability assessment for iot-based smart homes," *Sensors*, vol. 18, no. 3, p. 817, 3 2018. [Online]. Available: http://www.mdpi.com/1424-8220/18/3/817

[15] I. Sulistyowati and R. H. Ginardi, "Information security risk management with octave method and iso/eic 27001: 2013 (case study: Airlangga university)," *IPTEK Journal of Proceedings Series*, no. 1, pp. 32–38, 2019.

[16] S. Stephanus, "Implementation octave-s and iso 27001controls in risk management information systems," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 5, no. 2, p. 685, 2014.

[17] D. Leblanc, "Dreadful," 2007. [Online]. Available: https://blogs.msdn.microsoft.com/david-leblanc/2007/08/14/dreadful/

[18] N. Shevchenko, T. A. Chick, P. O'Riordan, T. P. Scanlon, and C. Woody, "Threat modeling: a summary of available methods," *no. July*, 2018.

[19] E. B. Fernandez, "Threat modeling in cyber-physical systems," in *Proc. nd Intl Conf Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech) 2016 IEEE 14th Intl Conf Dependable, Autonomic and Secure Computing, 14th Intl Conf Pervasive Intelligence and Computing*, Aug. 2016, pp. 448–453.

[20] M. Rekik, C. Gransart, and M. Berbineau, "Cyber-physical threats and vulnerabilities analysis for train control and monitoring systems," in *Proc. Computers and Communications (ISNCC) 2018 Int. Symp. Networks*, Jun. 2018, pp. 1–6.

[21] Y. Atif, Y. Jiang, D. Jianguo, M. Jeusfeld, B. Lindström, S. Andler, C. Brax, D. Haglund, and B. Lindström, "Cyber-threat analysis for cyber-physical systems," 2018.

Fast and Accurate Fish Detection Design with Improved YOLO-v3 Model and Transfer Learning

Kazim Raza[1], Song Hong[2]
Department of Ocean Science and Engineering
Ocean Opto, Electronics and Automation Lab
Zhejiang University, Hangzhou Zhejiang 310058, China

Abstract—**Object Detection is one of the problematic Computer Vision (CV) problems with countless applications. We proposed a real-time object detection algorithm based on Improved You Only Look Once version 3 (YOLOv3) for detecting fish. The demand for monitoring the marine ecosystem is increasing day by day for a vigorous automated system, which has been beneficial for all of the researchers in order to collect information about marine life. This proposed work mainly approached the CV technique to detect and classify marine life. In this paper, we proposed improved YOLOv3 by increasing detection scale from 3 to 4, apply k-means clustering to increase the anchor boxes, novel transfer learning technique, and improvement in loss function to improve the model performance. We performed object detection on four fish species custom datasets by applying YOLOv3 architecture. We got 87.56% mean Average Precision (mAP). Moreover, comparing to the experimental analysis of the original YOLOv3 model with the improved one, we observed the mAP increased from 87.17% to 91.30. It showed that improved version outperforms than the original YOLOv3 model.**

Keywords—Deep learning; computer vision; transfer learning; improved YOLOv3; anchor box; custom dataset

I. Introduction

Deep learning (DL) is the subfield of Machine learning (ML), which is built on artificial neural networks that can be unsupervised, semi-supervised, or supervised learning. The methods of DL are characterization learning methods that acquired from nonlinear modules to transform raw data representation into a higher level. The core aspect of DL is that layers acquired from the given data, unlike humans [1]. Researchers tried hard to train a deep multi-layer network for decades, but still, before 2006, there were not many successful experiments at that time where they only passed on effective results with one or two hidden layers. Those results were not producing substantial outcomes due to exploding gradients. DL is like a sensory system where the flow of information having internal connections with all of the neurons, and every neuron helps to process the information to the next one.

There is a massive difference between DL and ML, ML only relies on structured data, whereas DL required layers of the Artificial Neural networks. Szeged at el [2] Deformable Part Model (DPM) is one of the top techniques for object recognition that's implementation is established on the decomposition of the object and expressed in graphical mode. This model has only two layers that are not useful for the big dataset. Traditional ML classifiers likewise SVM, LDA, which is insufficient for huge dataset classification. The hierarchical Classification is quite exceptional than SVM because of its 4% accuracy results than a flat SVM classifier [3]. In the previous traditional methods, the researches never used the deep Convolutional Neural Network (CNN) design as they were using a tiny dataset that has a low range of images and a restricted number of fish species. Another critical point to remember, they were using handcrafted ways; that is why performance was not up to the mark. The implemented algorithm was inadequate for a big dataset, and resultantly the accuracy not achieved consequently. In the recent past, the Fast R-CNN, and faster R-CNN gain significant research performance, but these architectures have a very complex execution pipeline to perform recognition tasks. These architectures have less Frame Per Second (FPS) and accuracy as well. We proposed the YOLOv3 real time object detection model in our research work.

The major contribution of this work is given as follows:

- We improved the model by the addition of a 4th detection scale in the network to enhance the performance by obtaining finer-grained features.
- Applied K-means++ clustering on our dataset to get suitable anchor boxes and increase the anchor boxes from 9 to 12.
- Applied a novel transfer learning method to improve efficiency.
- We also changed the loss function for learning and convergence in the model.

The rest of the paper is described as follows:

Section II explains the background study. Section III explains the research methodology, including improvements in the methodology. Section IV explains the dataset composition and its structure. Section V explains the results and comparison of different state of the art object detectors. In the end, Section VI explains the conclusion and future direction.

II. Background Study

In the deep ocean, the movement of the fish is unpredictably quick and three-dimensionally; therefore, recognition is a difficult task. Fish recognition depicts to identify different types of fish species according to their features. It is essential to locate for other kinds of reasons, including contour and pattern matching, statistical, quality control, feature extraction, and determination of physical traits [4]. Larsen at el. [5] obtained the shape and texture feature from appearance model and testing on the dataset, which has been containing more than 100 images of three fish species and attaining the accuracy 76%. Helge Balk at el. [6] developed the Sonar5 post-processing program that covered interpretation, analysis, and acquisition stages of hydroacoustic fish detection. The fish-echoes, along with surrounding noise level, can be detected using this program due to its time variation in sonar's detection, so the overall accuracy was high. Fuming Xiang at el. [7] used CNN models pipeline, including VGG16 and SSD on 9 common species of fish in the Missouri river to classify into category and position. They have achieved 87.22% accuracy in the classification of the fish.

Recognizing fish is one of the possibilities that come out with DL, which helps to find the targeted underwater species, i.e., fish. There are hundreds of applications to recognize marine fish, and many practices have already been done to find the right one object, which helps people to solve the problem. Tracking and counting the fish is also crucial for fish industry and conservation purposes as well.

The exact quantity of slaughtering fish is not final yet. Still, there is an estimated figure that salmon, sea trout, and migratory char are 27.0% decreased in killing fish from 2017 to 2018, according to Statistics Norway [8]. As per the report, the global river catch has passed to almost 10 million right after the linear growth from the 1950s, which was under-reported on collecting the relevant data in the past [9]. There is no certainty on how much river fish caught, released, or slaughtered after catching from the river or ocean, so this thing needs some automation with an accuracy of data.

Moreover, the caught fish is healthy or not needs some consideration and observation to determine whether the fish is healthy as not all fishes can be healthy.

For all of such problems, the CNN does help in the classification of the marine system, observing the behavior of the underwater object, tracking an accurate object, automated, accurate counting of fish caught globally, localization, and controlling the environment.

There almost 20 deep neural networks have been trained for Salmon fish recognition that provides an in-depth discussion of each model with parameter tuning [10]. Moreover, SSD version 2 achieved 84.64% mAP, state-of-the-art accuracy with 3.75 FPS for salmon recognition. Background subtraction method used to detect and track fish in marine life with the help of a video sequence. They get an accurate 73% result from the real type of video though they get the best result by implementing the Viola-Jones method using Haar cascade [11].

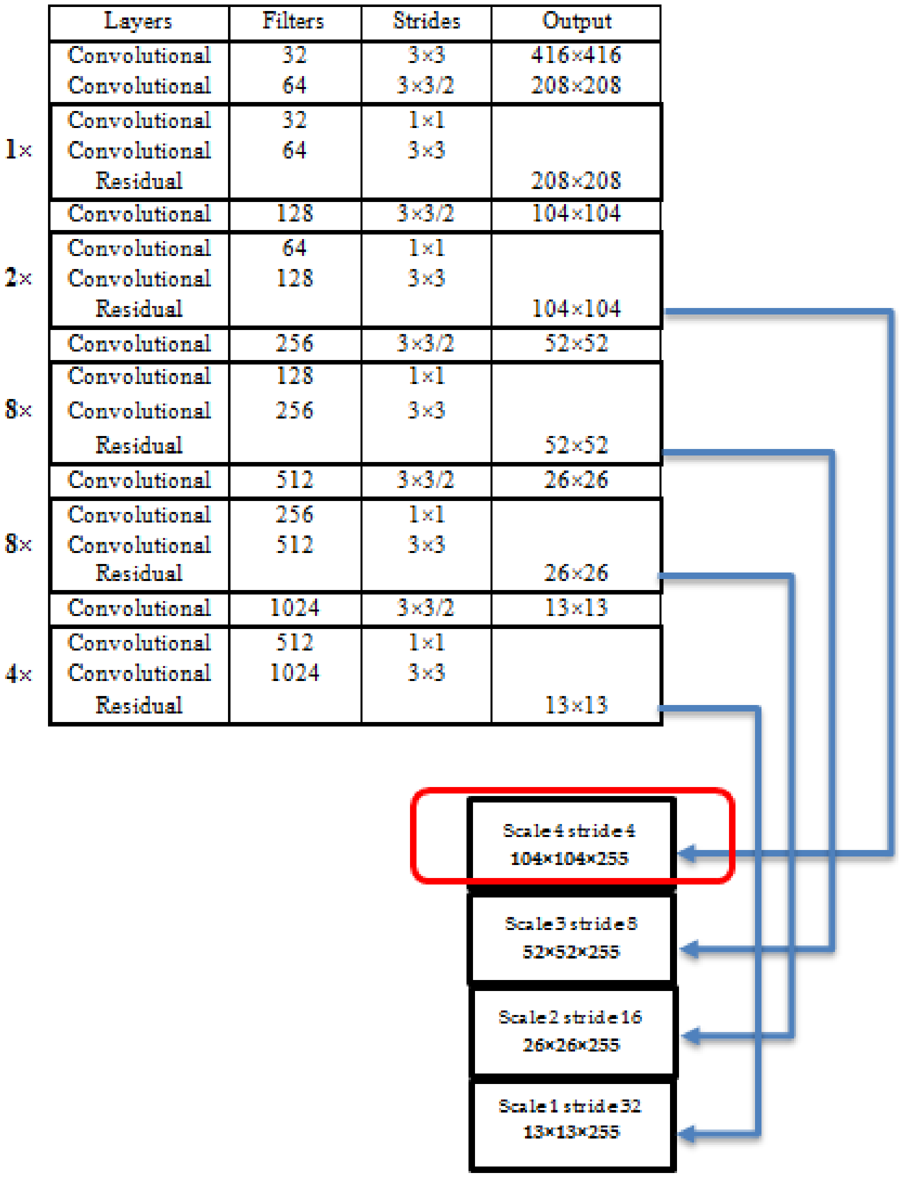
Undoubtedly, fish recognition is a complex task where some of the challenges like noise, distortion, overlap, occlusion, and segmentation error needs several techniques to get some accuracy in the result. Some of the techniques have already applied, and one of the SVM based techniques used on the two training sets on the fish features [12]. One was containing 74 fish testing set, and the other was about 76 fish. The final result based on SVM showed 78.59% accuracy in the fish classification. Dhruv Rathi et al. [13] derived a method based on CNN for the automation classification of fish species, which achieved 96.29% accuracy than other proposed systems.

III. Research Method

Object recognition and detection are important issues in CV problems. Based on the detection pipeline and backbone architecture, the object detector algorithm classified into two types (1) two-stage object detectors such as fast R-CNN [14], faster R-CNN [15], Mask R-CNN [16], and (2) single-stage object detectors such as SSD [17], YOLO [18], YOLOv2 [19], YOLOv3[20]. The two-stage detection algorithm computationally very complex because they have separate backbone architecture. The single-stage object detector models are computationally less complex than that of the two-stage detector. The single stage detection algorithm like YOLOv3 is much faster, and the accuracy of YOLOv3 and faster R-CNN have no larger difference. So we implement the YOLOv3 object detection model in this paper, which is a fast and real- time object detection model. For the feature extraction, YOLOv3 use darknet-53 as a backbone architecture. The first and second versions of YOLOv3 architecture struggle with small object recognition. As we detect fishes so this 53 convolutional layers’ architecture for feature extractor is the best choice. The backbone architecture of YOLOv3 still performs better than ResNet-101 and ResNet-152.

The backbone darknet-53 holds 23 residual units, and every such unit performs the 1 × 1 and 3 × 3 convolutional. At the end of every residual unit, an element-wise addition carried out between the input and output vectors. Every convolutional layer pursued by the Leaky ReLU activation function, where Batch Normalization is using. The downsampling runs with a stride of 2 at five separate convolutional layers.

YOLOv3 implements a Feature Pyramid Network (FPN) that used to detect the objects at different scales that constructs FPN on top of backbone architecture and build a pyramid with downsampling strides, 8, 16, and 32 in order to detect all-sized objects. The improved network structure of Darknet-53 shown in Fig. 1. We proposed the 4th scale to increase the detection performance where the red box represents the 4th detection scale, which helps to increase the detection of extra small objects with the downsampling stride of 4×. It helps us to get more exceptional grained features to detect extra small size targeted objects. The experimental scheme of the proposed work is shown in Fig. 2, which illustrates the initial dataset collection, pre-processing, and labeling of the dataset. Then we applied transfer learning on our custom dataset and fine-tuned the model to get better results on the custom dataset. We trained our model as much as it is converged and finally checked the visualization detection results and evaluation of the model.



	Layers	Filters	Strides	Output
	Convolutional	32	3×3	416×416
	Convolutional	64	3×3/2	208×208
1×	Convolutional	32	1×1	
	Convolutional	64	3×3	
	Residual			208×208
	Convolutional	128	3×3/2	104×104
2×	Convolutional	64	1×1	
	Convolutional	128	3×3	
	Residual			104×104
	Convolutional	256	3×3/2	52×52
8×	Convolutional	128	1×1	
	Convolutional	256	3×3	
	Residual			52×52
	Convolutional	512	3×3/2	26×26
8×	Convolutional	256	1×1	
	Convolutional	512	3×3	
	Residual			26×26
	Convolutional	1024	3×3/2	13×13
4×	Convolutional	512	1×1	
	Convolutional	1024	3×3	
	Residual			13×13

Fig. 1. Improved Network Structure of YOLOv3.

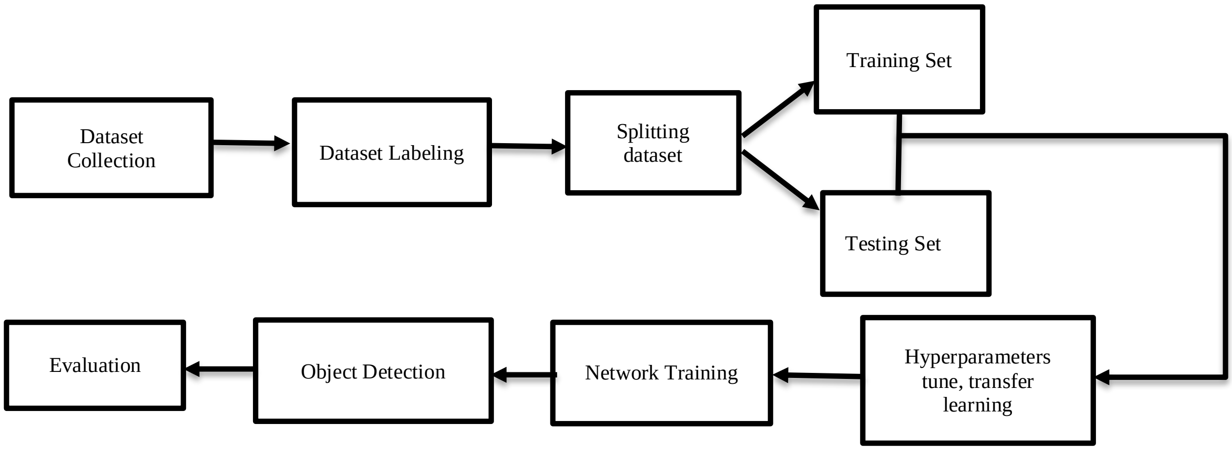


Fig. 2. Dataset Management and Detection Flow Diagram.

A. Algorithm Implementation Parameters

In improved YOLOv3, we have changed several default parameters to make the algorithm more robust. Unlike YOLOv3 in improved YOLOv3, we enhance the FPN because of increasing detection scales. Addition of 3 types of data augmentation in the algorithm for better training and testing results, improvement in the loss function, increase in anchor boxes, configuring of the tensor board to visualize the entire network performance. The algorithm parameters are shown in Table I.

B. K-means ++ Clustering

YOLOv3 used the idea of anchor boxes during the prediction of a bounding box. We increased the detection scale from 3 to 4 and used a custom dataset. With these effects in the network model, we ran a k- means++ clustering algorithm on our dataset to get the suitable size of anchor boxes for more improvement in detection accuracy. Besides, we increased the anchor boxes from 9 to 12 because we increased the detection scales from 3 to 4. Assign 3 anchor boxes to each detection scale depending upon the size of the object. The 12 anchor boxes generated by running the k-means++ clustering on our dataset are: (38, 23), (78, 52), (112, 84), (127, 117), (194, 98), (165, 139), (243, 155), (199, 205), (297, 237), (302, 280), (286, 343), (318, 374).

C. Improved YOLOv3 Loss Function

In the original paper of YOLOv3, the author used logistic regression to predict an objectness score for each bounding box that calculated the cost function. The objectness score is 1 if the anchor box overlaps the ground truth by more than or equal to a specific threshold value. On the other hand, if it still overlaps ground truth by less threshold value, that will not be considered the best bounding box. In Equation (1), we can see how the network output is changed by bounding box predictions where coordinates tx, ty, tw, th are responsible for computing the prediction.

$$b_x = \sigma\ (t_x)+c_x$$

$$b_y = \sigma\ (t_y)+c_y \qquad (1)$$

$$b_w = p_w e^{tw}$$

$$b_h = p_h e^{th}$$

The loss function is responsible for calculating the error between the real values and predicted one. The YOLOv3 loss function is the total sum of the coordinate loss, class loss, and confidence loss defined in equations (2), (3), and (4).

$$\mathrm{Loss}_{\mathrm{coord}} = \lambda_{\mathrm{coord}} \textstyle\sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[\left(x_i - \widehat{x}_i\right)^2 + \left(y_i - \widehat{y}_i\right)^2\right],$$

$$+\lambda_{\mathrm{coord}} \textstyle\sum_{i=0}^{s^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left[\left(\frac{w_i - \widehat{w_i}}{\widehat{w_i}}\right)^2 + \left(\frac{h_i - \widehat{h_i}}{\widehat{h_i}}\right)^2\right] \qquad (2)$$

$$\mathrm{Conf}_{\mathrm{loss}} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} \left(C_i - \widehat{C}_i\right)^2$$

$$+\ \lambda_{\mathrm{noobj}}\ \textstyle\sum_{i=0}^{S^2}\ \sum_{j=0}^{B} 1_{ij}^{noobj} \left(C_i - \widehat{C}_i\right)^2 \qquad (3)$$

$$\mathrm{Class}_{\mathrm{loss}} = \textstyle\sum_{i=0}^{S^2} 1_i^{obj} \sum_{c \in \mathrm{classes}} \left(p_i(c) - \hat{p}_i(c)\right)^2 \qquad (4)$$

Above loss function, x_i, y_i are the center coordinates of the ith box grid cell. $h_{i.}$ w_i are the height and width and height of the i-th grid cell, respectively x_i, y_i, w_i and h_i are the real value and $\widehat{x_\iota}$, $\widehat{y_\iota}$, $\widehat{w_\iota}$ and $\widehat{h_\iota}$ are the predicted values. $(p_i(c)$ is the probability of a class and $(p_i\widehat{(c)}$ is the corresponding prediction value. λ_{coord} coordinate loss of weights and λ_{noobj} is the bounding box object loss without weights. 1_{ii}^{obj} denotes that the j-th box predictor in cell i is “responsible” for that prediction. We used $\left(\frac{w_i - \widehat{w_i}}{\widehat{w_i}}\right)^2$ and $\left(\frac{h_i - \widehat{h_i}}{\widehat{h_i}}\right)^2$ rather than $w^i - \widehat{w}^i$ and $h^i - \hat{h}^i$ which helps to reduce the effect of different sizes of an object of the same kind. S^2 Denotes the grid cell B denotes the bounding boxes and 1_i^{obj} denotes the object existence in cell i or not.

D. Transfer Learning

A transfer learning method developed to attain better performance with more transferred feature layers. Transfer learning is being used to extract the features from a custom dataset automatically with the help of using pre-trained models. It is a suitable way to apply transfer learning without considering substantial datasets, training, and calculation, which only consumed the time. Transfer learning is an adequate method if one has a small-scale sample dataset. Transfer learning used pre-trained CNN architecture, where almost 1.2 million samples of ImageNet dataset and 1000 classes have trained with powerful features extraction potential.

TABLE. I. ALGORITHMS PARAMETERS

Algorithms parameters	
YOLOv3	***Improved YOLOv3***
9 anchor boxes	12 anchor boxes
Default loss function	Improved loss functions
Darknet53, 53 Conv layers with 3 YOLO layers	Darknet53, 53 Conv layers, with 4 YOLO layers
Configure Upsampling layers,	Configure Upsampling layers,
Configure residual blocks, Conv 3×3, 1×1, concatenate	Configure residual blocks, Conv 3×3, 1×1, concatenate
Fine-Grained Features	Deep Fine-Grained Features
Batch Normalization	Batch Normalization
CUDA implement in the algorithm	CUDA implement in the algorithm
Train from scratch	Train using transfer learning,fine tuning
No FPS	Compute FPS in the algorithm
Single image size training	Multiscale image training
Train without data augmentation	Train with data Augmentation
Default batch size	Change batch size
No tensorboard Visualization	Tensorboard visualization in code
3 detection scales	4 detection scales
Configuring IOU, mAP	Configuring IOU, mAP

We proposed and trained darknet-53 backbone architecture, which is pre-trained on the ImageNet dataset to extract the features. Then we performed target detection on the COCO dataset by fine-tuning. During the fine-tuning, we adjust several parameters, including the multi-scale size of input images, learning rate, batch size, to boost and enhance the accuracy and performance.

IV. Dataset Composition

The dataset is a key for object detection, and the collection of the dataset is an important, challenging milestone for object recognition. We used four kinds of fish, including anemone-fish, jelly-fish, star-fish, and shark. The samples of the dataset collected from various resources. All the samples of the dataset have varying sizes, such as 320×320, 416×416, and 480×480. The sample of the collected dataset is shown in Table II.

Dataset annotation is a very time consuming process that takes much time than usual. As we know that the fish postures slightly and haphazardly change due to their free and multiple dimensional rotations, so the bounding box labeling inserts with much care and accurate for mAP. Fish move freely, so we need to insert bounding box labeling in each direction for precise detection. We use a labeling tool for dataset annotation, Labelimg.

TABLE. II. The Number of Training and Testing Dataset of Fish Species

Class	Training images	Testing Images	Total Images
Anemone Fish	950	200	1150
Jelly Fish	1005	200	1205
Star Fish	1100	200	1300
Shark	950	200	1150

V. Results and Comparison

The experiment performed by the DL open-source library TensorFlow 1.11, OpenCV 4.1.1, and coding concluded with the high-level language python 3.5 at Ubuntu 18.04 operating system. Training and testing performed on the system intel core i-7-7700, GPU GTX 1080 with 12 GB of memory. Libraries, packages, and hardware specifications are shown in Table III.

We used the MS-COCO dataset for restoring and initialization of darknet-53 backbone architecture for Fish detection tasks. We set the resolution of the image is 608×608 during training the model. At the training stage, the initial and end learning rate set to 1e-4 and 1e-6, respectively, Intersection over Union (IOU) threshold value 0.5, average decay 0.9, and the batch size is 4. We trained our model to 100 epochs. To prevent the model from non-convergence, the learning rate during the training process changed gradually. The hyperparameters showed in Table IV.

In the experiment, we used custom fish detection dataset that consists of 4 classes, such as anemone-fish, star-fish, jelly-fish, and shark. The total number of training images is 4005, and images for testing are 800. The mAP of the proposed model increased, with improved detection scale, k-means++ clustering, loss function, and transfer learning technique of improved YOLOv3, by 4.13% compared to that of baseline YOLOv3, and the detection speed is 39 FPS, which enables real-time detection of YOLOv3. Some state-of-the-art architectures and detectors were choosing for comparisons such as Faster RCNN and YOLOv2 with our improved YOLOv3 model. The mAP with input image sizes of diverse network structures is shown in Table V, and the AP% comparison of YOLOv3 and improved YOLOv3 with our custom dataset and brackish dataset [21] is shown in Table VI.

TABLE. III. Software and Libraries

Tensor flow	1.12
OpenCV	4.1.1
Python	3.6.5
Matplotlib	3.1.2
Numpy	1.16.4
System	Intel Core i7-7700
CPU	3.6 Ghz
GPU	GeForce GTX 1080 Ti Memory 11 GB
CUDA	9.2, 10.0
cuDNN	7.6.0

TABLE. IV. The Hyperparameters

Parameters	Values
Initial learning rate	1e-4
End learning rate	1e-7
Total epochs	100
Warm-up epochs	2
1^{st} stage epochs	50
2nd stage epochs	50
Batch size	4
Image train size	608×608
IOU threshold value	0.5
average decay	0.9995
Gradient optimizer	Adam optimizer
Train mode	GPU

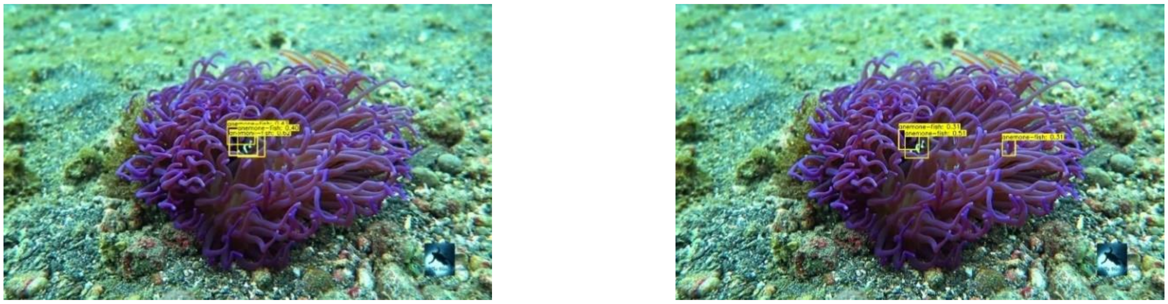
TABLE. V. YOLOv3 Comparison with others object Detector Models

Detection Model	Faster R-CNN	YOLOv2	YOLOv3	YOLOv3 Improved
Input Image Size	480	416	608	608
mAP	77.4%	81.63%	87.17%	91.30%

TABLE. VI. AP (%) of different Fish Dataset Species Comparison between YOLOv3 and Improved YOLOv3

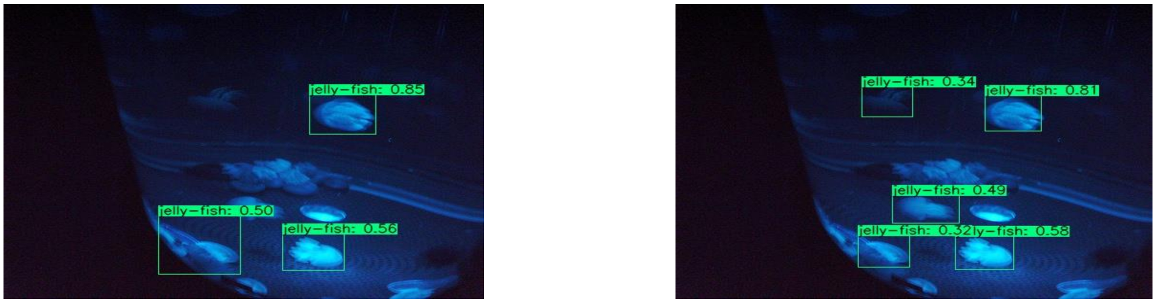
Model	Anemone-fish AP%	Jelly-fish AP%	Star-fish AP%	Shark AP%	mAP
YOLOv3	83.63%	88.21%	88.97%	87.89%	87.17%
YOLOv3 (brackish dataset)	89.99%	82.05%	93.67%		82.17%
YOLOv3 Improved	94.42%	86.14%	98.27%	86.35%	91.30%

The brackish dataset is a publically open dataset that collected from turbid water. Due to its turbidity, small size, the mAP evaluation on this dataset is comparatively less than our custom dataset. The brackish dataset contains 6 classes. We choose 3 classes among them, such as anemone-fish, jelly-fish, and star-fish to check the mAP and AP% on each class at the YOLOv3 detection model. The visualization comparison results between YOLOv3 and improved YOLOv3 are illustrating in Fig. 3, Fig. 4, Fig. 5 and Fig. 6. We draw the curves of model learning loss, confidence loss, and probability loss in Fig. 7, Fig. 8 and Fig. 9. The confidence loss curve of the trained model expresses the object confidence loss at each iteration, which is gradually improving after every iteration. The probability loss curve expresses the probability of an object, either the object belongs to anemone-fish or star-fish. The total loss curve expresses the feature extraction ability of model and model convergence.



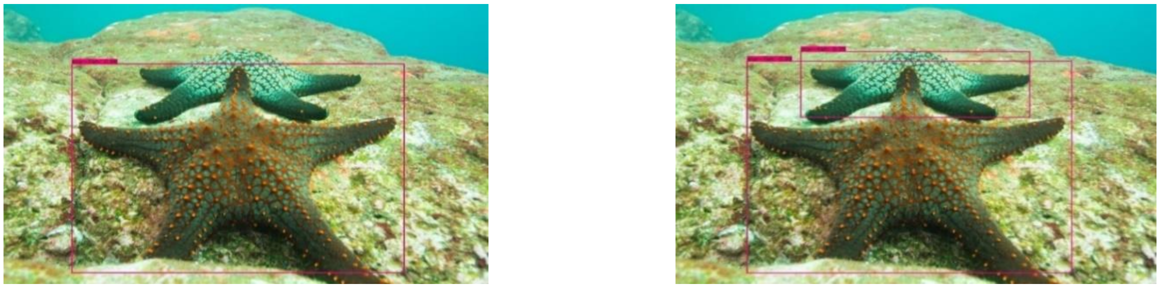
(a) (b)

Fig. 3. (a) Anemone Fish Result of Original YOLOv3, (b) Anemone Fish Result of Improved YOLOv3.



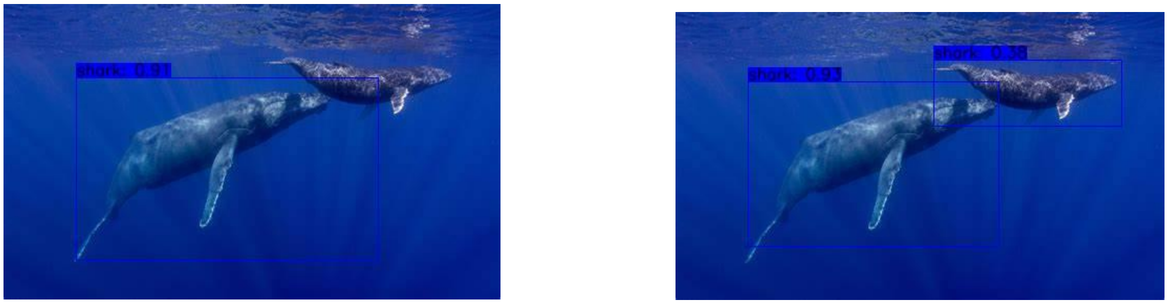
(a) (b)

Fig. 4. (a) Jelly Fish Result of Original YOLOv3, (b) Jelly Fish Result of Improved YOLOv3.



(a) (b)

Fig. 5. (a) Star-Fish Result of Original YOLOv3, (b) Star-Fish Result of Improved YOLOv3.



(a) (b)

Fig. 6. (a) Star-Fish Result of Original YOLOv3, (b) Shark Result of Improved YOLOv3.

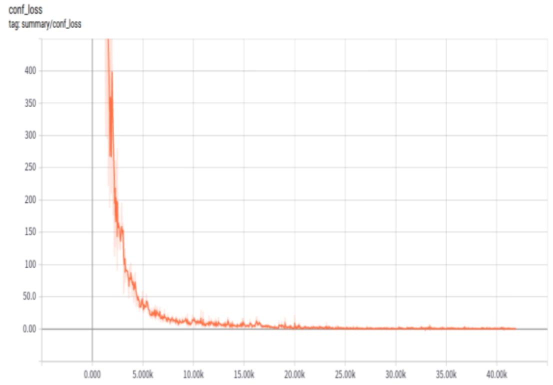


Fig. 7. Confidence Loss Curve of the Trained Model.

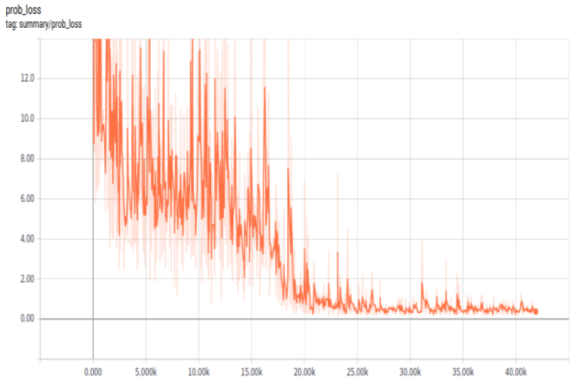


Fig. 8. Probability Loss Curve of the Trained Model.

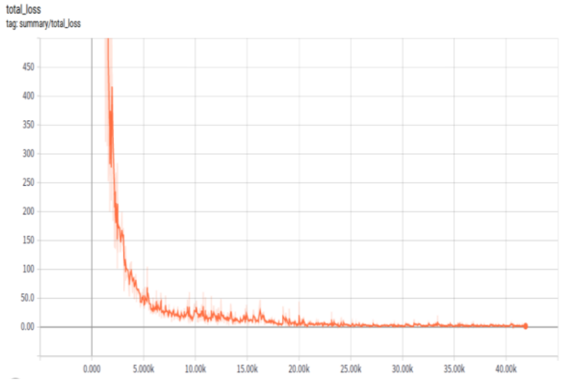


Fig. 9. Total Loss Curve of the Trained Model.

A. *Evaluation Matrices*

Intersection over union (IOU) and precision, recall are important metrics for model evaluation. IOU is the difference between ground truth and the predicted value which is defined as.

$$IOU=\frac{B\cap C}{B\cup C} \quad (5)$$

Where B is the ground truth value of an object, and C is the predicted value. From the results, it can be clear that the IOU of small and medium size objects is improved by the improved YOLOv3 model than the baseline YOLOv3 model with the addition of a 4th detection scale. Because it has the ability to extract finer grained features of small objects. In summary, the IOU value of the improved YOLOv3 model has greatly improved and better compared to the baseline YOLOv3 detection model. The IOU values of original YOLOv3 and improved YOLOv3 of various objects are shown in Fig. 10.

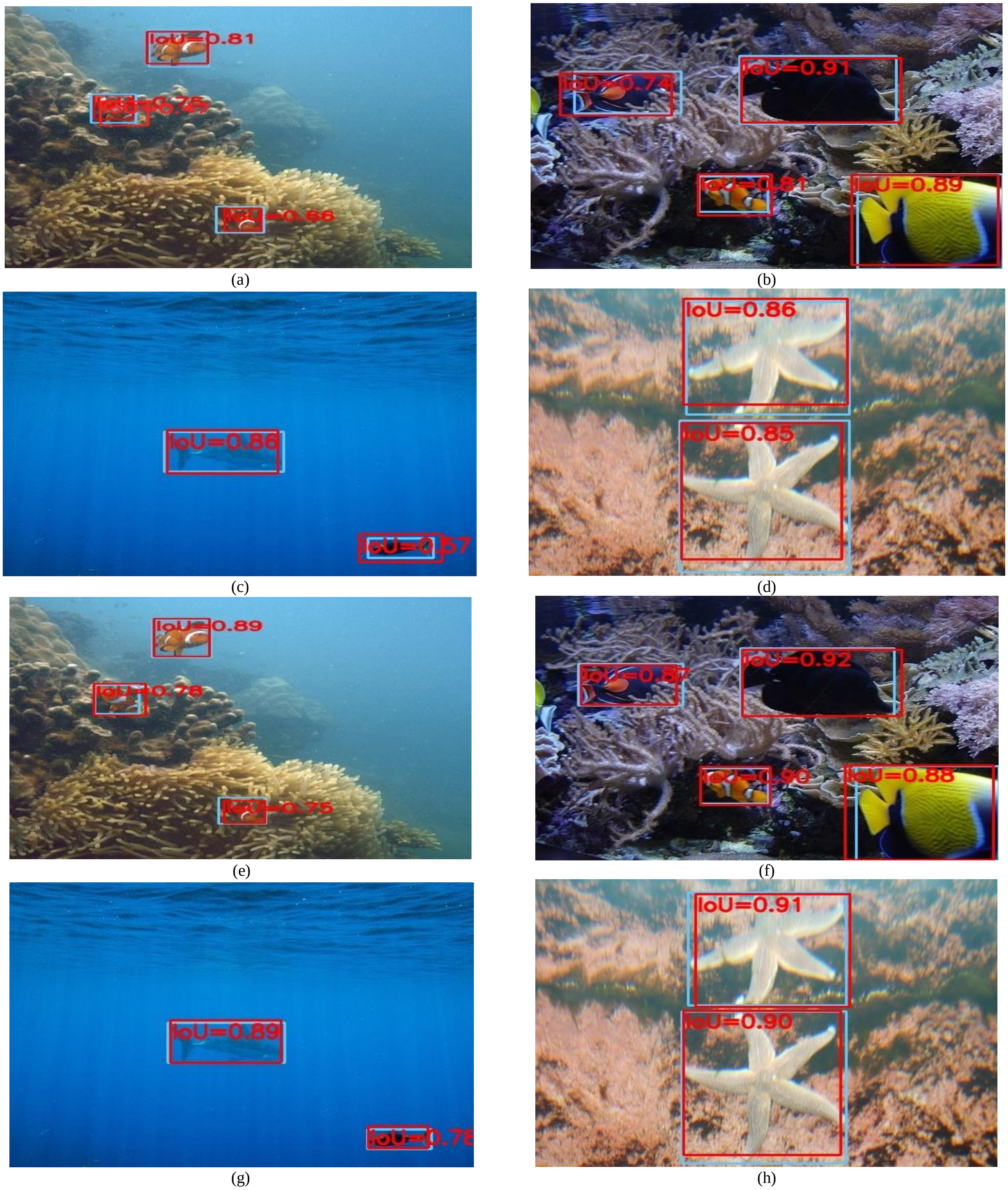
The precision (P) and recall (R) have been calculated on the basis of true positive (TP) false positvie (FP) and false negative (FN). The precision and recall are defined in equations (6) and (7).

$$\mathrm{P}=\frac{\mathrm{TP}}{\mathrm{FP+TP}} \quad (6)$$

$$R=\frac{\mathrm{TP}}{\mathrm{FN+TP}} \quad (7)$$

Where TP is the detection of an object correctly with a positive sample, and FP is the detection of an object negatively by the mistake of a positive sample. FN is not detected of an object with a positve sample.

The trade-off between precision-recall is a complicated problem. The precision-recall is one of the significant measures to evaluate the network performance at the testing dataset. In addition, precision is measured with respect to relevancy in results, while recall measures the total number of true, relevant results. The precision-recall curve expresses in the y-axis and x-axis, respectively. The precision-recall curve showed the trade-off at fixed IOU thresholding value 0.5. It is clear from the results with higher precision the recall rate also goes higher, which shows that our model is efficient and converges well. We noticed that the improved YOLOv3 model precision-recall of anemone-fish and star fish is much better than the baseline YOLOv3 model because these two classes have small and extra small objects. In the case of the jelly-fish, the precision-recall curve has not a big difference, and the shark precision-recall curve has been decreased because the size of the object of the shark class is comparatively big than other classes. The precision-recall curves of all fish classes, both YOLOv3 and improved YOLOv3, are shown in Fig. 11.



(a)

(b)

(c)

(d)

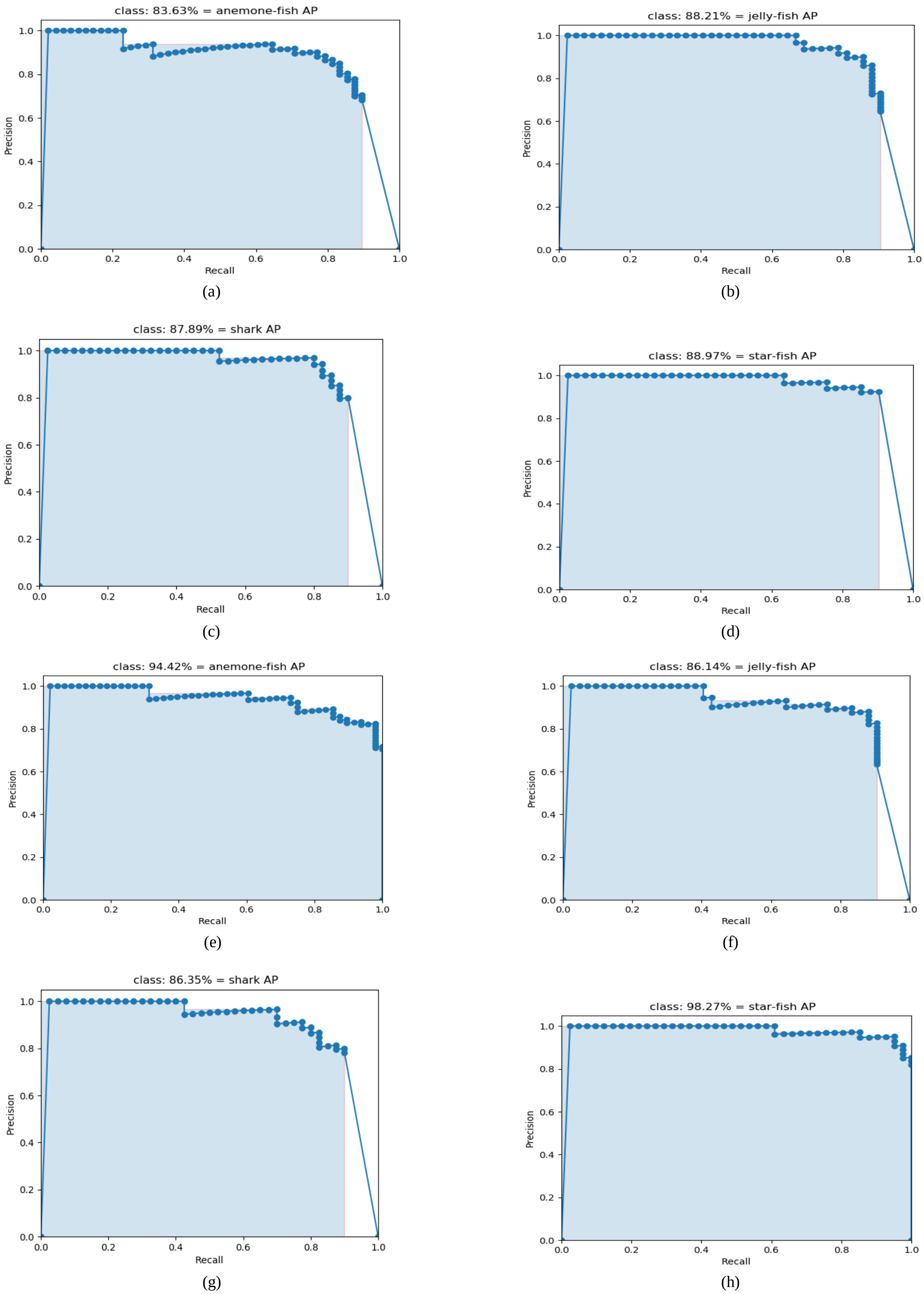
(e)

(f)

(g)

(h)

Fig. 10. (a,b,c,d) IOU Values of Original YOLOv3 and in Fig 10. (e,f,g,h) IOU Values of Improved YOLOv3.



class: 83.63% = anemone-fish AP

(a)

class: 88.21% = jelly-fish AP

(b)

class: 87.89% = shark AP

(c)

class: 88.97% = star-fish AP

(d)

class: 94.42% = anemone-fish AP

(e)

class: 86.14% = jelly-fish AP

(f)

class: 86.35% = shark AP

(g)

class: 98.27% = star-fish AP

(h)

Fig. 11. (a,b,c,d) AP% Values of Original YOLOv3 and in Fig 11. (e,f,g,h) AP% Values of Improved YOLOv3.

VI. Conclusion and Future Work

Mainly, we introduced how DL could be beneficial for the underwater species analysis at a large-scale dataset. The detection results showed how DL could be achieved excellent results for fish detection. In this paper, we improved YOLOv3 for fish detection. To obtain better results, we increased the detection scale to detect very small size objects. Apply k-means++ clustering to get suitable clusters, as well as transfer learning and improvement in the loss function. The improved YOLOv3 model proves that it outperforms than that of the baseline YOLOv3 model by improving the mAP of 4.13%.

In future work, we will collect large and live datasets, both images and video formats from different underwater conditions. We will improve this model by changing in backbone architecture to make it lightweight architecture and move this model on the embedded system, portable devices for live underwater marine animal detection.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning.” In: Nature 521.7553 (2015), pages 436–444 (cited on pages 21, 126, 128).

[2] Szegedy, C., Toshev, A., & Erhan, D. (2013). Deep neural networks for object detection. In Advances in neural information processing systems (pp. 2553-2561).

[3] Huang, Phoenix X., Bastiaan J. Boom, and Robert B. Fisher. "Hierarchical classification for live fish recognition." In BMVC student workshop paper. 2012.

[4] Bermejo S. (2007). “Fish age classification based on length, weight, sex, and otolith morphological features.”Fish. Res. 84.

[5] R. Larsen, H. Olafsdottir, B.K. Ersbøll, Shape and texture based classification of fish species, Image Anal., 2009, 745-749.

[6] Helge Balk, Development of hydro acoustic methods for fish detection in shallow water, 2001, pg 28.

[7] Fuming Xiang, Application of Deep Learning to Fish Recognition, 2018, pg 53.

[8] River catch of salmon, sea trout and migratory char”, 2019, Available: https://www.ssb.no/en/elvefiske. Accessed on: Dec. 10, 2019.

[9] D.H Hubel and T.N Wiesel. Receptive Fields, Binocular Interaction, and Functional Architecture in the Cat's Visual Cortex. Pages 151-152, 1961.

[10] Adrian Reithaug, Employing Deep Learning for Fish Recognition, 2018, pg 85.

[11] Ekaterina Lantsova, Automatic Recognition of Fish from Video Sequence, 2015, pg 49.

[12] S.O. Ogunlana, O. Olabode , S.A. A. Oluwadare & G. B. Iwasokun, Fish Classification Using Support Vector Machine, 2015, pg 75.

[13] Dhruv Rathi, Sushant Jain, Dr. S. Indu, Underwater Fish Species Classification using Convolutional Neural Network and Deep Learning,

[14] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[15] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[16] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[17] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[18] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[19] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).

[20] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[21] Pedersen, M. (Creator), Haurum, J. B. (Creator), Gade, R. (Creator), Moeslund, T. B. (Creator), Madsen, N. (Creator) (2019). The Brackish Dataset. Kaggle.

Drying Process Simulation Methodology based on Chemical Kinetics Laws

Vladimir M. Arapov[1], Dmitriy A. Kazartsev[2], Igor A. Nikitin[3], Maria V. Babaeva[4], Svetlana V. Zhukovskaya[5]
Svetlana N. Tefikova[6], Galina V. Posnova[7], Igor V. Zavalishin[8]
Voronezh State University of Engineering Technologies, Voronezh, Russian Federation[1]
K.G. Razumovsky Moscow State University of Technologies and Management
(The First Cossack University) Moscow, Russian Federation[2, 3, 4, 5, 6, 7, 8]

***Abstract*—It is shown that the existing approaches to drying process modeling, based on a system of interconnected differential equations of heat and mass transfer or on statistical processing of drying process experimental data, have significant drawbacks. It greatly complicates the development of computer means for controlling production processes. During modeling it is proposed to consider drying process from the standpoint of physical chemistry as a quasi-topochemical heterogeneous reaction and perform mathematical modeling of this process based on the laws of chemical kinetics. The basic issues of methodology of drying process modeling based on the laws of chemical kinetics are reviewed: the study of the equation of drying rate during the removal of free and bound moisture; methods for determining composition of the aqueous fractions with different forms and energy of moisture in materials; methods of determination of an activation energy of moisture; the influence of the concentration of moisture and other process factors on the drying speed. The methodological approach considered in the article allows developing reliable mathematical models of drying kinetics for the purposes of computer technologies for managing production processes and avoiding the errors that the authors note in previously published works**.

Keywords—Modeling; drying; chemical kinetics; activation energy

I. Introduction

The current stage of industrial development is characterized by the transition to computer technology for managing production processes. The pace of this development largely depends on the level of mathematical modeling of technological processes - the completeness and accuracy of reflection by a mathematical model of all the phenomena accompanying the technological process. Computer technology allows not only to increase the culture of production and labor productivity, but also to optimize energy and material costs for production [1, 2]. Therefore, the development of reliable mathematical models of modern technological processes is an urgent task for many industries. In particular, this problem is relevant for the drying process.

When modeling drying process, it is necessary to consider a number of complex interrelated phenomena [3]. The diversity of these phenomena and their interrelation cause the process to be considered from different points of view. From the standpoint of the classical theory of drying, this process represents a set of simultaneously occurring phenomena of heat transfer and mass transfer [4, 5, 6]: the transfer of heat from gas to the material through a boundary layer of gas; the evaporation of moisture from the surface of the material; the movement of moisture vapor from the material surface in the environment of the drying agent through the boundary layer of gas; transfer of heat and moisture inside the material.

Most products consist of three phases: solid, liquid, and gaseous. The mass of the gaseous phase is very small. Therefore, it is believed that a wet body consists of absolutely dry material and moisture. When interacting with the solid phase, moisture molecules form a bond of a certain strength. The drying process is accompanied by a violation of this connection, which requires a certain amount of energy. The binding energy of moisture to the material depends on the nature of the formation of bond forms, which may have a chemical, physical-chemical or physical-mechanical origin. The connection of moisture with the material is significantly affected by the phenomenon of adsorption. The mechanism of physical adsorption is very complex. Different areas of the surface have different adsorption potentials, and there is no strictly justified complete theory of this phenomenon, which significantly complicates the analytical determination of the energy of the connection of moisture with the material and mathematical modeling of drying.

The mechanism of moisture transfer in capillary-porous bodies, which many products belong to, is very complex and includes the phenomena of convective filtration transfer in macro-and micro-capillaries, the phenomena of diffusion, effusion, thermodiffusion, and others [5]. But even in macrocapillaries, the molar and diffusive transfer of moisture is of different nature, so mathematical modeling of the total flow of moisture, taking into account the totality of all phenomena, is a difficult task.

In the process of thermal drying a gradient of temperature and moisture content occurs in the material. This causes the movement of moisture in the material both due to the temperature gradient (the phenomenon of thermal conductivity or thermal diffusion) and due to the humidity gradient (the phenomenon of moisture conductivity or concentration diffusion) [3, 4]. Due to the humidity gradient, moisture moves to the evaporation surface, and under the influence of the temperature gradient, it can move from the surface layers to the inside of the material. Thermal and moisture conductivity is a complex physical phenomenon that consists of a number of processes and its mathematical description requires a large amount of information about the properties of the product as an object of drying. Thus, the phenomena of mass transfer during heat drying are not only complex in their thematic description, but also have a contradictory character.

Heat transfer and mass transfer phenomena are no less complex during heat drying [3, 4, 5, 6]. Thus, in general, it is not possible to obtain an accurate analytical solution for calculating heat transfer and mass transfer of heat drying, since for a mathematical description of all phenomena, it is necessary to know a large amount of information about the properties and parameters of the drying agent and product, which may not change deterministically.

Possible approaches to modeling drying are analyzed below. A relatively new approach to modeling based on the laws of chemical kinetics is considered. A critical analysis of published works in this area is given, and the methodology for modeling drying based on the laws of chemical kinetics is briefly described.

II. Analytical Review

Over the past decades scientific provisions on drying have been continuously developed. A large number of mathematical models have been published for various drying methods, most of which are based either on a system of differential interrelated equations of mass and energy transfer [6, 13], or on statistical processing of experimental research data for a specific product [3, 7, 8, 9, 10, 11, 12].

Sufficient information about the main directions of this development can be found at some practical conferences [6]. Modern mathematical models of drying are computer models, since their application to calculation and management of the process is carried out using a computer due to the complexity. However despite the use of computer technology, models in the form of a system of differential equations are difficult for practical use since they contain a large number of coefficients and characteristics of the properties of the processed product, these coefficients for each product and drying mode are often impossible to predict [3, 4]. The value of the coefficients depends on many random factors in preceding drying production stages. All this significantly reduces the reliability of these models. The development of mathematical modeling of drying processes with these approaches is hindered by a lack of knowledge about the mechanism of interaction of phases, as well as about the properties of the product that change during the drying process. Further development of drying modeling, as the author rightly notes [13], is possible on the application of achievements in the field of physical chemistry.

As is known from the theory of similarity and modeling, the application of models based on experimental studies is limited by the range of the study. The accuracy and reliability of these models is not high. These circumstances determine the relevance of the search for new approaches to drying process modeling and obtaining more reliable mathematical models of process kinetics.

III. Physicochemical Approach to Drying Process Modeling

From our point of view very promising in this direction is a physicochemical approach based on the idea of drying as a quasi-topochemical heterogeneous reaction. Here a dry residue and a vaporous phase, that has transferred to a drying agent, appear from the initial product as a result of physicochemical and phase transformations. Such a representation of drying allows applying the laws of kinetics of topochemical reactions of heterogeneous processes to mathematical modeling. First of all, two fundamental provisions of formal chemical kinetics: the law of acting masses and the Arrhenius kinetic equation are considered [14].

The law of acting masses is based on the assumption that the reaction rate depends on the number of collisions of reacting molecules, which is proportional to the product of the concentrations of reacting substances in degrees, whose indices correspond to stoichiometric coefficients of the chemical reaction equation.

The Arrhenius equation allows us to determine the value of the reaction rate constant k, and in practical application it usually has the form:

$$k = A\exp\left(-\frac{E}{RT}\right), \tag{1}$$

where A is the preexponential factor, depending on the physicochemical properties of the reaction system, E is the activation energy, T is the absolute reaction temperature.

One of the first such works [15] some advantages of the physicochemical approach to drying are shown. The main advantage of this approach is to establish an explicit form of the drying rate equation, which does not require complex mathematical calculations for practical use in engineering calculations. Further studies [16, 17, 18] confirmed the prospects of this approach to mathematical modeling of drying. An indisputable advantage of this approach is the possibility of using modern high-precision thermal analysis instruments for the numerical determination of process characteristics that are part of the mathematical model of the drying process, for example, the E value.

However, in some recent scientific papers [19, 20, 21, 22] based on the physicochemical approach to the kinetics of drying process modeling, some research results, in our opinion, significantly contradict the generally accepted provisions of the theory of drying. So in [19], the author distinguishes two stages of the process of free water removal during wood drying. In this case, the activation energy of free water molecules at individual stages differs almost twofold at a sample temperature of 295 ... 296 K, the E/R value is 55932 ... 32608 K, which corresponds to an activation energy of 465019 ... 271103 kJ/kmol, and at a temperature of 303 ... 298 K, the E/R value is 34060 ... 62035 K, which corresponds to an activation energy of 283175 ... 515759 kJ/kmol. In addition, the indicated values of the activation energy exceed the enthalpy of saturated water vapor at given temperatures by an order of magnitude (the enthalpy of saturated water vapor at a temperature of 303.15 K is 46008 kJ / kmol), which cannot but raise doubts about the reliability of the author's results [19]. And according to [10], the activation energy for drying mushrooms in the temperature range 65 ... 85 ° C is 22 kJ / mol. These facts indicate that there is no scientifically based method for determining the activation energy of water in relation to drying processes.

In [19, 20] the question of the influence of relative air humidity on the value of activation energy is considered. But the relative humidity of the air, as is known [3], has an effect on the external heat and mass transfer, when water molecules have already passed into a vaporous state. It is no coincidence that in the graphs of the dependence of the activation energy on the relative air humidity [19, 20] it is difficult to distinguish the indicated functional dependence due to the significant scatter of the experimental data, and the correlation estimate is not given to the corresponding equations.

A number of authors believe that the destruction of the moisture bond with the dry part of the product occurs in “their (different) temperature ranges” [19]. This means that the removal of water fractions with different binding energies during drying should occur in a strictly defined temperature range. But from the practice of drying, it is known that bound moisture is removed from the product even at relatively low temperatures, if the air is dried. We believe that the contradictory information available in the scientific literature on the issue under consideration indicates the absence of a reliable methodological approach to the application of laws of chemical kinetics to the analysis and modeling of drying processes. In this regard, the goal of this work is to substantiate the main approaches to drying processes modeling based on the laws of chemical kinetics. The object of research is the kinetics of products thermal drying. The subject of the study is the methodology of applying the laws of chemical kinetics to the modeling of drying processes.

IV. Methodology of a Physicochemical Approach To Drying Process Modeling

First of all, in the general case, it should be considered that the moisture in the material has a many fractional composition. Water fractions are distinguished by the binding energy to the dry part of the product. To determine the boundaries of each water fraction, appropriate studies should be carried out. The most complete and reliable information on the fractional composition of water in materials is provided by methods based on nuclear magnetic resonance [23] and thermal analysis methods [14, 24, 25]. A quantitative assessment of the bond strength of moisture with the dry part of the product can be obtained by the method of [26].

The drying rate of each water fraction based on the law of masses and the Arrhenius equation can be represented as:

$$\frac{d\alpha_i}{d\tau} = f_i(\alpha_i) A_i \exp\left(-\frac{E_i}{RT}\right), \qquad (2)$$

where $\alpha_i = \frac{U_i^{\text{н}} - U_i}{U_i^{\text{н}} - U_i^{p}}$ – the degree of conversion of the *i*-th aqueous fraction;

$f_i(\alpha_i)$– function of the degree of conversion of the *i*-th aqueous fraction;

A_i – preexponential factor of the *i*-th water fraction, sec^{-1};

E_i – activation energy of the *i*-th aqueous fraction, J / (mol K);

T –the absolute temperature of the *i*-th aqueous fraction equal to the temperature of the material, K.

$U_i^{\text{н}}$, U_i^{p}, U_i –the moisture content of the material, calculated as the ratio of the amount of the *i*-th aqueous fraction to the amount of the dry part of the product, respectively, the initial state, equilibrium state and the state considered at a given moment in time.

In detailing equation (2), it is necessary to determine the sequence of removal of water fractions based on information on the forms of moisture bonding in a material. Can the removal of water fractions be considered as parallel independent chemical reactions, or as sequential reactions, or is there some other order possible? If water clusters with different binding energies can be in different product areas, then the option of parallel removal of water fractions is possible. In this case, the equation for the drying rate will be:

$$\frac{d\propto}{d\tau} = \sum f_i\,(\alpha_i) A_i \exp\left(-\frac{E_i}{RT}\right) \qquad (3)$$

If moisture is retained layer by layer by adsorption forces (moisture of monoadsorption and polyadsorption layers) and differs significantly in different layers in terms of activation energy, then at moderate drying temperatures, a model of successive removal of fractions is likely. If there is a film of free moisture on the surface of the plant cell, then the simultaneous removal of film and intracellular moisture is unlikely. Deletion of these fractions should be considered sequentially. In this case, the drying rate will be determined by the removal rate of the next aqueous fraction currently being considered:

$$\frac{d\propto}{d\tau} = f_i(\alpha_i) A_i \exp\left(-\frac{E_i}{RT}\right) \qquad (4)$$

More complex options are also possible. For example, when drying casein [15], protein denaturation and the transition of bound moisture to a free state are possible.

Difficulties arise in determining the explicit form of the function of the degree of transformation$f_i(\alpha_i)$. For this, it is necessary to compare the studied drying method with the already known topochemical reactions [14] and select the closest option. Otherwise, you can take the option that is often encountered in practice:

$$f_i(\alpha_i) = (1 - \alpha_i)^n, \qquad (5)$$

where *n* is a value showing the order of the reaction.

When choosing the type $f_i(\alpha_i)$ for free water, it is necessary to take into account the position known in the theory of drying: with constant heat supply to the material, the rate of free water removal is constant and does not depend on moisture content [3]. Therefore, for this fraction, the reaction order should be taken equal to zero. In addition, during this drying period, the temperature of the product is comparable to the temperature of the wet thermometer under these drying conditions [3]. Therefore, when free water is removed, the mathematical model of the drying rate has the form:

$$\frac{d\alpha}{d\tau} = A_{\text{св}} \exp\left(-\frac{E}{RT_{\text{М}}}\right) \quad (6)$$

where $A_{\text{св}}$ - preexponential factor when removing free moisture, sec^{-1};

$T_{\text{М}}$ - absolute temperature of the drying agent by wet thermometer, K.

If we assume that for the transition of a free water molecule to a vapor state, additional energy is required equal to only the latent heat of vaporization *r*, then in this case we can take $E = r$.

Unfortunately, this was not taken into account in [19].

In our opinion, the order of magnitude of the activation energy can be estimated theoretically. Firstly, for a water molecule to pass from a liquid to a vapor state, it is necessary to break its connection with the dry part of the product, i.e. to give additional energy to the water molecule equal to (and possibly somewhat larger) the binding energy of moisture with the material $E_i^{\text{св}}$. For the transition of the molecule into a vaporous state, additional energy is required equal to the latent heat of vaporization. If the generated water vapor is not removed from the external surface of the product, but from the deepened evaporation zone [3], then additional energy will be needed to move the vapor through the dry layer of the product. During this movement, the initially formed saturated steam may overheat. Thus, the activation energy of the *i*-th aqueous fraction can be estimated as:

$$E_i = E_i^{\text{св}} + r + E_i^{\text{дв}}, \quad (7)$$

whereE_i – activation energy of the *i*-th aqueous fraction, J/mol;

$E_i^{\text{св}}$ – binding energy of the *i*-th moisture fraction with the material, J/mol;

r –latent heat of vaporization, J/mol;

$E_i^{\text{дв}}$ – energy required to move moisture to the outer surface of the product, J/mol.

Due to the fact that the heat of vaporization decreases with increasing temperature, then for $r \gg E_i^{\text{св}} и r \gg E_i^{\text{дв}}$, a decrease in the activation energy of the i-fraction is resulted with an increase in the temperature of the product and, as a consequence of this, an increase of the drying rate.

P.A. Rebinder proposed to determine the binding energy by the equation$E_{\text{св}}$.

$$E_{\text{св}} = -RTln[\varphi(U,T)], \quad (8)$$

where $\varphi(U,T)$ – the relative humidity of the air in equilibrium with the material having a moisture content of *U* and an absolute temperature of *T*;

T – absolute temperature of the product, K;

R – universal gas constant, J/(mol·K).

Equation (8) characterizes the binding energy as the free energy of the isothermal separation of one mole of water from the dry part of the product without changing the state of aggregation [27]. It shows that $E_{\text{св}}$ is a function of the moisture content and temperature of the product, and not the relative humidity of the air supplied to the dryer, since each value of the moisture content of the product at temperature T corresponds to a strictly defined value of the relative humidity of the air in the equilibrium system: air - product.

However, drying in an industrial apparatus is an irreversible and nonequilibrium process. Therefore, the actual energy consumption for breaking the bond of moisture with the material will probably be greater than the value calculated by equation (8). The value of $E_i^{\text{дв}}$ can be approximately determined as the energy cost of overcoming the hydraulic resistance during the movement of water vapor through capillaries of the dry layer of the product.

We believe that in order to establish the explicit form of mathematical models (4), (5), (6), the value of the activation energy should be determined experimentally on the basis of existing [14, 24] or the development of new thermal analysis methods.

The most difficult task is to determine the value of the preexponential factor A_i, since the kinetics of topochemical reactions (as well as during drying) are significantly affected by geometric factors of shape, crystal structure, interface and relative phase surface velocity, and others [2]. Therefore, without taking these factors into account, the results of determining the value of A_i, for example, obtained using thermal analysis devices [19, 20] cannot be transferred to mathematical models of industrial processes. In its physical meaning, the coefficient $A_i$represents the maximum value of the drying rate at a very high temperature, when neither the activation energy nor the moisture content influence on the speed process (*f(α)*=1; $\exp\left(-\frac{E}{RT}\right) = 1$). In this regard, the determination of the value of A_i is a separate task of scientific research. In the simplest case, information on the coefficient A_i can be obtained experimentally, as shown in [5].

Thus, the development of a mathematical model of drying based on the laws of chemical kinetics should be performed in the following sequence:

1) The fractional composition of moisture in the product is investigated and the boundaries of water fractions are determined.

2) The activation energy of each water fraction is determined.

3) Based on information about the forms of moisture coupling in the material, the sequence of removal of water fractions (sequential or parallel) and the type of kinetic equation for periods of constant and decreasing drying speed are determined.

4) The type of function of the degree of water fractions transformation is determined.

5) We conduct a theoretical or experimental study to determine the value of the pre-exponential multiplier of the kinetic equation.

6) The resulting drying model is examined for suitability for practical use.

V. Conclusion

Drying process can be considered from the standpoint of physical chemistry as a quasitopochemical heterogeneous reaction. Mathematical modeling of the process can be based on application of laws of chemical kinetics.

The type of such a quasitopochemical heterogeneous reaction (sequential or parallel removal of water fractions) is determined on the basis of studying the forms and energy of water binding to the dry part of the product. The boundaries of aqueous fractions, which differ in the binding energy of moisture, can be reliably determined by nuclear magnetic resonance and thermal analysis.

The value of activation energy can be defined as the sum of the binding energy of moisture with the dry part of the product, the latent heat of vaporization and the energy needed to move water vapor through the capillaries in the dry layer of the product. Activation energy is a function of the temperature and moisture content of the product. The value of the activation energy can be determined by the thermal analysis method.

To establish the dependence of pre-exponential coefficients of the Arrhenius equation on the geometric factors of the form, the relative motion of the phases and other factors, additional research is needed.

The considered methodology is applicable primarily to convective drying of dispersed products with an active hydrodynamic regime, when the evaporation zone can be considered from the point of view of chemical kinetics as a kinetic region for small particle sizes. Therefore, as the next research tasks, it is advisable to justify the type of kinetic equation and develop a simulation algorithm for other drying methods: drying with heat input by high frequency currents, infrared rays, etc. Further improvement of mathematical models of drying is possible on the basis of a detailed study of the transformation function of water fractions and the pre-exponential coefficients of the Arrhenius equation.

The implementation of the physicochemical approach to drying process modeling allows one to obtain simple and reliable mathematical models of drying kinetics for engineering practice and to avoid errors that the authors note in previously published works.

VI. Practical Implementation of Scientific Research Results

For the practical implementation and development of the idea of applying the laws of chemical kinetics to mathematical modeling of drying processes, various classification groups of products should be studied according to the indicated methodology: colloidal, capillary-porous, and others. Also it is also needed to establish the kinetic laws of drying explicitly for each group of products and, on this basis, to develop algorithms for computer technology for controlling production processes.

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] Patent of the Russian Federation No. 2444689, PMK F26B 25/22, C1. A method for automatic control of the drying process of food products in a belt dryer using convective and microwave energy supply / S.T. Antipov, D.A. Kazartsev, A.V. Zhuravlev, T.V. Kalinina, I.S. Yurova, A.B. Emelyanov. - Application: No. 2010135851/06; 08/26/2010, publ. 03/10/2012. Bull. Number 7.

[2] Patent of the Russian Federation No. 2547345, PMK F26B 25/22, C1. A method for automatic control of the drying process of dispersed materials in a swirling coolant flow with microwave energy supply / D.A. Kazartsev, S.T. Antipov, A.V. Zhuravlev, D.A. Nesterov, A.V. Bo-Rodkina, S.A. Vinichenko. - Application: No. 2013156470/06; 12/19/2013, publ. 04/10/2015. Bull. Number 10.

[3] Sazhin, B.S. Scientific principles of drying technology / B.S. Sazhin, V.B. Sazhin. - M .: Nauka, 1997 .- 448 p.

[4] Rudobashta S. P. Using the theoretical propositions of academician A.V. Lykov in modern models of heat and mass transfer during drying / Actual problems of drying and heat-humidity processing of materials in various industries and agro-industrial complex / / collection of scientific articles of the First international Lykov scientific readings. Kursk, 2015. P. 21 – 28.

[5] Fedosov S. V. Heat and mass transfer in technological processes of the construction industry. Ivanovo: IPK "Pressto". 2010. – 364 p.

[6] Rudobashta S. P. New Russian studies in the field of drying and thermal-moisture processes / / Trudy 3 mezhdunarod. science. practical. Conf. "Modern energy-saving thermal technologies (Drying and thermal processes SETT-2008)". Vol. 1. Moscow, 2008. - Pp. 5-18.

[7] Antipov, S.T. Development of a mathematical model of the drying process of blackcurrant fruits in a vacuum apparatus with microwave energy supply / S.T. Antipov, D.A. Kazartsev, A.V. Zhuravlev, S.A. Vinichenko // Bulletin of the Voronezh State University of Engineering Technologies - Voronezh: Publishing House of the Voronezh State. University of Technology, 2014. - P. 7–12.

[8] Yurova, I.S., Heat and mass transfer during drying of milk thistle seeds in a vortex chamber with microwave energy supply: Monograph / I.S. Yurova, I.T. Kretov, A.V. Zhuravlev, D.A. Kazartsev. - Voronezh: Publishing house of the Voronezh State University of Engineering Technology, 2012. - 192 p.

[9] CFD simulation of spray dryers / M.W. Woo, L.X. Huang, A.S. Mujumdar, W.R.W. Daud ; Ed.M.W. Woo, A.S. Mujumdar, W.R.W. Daud. – Singapore, 2010. – V. 1. – P. 1 – 36. – ISBN 978-981- 08-6270-1.

[10] Yoon, S.S. Lagrangian-based stochastic dilute spray modelling for drying applications / S.S. Yoon ; Ed.M.W. Woo, A.S. Mujumdar, W.R.W. Daud. – Singapore, 2010. – V. 1. – P. 77 – 112. – ISBN 978-981- 08-6270-1.

[11] Julklang, W. Analysis of Slurry Drying in a Spray Dryer / W. Julklang, B. Golman // International Journal of Engineering and Technology (IJET). – Dec 2013 – Jan 2014. – V. 5. – N. 6. – P. 5178 – 5189. – DOI:10.1007/s11671-010-9793-9.

[12] Handbook of Industrial Drying / Fourth Edition Edited by Arun S. Mujumdar // CRC Press Taylor & Francis Group. – 2015. – P. 1334.

[13] Dornyak O. R. Modern problems of mathematical modeling of thermal-humidity processing of materials / collection of scientific articles of the First international Lykov scientific readings. Kursk. P. 36 - 42.

[14] Dyachenko, A.N. Chemical kinetics of heterogeneous processes: a training manual / / A.N. Dyachenko, V.V. Shagalov // Tomsk Polytechnic University. - Tomsk: Publishing House of Tomsk Polytechnic University, 2014. - 102 p.

[15] Arapov, V.M. Improving the drying of casein Spec. 05.18.12 - Processes and Food Production Equipment / Diss. for Ph.D. Moscow, 1985 .- 146 p.

[16] Drannikov, A.V. Study of the drying process of beet pulp with superheated steam Specialty: 05.18.12-Processes and food production apparatus / Diss. for Ph.D. Voronezh - 2003. - 164 p.

[17] Antipov, S.T., Heat and mass transfer during drying of coriander seeds in an apparatus with microwave energy supply: Monograph / S.T. Antipov, D.A. Kazartsev. - Voronezh: Publishing house of the Voronezh state. Technological Acad., 2007. - 142 p.

[18] Mamontov M.V. Development and research of drying finely chopped carrots during its complex processing/ Specialty: 05.18.12-Processes and food production apparatus / Diss. for Ph.D. Voronezh - 2009. - 184 p.

[19] Ermochenkov M.G. Kinetic parameters of the wood drying process // Lesn. journal 2017. No. 6. P. 114–125. (Izv. Higher education. Institutions).

[20] Kuvik T.Ye. Kinetics of evaporation of bound moisture and wood destruction during thermal modification / Diss. abstract for Ph.D. Moscow - 2013. - 20 p.

[21] Ermochenkov M.G. Internal sources of mass during wood drying / M.G. Ermochenkov // Materials of the VI B.N. Ugolev international symposium, dedicated to the 50th anniversary of the Regional Coordinating Council on Contemporary Problems of Wood Science (Krasnoyarsk, September 10–16, 2018) - Novosibirsk: Publishing House of the SB RAS, 2018.- P. 80 - 82.

[22] Kholmansky A.S., Tilov A.Z., Sorokina E.Yu. Physicochemical modeling of the drying process of vegetables and fruits / A.S. Kholmansky, A.Z. Tilov, E.Yu. Sorokina // Modern problems of science and education. - 2012. - No. 5; URL: http://www.science-education.ru.

[23] Patent of the Russian Federation No. 2204822, PMK G01N 24/00, C2. A method for determining the amount of monomolecular - adsorption and polymolecular - adsorption moisture / I.T. Kretov, V.M. Arapov, S.V. Shakhov. - Application: No. 2001123030/28, 15. 08. 2001; publ. 20. 05. 2003 Bull. No. 14.

[24] Patent of the Russian Federation No. 2296974, PMK G01N 15/00, C1. A method for determining the fractional composition of moisture in materials / V.M. Arapov, M.V. Mamontov, M.V. Arapov. - Application: No. 2005122966/28, 19. 07. 2005; publ. 10.04. 2007 Bul. No. 1.

[25] Patent of the Russian Federation No. 2312328, PMK G01N 25/56, C2. A method for determining the amount of aqueous fractions that differ in the binding energy of moisture with a substance / V.M. Arapov, S.V. Shakhov, M.V. Arapov, S.V. Buturlin. - Application: No. 2006100224/13, 10. 01. 2006; publ. 10. 12. 2007 Bull. Number 34.

[26] Patent of the Russian Federation No. 2230311, PMK G01N 25/56, C1 Method for determining the bond strength of moisture with a substance / V.M. Arapov, D.A. Kazartsev, M.V. Arapov. - Application: No. 2003103805/28, 10. 02. 2003; publ. 10.06. 2004 Bul. No. 16.

[27] Patent of the Russian Federation No. 2292018, PMK G01J 5/56, C1. A method for determining the binding energy of moisture with a substance / V. M. Arapov, M.V. Mamontov, M.V. Arapov. - Application: No. 2005121839/28; 07/11/2005, publ. 20. 01. 2007 Bull. No. 2.

Vision-based Indoor Localization Algorithm using Improved ResNet

Zeyad Farisi[1], Tian Lianfang[2], Li Xiangyang[3], Zhu Bin[4]
School of Automation Science and Engineering, South China University of Technology, Guangzhou, China[1, 3]
College of Community Service Department of Engineering and Science. Tabah University, Medinah, Saudi Arabia[1]
School of Automation Science and Engineering, South China University of Technology[2]
Research Institute of Modern Industrial Innovation, South China University of Technology[2]
Key Laboratory of Autonomous Systems and Network Control of Ministry of Education. Guangzhou, China[2]
School of Mechanical and Electronic Engineering, Jiangxi college of applied technology, Ganzhou, China[4]

***Abstract*—The output of the residual network fluctuates greatly with the change of the weight parameters, which greatly affects the performance of the residual network. For dealing with this problem, an improved residual network is proposed. Based on the classical residual network, batch normalization, adaptive -dropout random deactivation function and a new loss function are added into the proposed model. Batch normalization is applied to avoid vanishing/exploding gradients. -dropout is applied to increase the stability of the model, which we select different dropout method adaptively by adjusting parameter. The new loss function is composed by cross entropy loss function and center loss function to enhance the inter class dispersion and intra class aggregation. The proposed model is applied to the indoor positioning of mobile robot in the factory environment. The experimental results show that the algorithm can achieve high indoor positioning accuracy under the premise of small training dataset. In the real-time positioning experiment, the accuracy can reach 95.37.**

***Keywords*—Deep learning; residual network; loss function; dropout; indoor localization**

I. Introduction

With the development of artificial intelligence technology, various types of robots have been widely used. In the application of mobile robots, real-time detecting and monitoring the location of robots is the prerequisite for better service to human beings. For indoor localization assignments, Wi-Fi based method [1], Bluetooth based method [2] and Radio Frequency identification technology [3] were proposed and widely used. However, bottlenecks exist in these methods. Wi-Fi-based methods are vulnerable to multi-path effects, Bluetooth-based methods exist mutually interference and RF-based methods require expensive equipment support. Vision-based methods [4][5] which can realize real-time positioning only by a normal RGB camera, avoid all these bottlenecks mentioned above and provide a new way for indoor positioning.

In recent years, deep learning technology has been greatly developed and widely used in image processing, especially in image-based classification tasks. Compared to many traditional algorithms, deep learning technology, which uses massive training dataset to learn prior knowledge, has stronger generalization ability and more complex parametric expression. Since 2012 [6], Hinton et al put forward the outstanding performance of Alexnet with five convolution layers and three full connection layers in the ImageNet image classification competition. More and more scholars began to study convolution neural network to solve various practical problems. It was found that the accuracy can be improved by increasing the depth of CNN (Convolutional Neural Network). The deeper the network, the more features can be obtained, and the stronger the expression ability of the network. What's more, the deeper the network, the more abstract semantic features can be extracted [7-10]. However, simply increasing the number of layers of neural network will lead to the problems of gradient disappearance, gradient explosion and model degradation. In 2016, He et al proposed a 152 layer Res-Net [11], which the residual structure is used in the deep neural network. Res-Net can solve the degradation problem and the residual structure makes the model easier to optimize, and can get better training results under the premise of smaller training dataset, but the learning results of the network are very sensitive to the fluctuation of the network weight, that is, the slight change of the network weight will cause a greater change of the output. The model would be affected badly by this shortcoming in the process of model training and testing. In [12-16], a serious of improvements have been made to ResNet. But none of them can solve the problem well.

In order to solve the stability problem of the ResNet, an improved residual network is proposed. Based on the classical residual network, batch normalization, adaptive β -dropout random deactivation function and a new loss function are added into the proposed model. Batch normalization is applied to avoid vanishing/exploding gradients. β -dropout is applied to increase the stability of the model, which we select different dropout method adaptively by adjusting parameter β . The new loss function is composed by cross entropy loss function and center loss function to enhance the inter class dispersion and intra class aggregation. The proposed model is applied to the indoor positioning of mobile robot in the factory environment. The experimental results show that the algorithm can achieve high indoor positioning accuracy under the premise of small training dataset. In the real-time positioning experiment, the accuracy can reach 95.37.

II. The Improved Resnet

We use 50 layers residual network in our assignment, to enhance the performance of our model, batch normalization layer, β -dropout layer and improved loss function are added into our model. The structure of the improved ResNet is as follow Table I:

The residual structure is composed of image preprocessing convolutional layer conv1, convolutional blocks conv2_3, conv3_4, conv4_6 and conv5_3 and full connection layer conv6. Each block is composed of three convolution layers, they are duplicated 3 times, 4 times, 6 times and 3 times, respectively. Batch Normalization layers are placed in front and back of each block and residual structure is applied in each block. Between conv5_3 and conv6, Average pool is applied to extract deep image feature and β -Dropout is applied to simplify the network. After conv6, loss function is applied, weight parameters are adjusted by stochastic gradient descent (SGD) of loss function with back-propagation, mini batch size is 256. The learning rate is 0.1 at the beginning and is divided by 10 when the error rate stops falling. We have 18 localization centers, so the final result of the loss function is, and we select the biggest one in these 18 number.

The residual structure is shown in Fig. 1， where x is the input of the convolutional block, the output of the block is $H(x)=F(x)+x$. Compared to $F(x)$, Nonlinear function $F(x)=H(x)-x$ is more easier to be optimized. Branch x is sent to the next block directly, which can be studied easily. Under this structure, Back propagation is easier to go on.

A. Batch Normalization

Batch normalization is used to regulate the input into a reasonable scope, which can avoid the vanishing/exploding of gradients caused by the increase of the layer of deep neural network.

$$\tilde{x}=\frac{x-E(x)}{\sqrt{Var(x)+\varepsilon}} \tag{1}$$

TABLE. I. Improved Resnet

Layer name	The configuration of each layer	Output size
conv1	Ksize=(7,7), stride=2, filter =64 max pool , batch normalization	112*112*64
conv2_3 3 layers	[1*1, 64; 3*3, 64, 1*1, 128]*3 Batch normalization	56*56*64 56*56*256
Conv3_4 4 layers	[1*1, 128; 3*3, 128, 1*1, 512]*4 Batch normalization	28*28*512 28*28*512
conv4_6 6 layers	[1*1, 256; 3*3, 256, 1*1, 1024]*6 Batch normalization	14*14*1024 14*14*1024
conv5_3 3 layers	[1*1, 512; 3*3, 512, 1*1, 2048]*6 Batch normalization	7*7*2048 7*7*2048
Aver pool	Ksize =(7,7)，stride=7	1*1*2048
β -Dropout	β changes adaptively	1*1*2048
conv6	Ksize =(1,1)，filter =2048，stride=1	1*1*2048
Loss function	Two loss function combined	1*1*18

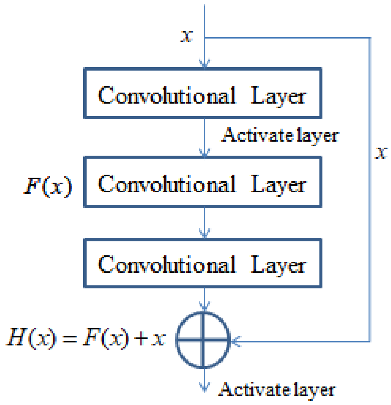


Fig. 1. Residual Structure.

Where x is the activate value of each node, $E(x)$ is the mean of one layer, $Var(x)$ is the variance. ε is a small number used to avoid denominator equal to zero. $\tilde{x}$ is the activate value after normalization.

B. Dropout

With the increase of the layer, depth neural network is easy to cause over-fitting, dropout [17-18] is a commonly used technology to alleviate this problem. The specific method is to discard a neural network node according to a certain probability in the training process of deep learning network, that is, to set the activate value of the node to zero. To enhance the stability of the model, β -dropout is applied. Adjusting the value of β adaptively, we can generate different kinds of distribution of dropout.

$$r^{(l)} \sim Beta(x:\beta,\beta) \tag{2}$$

$$\tilde{y}^{(l)}=r^{(l)}y^{(l)} \tag{3}$$

$$z_i^{(l+1)}=w_i^{(l+1)}\tilde{y}^{(l)}+b_i^{(l+1)} \tag{4}$$

$$y_i^{(l+1)}=f(z_i^{(l+1)}) \tag{5}$$

$$Beta(x;\alpha,\ \beta)=\frac{1}{B(\alpha,\ \beta)}x^{\alpha-1}(1-x)^{\beta\text{-}1} \tag{6}$$

$$Beta(\alpha,\ \beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{7}$$

$$\Gamma(\beta)=\int_0^{\infty} e^{-t}t^{\beta-1}dt \tag{8}$$

Where $r^{(l)}$ obey Beta distribution, $y^{(l)}$ is the activate value of l-th layer, $\tilde{y}^{(l)}$ is the output value of l-th layer, $w_i^{(l+1)}$ is the weight of $l+1$-th layer, $b_i^{(l+1)}$ is the bias of $l+1$-th layer. $z_i^{(l+1)}$ is the input of $l+1$-th layer, $f()$ is the activate function. $\Gamma(\alpha)$, $\Gamma(\beta)$ is Gamma function, we set $\alpha=\beta$, $Beta(\alpha, \beta)$ equal to $Beta(\beta)$, is a symmetry distribution. Adjusting parameter β, $Beta(x:\beta,\beta)$ can generate Bernoulli distribution, uniform distribution and Gaussian distribution. At the beginning of model training stage, the Bernoulli distribution is applied to delete some unimportant nodes, we set $\beta=0.001$; in the middle of model training stage, 0-1 uniform distribution is applied to smooth each node, we set $\beta=1$; at the end of model training stage, gauss distribution is applied to highlight important nodes, we set $\beta=3$ at this time.

C. Loss Function

In indoor localization algorithm, the image location features of the adjacent location points are similar, that is, the spacing between different classes is very small. In order to increase spacing between different classes and reduce spacing in one class, a loss function combined center loss and cross entropy loss is applied. The loss function can be described as follows:

$$L = L_s + \lambda L_c \tag{9}$$

Where L_c is the center loss function, L_s is the cross entropy loss function, λ is a weight used for balancing the two loss functions. The structure of our loss function is shown in Fig. 2.

The cross entropy loss function can be seen in [17]. We establish a class center in the feature space for each class. The center loss function is the sum of the distance between features of the sample and features of the class center in the feature space.

$$L_c = \frac{1}{2}\sum_{i=1}^{m} \| x_i - c_{y_i} \|_2^2 \tag{10}$$

L_c represents the center loss function, x_i is the depth feature of i-th class, c_{y_i} is the deep feature center of i-th class. m is the size of mini-batch.

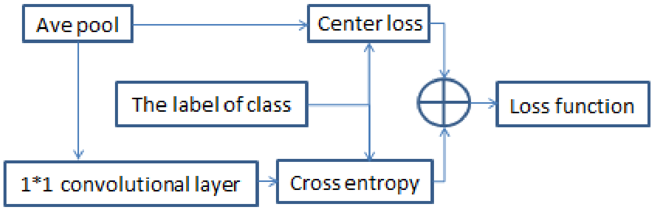


Fig. 2. The Proposed Loss Function.

III. Experimental Results and Analysis

A. ImageNet Classification

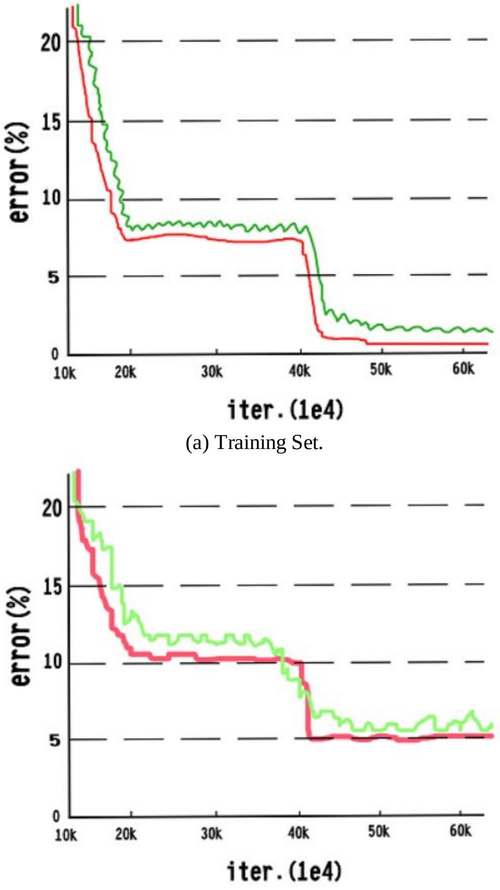
For testify the superiority of our algorithm, ImageNet 2012 classification dataset [19], which include 1.3 million images and 1000 classes, is applied. Our improved ResNet is trained by 1.15 million images, evaluated by 50k images, and eventually tested by 100k images.

The training error rate and test error rate of classical ResNet and our improved ResNet can be found in Fig. 3. It can be seen that curves of classical ResNet both in training set and test set fluctuate badly and curves of improved ResNet changes smoothly with the increase of iteration times.

Top-1 and top-5 error rates of ResNet50 and our improved ResNet50 are shown in Table II. It can be seen that our improved ResNet is a little lower than ResNet.

B. Indoor Localization

Indoor localization experiments are done in a factory environment where 18 regions and location centers are placed. Fig. 4 shows the factory plan. An RGB camera, mobile control devices and a mini-computer are equipped in a mobile robot in which setup our trained ResNet. The experimental platform can be seen in Fig. 5. Using the improved ResNet, the position of a mobile robot can be classified by images taking in real time.



(a) Training Set.

(b) Testset.

Fig. 3. The Error rate of ResNet and Improved ResNet.

TABLE. II. IMPROVED RESNET

model		Top-1 error	Top-5 error
Vgg-16 [20]		22.58	8.43
Googlenet [21]		22.34	7.89
Prelu-net [22]		21.59	5.71
Inception [23]		21.99	5.81
ResNet [11]		20.74	5.25
Improved ResNet		20.35	5.22

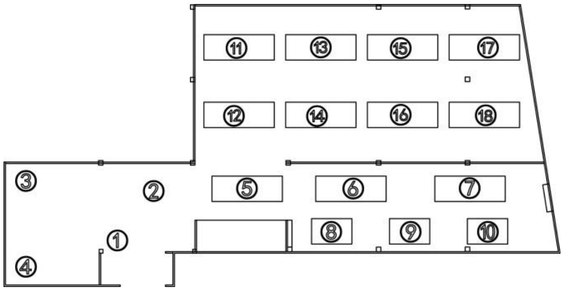


Fig. 4. Floor Plan of the Experimental Scene.

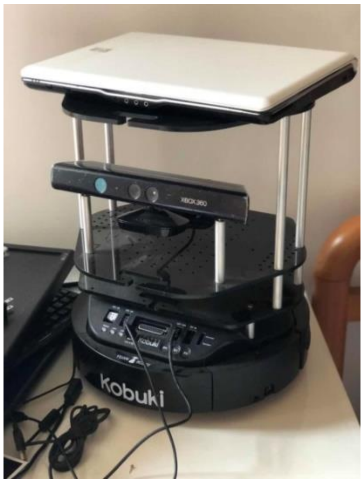
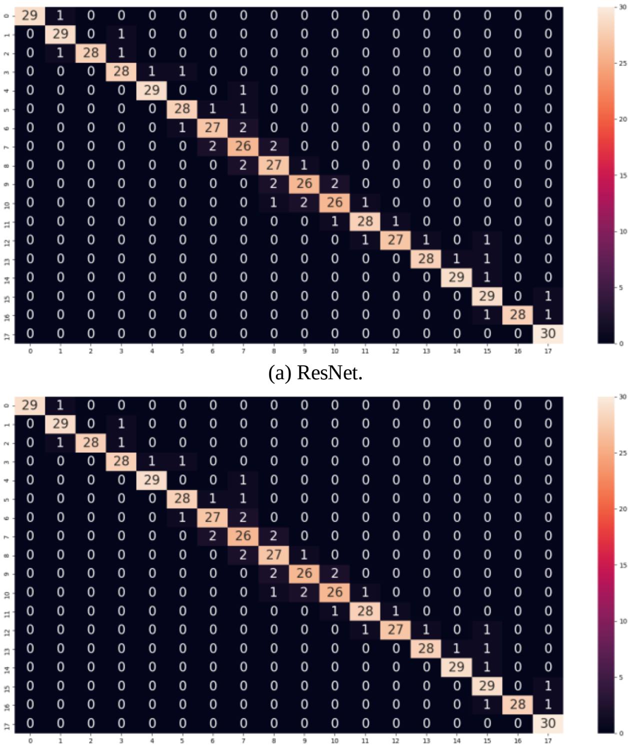


Fig. 5. The Experimental Platform.

We constructed the dataset [24] in a factory environment where we test our algorithm. 1800 samples labeled location information are included in the dataset, we have 100 samples for each location point with different shooting angles.

Confusion matrix is employed to describe localization result, 30 images of each location region were used for testing these experiment results that can be seen in Fig. 6.

We can see in Fig. 6(a) that correct classification time of method1 is 504 and wrong classification time is 36, the accuracy of method1 is 93.33%, more errors happened in middle regions of the scene that is because the location feature of these nearby regions are similar and hard to distinguish, and when the input location feature fluctuate, the output would go to the wrong location. When comes to our improved ResNet, the output would be more stable when the location feature of the input changes, so the accuracy increases. In Fig. 6(b), correct classification time of method1 is 515 and wrong classification time is 25, the accuracy of improved ResNet is 95.37%, the accuracy increased by 2.04%.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	29	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	29	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	28	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	28	1	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	29	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	28	1	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	27	2	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	2	26	2	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	2	27	1	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	2	26	2	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	1	2	26	1	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	1	28	1	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	1	27	1	0	1	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	28	1	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	1	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	28	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30

(a) ResNet.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	29	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	29	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	28	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	28	1	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	29	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	28	1	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	27	2	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	2	26	2	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	2	27	1	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	2	26	2	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	1	2	26	1	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	1	28	1	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	1	27	1	0	1	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	28	1	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	1	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	28	1
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30

(b) Improved ResNet.

Fig. 6. Confusion Matrix of the Localization Results.

IV. CONCLUSION

An improved residual network is proposed in this paper to enhance the stability of classical ResNet. Based on the classical residual network, batch normalization, adaptive β -dropout random deactivation function and a new loss function are added into the proposed model. Batch normalization is applied to avoid vanishing/exploding gradients. β -dropout is applied to increase the stability of the model, which we select different dropout method adaptively by adjusting parameter β . The new loss function is composed by cross entropy loss function and center loss function to enhance the inter class dispersion and intra class aggregation. The improved ResNet50 is then applied to the indoor positioning of mobile robot in a factory environment. The experimental results show that the algorithm can achieve high indoor positioning accuracy under the premise of small training dataset. Future work will focus on the temptation and improvement of other neural-networks to improve the accuracy of the indoor localization system.

REFERENCES

[1] Moustafa A, Moustafa E, Marwan T. “WiDeep: WiFi-based Accurate and Robust Indoor Localization System using Deep Learning”, 2019 IEEE International Conference on Pervasive Computing and Communications, Kyoto, Japan, IEEE, March 2019: 1883-1890.

[2] Cabrera E, Camacho D. “Towards a Bluetooth Indoor Positioning System with Android Consumer Devices,” The IEEE International Conference on Information Systems and Computer science, Quito, Ecuador, 2017: 56-59.

[3] Qiu L, Huang Z, Wirstrom N, et al. “3DinSAR: Object 3D localization for indoor RFID applications,” The IEEE International Conference on RFID, Orlando, USA, IEEE, 2016:101-108.

[4] Desai A, Ghagare N, Donde S. “Optimal Robot Localization Techniques for Real World Scenarios”, 2018 Fourth International Conference on Computing Communication Control and Automation, Pune, India, IEEE, Aug, 2018: 1861-1868.

[5] Walch F, Hazirbas C, Sattler T, et al. “Image-based localization using LSTMs for Structured Feature correlation,” The IEEE International Conference on Computer Vision, Venice, Italy, IEEE, 2017: 627-637.

[6] Krizhevsky A, Sutskever L, and Hinton G. “AlexNet: Imagenet classification with deep convolutional neural networks”, 2012 International Conference and Workshop on Neural Information Processing Systems, Spanish, Lake Tahoe, NIPS, 2012: 3546-3559.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.” International conference on computer vision, IEEE, 2015: 345-367.

[9] S. Ioffe and C. Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” International conference on machine learning, IEEE, 2015:1245-1263.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. “Rabinovich. Going deeper with convolutions.” International conference on computer vision and pattern recognition, IEEE, 2015: 784-796.

[11] He K , Zhang X , Ren S , et al. “Deep Residual Learning for Image Recognition”, 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vega, USA, IEEE, 2016: 770-778.

[12] S. Xie, R. Girshick, P. Dollar, et al. Aggegated residual transformations for deep neural networks. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1492-1500.

[13] J. Hu, L. Shen, G. Sun. “squeeze-and-excitation networks.” IEEE Conference on computer vision and pattern recognition. 2018: 7132-7141.

[14] Y. Chen, J. Li, H. Xiao, et al. “Dual path networks”. Computer Sciences , Arxiv: 1707.01629. 2017: 4467-4475.

[15] S. Zagoruyko, N. Komodakis. “Wide residual networks”. Computer Sciences, Arxiv: 1605.07146. 2016: 2378-2386.

[16] G. Hinton, E. Osindero, et al. “A Fast Learning A lgorithm for Deep Belief Nets”. Neural Computation, 2006, 18(7): 1527-1554.

[17] Chen X J, Guo R Q, Luo W, et al. “Visual Crowd Counting with Improved Inception-ResNet-A Module”, 2018 IEEE International Conference on Robotics and Biomimetics, Kuala Lumpur, Malaysia, IEEE, 2018.

[18] Li B Q, He Y. “An Improved ResNet Based on the Adjustable Shortcut Connections”, IEEE Access, 2018, 5(99): 1348-1356.

[19] http://www.image-net.org/

[20] Simonyan K, Zisserman A. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. Computer Science, 2014, 35(4): 386-400.

[21] Szegedy C, Liu W, Jia Y, et al. “Going deeper with convolutions”, 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015.

[22] He K, Zhang S R and Sun J. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”, 2015 IEEE International Conference on Computer Vision, Santiago, USA, 2015.

[23] Ioffe S , Szegedy C . “Batch normalization: accelerating deep network training by reducing internal covariate shift”, 2015 International Conference on Machine Learning and Cybernetics, Guangzhou, China, JMLR, 2015.

[24] https://pan.baidu.com/s/1oR7fg_sZHe_qHtpSH7PUzg, password: ix4q.

Predicting Book Sales Trend using Deep Learning Framework

Tan Qin Feng[1], Murphy Choy[2]
University of Stirling
Singapore

Ma Nang Laik[3]
Singapore University of Social Science
Singapore

Abstract—**A deep learning framework like Generative Adversarial Network (GAN) has gained popularity in recent years for handling many different computer visions related problems. In this research, instead of focusing on generating the near-real images using GAN, the aim is to develop a comprehensive GAN framework for book sales ranks prediction, based on the historical sales rankings and different attributes collected from the Amazon site. Different analysis stages have been conducted in the research. In this research, a comprehensive data preprocessing is required before the modeling and evaluation. Extensive predevelopment on the data, related features selections for predicting the sales rankings, and several data transformation techniques are being applied before generating the models. Later then various models are being trained and evaluated on prediction results. In the GAN architecture, the generator network that used to generate the features is being built, and the discriminator network that used to differentiate between real and fake features is being trained before the predictions. Lastly, the regression GAN model prediction results are compared against the different neural network models like multilayer perceptron, deep belief network, convolution neural network.**

Keywords—Generative adversarial network; deep learning framework; book sales forecasting; regression

I. Introduction

In the year 2018, the US book publishing industry achieved a net revenue of 25.82 billion USD[1]. With the invention of the internet and the creation of an online purchasing platform, finding and buying books have become so much easier and convenient. Such existence has pushed up the hardcopy and softcopy books' sales. The introduction of digital copies of books, direct publishing has made the publishing process so much simpler and accessible by both authors and readers. Not only that, the digital E-books' competitive pricing helps to stir up the book sales in the region. The book authors also enjoy getting a bigger fraction of the sales split from the self-publishing. These factors mentioned above encouraged many people to write and publish on the internet [1].

According to the Statista website, there are 45,210 writers and authors in the US and the total number of self-publication released in the US exceeded 1 million books in the year 2018[2]. With such a large number of authors available in the US and the mass amount of new book titles being published every year, the market has then become very competitive, especially for the new entry authors. This has become a challenge for them to draw attention from the mass market to their publications, and subsequently to attract readers to purchase their books.

With the improvement in computing power over the past decades, there has been increasing interest from many individuals and companies to use data science approach to predict the demand and sales across various industries. Just from the year 2009 to 2017, the most common machine learning algorithms used in new books sales forecasting are Extreme Learning Machine, K-nearest-Neighbour, Decision Tree, Artificial Neural Network, Random Forest and Service Vector Machine [2]. By mastering the demand and sales through data science, they can gain better foresight and a huge advantage in positioning their resources.

Chen found that the number of readers review is positively related to online book sales. The higher the number of reviews from different users, the better the book performed in the sales. Interestingly, her team also found that the readers' ratings on the books have no relation to the book sales [3]. In their opinion, most of the books received high ratings and dilute the trust and interest of the readers. Contradict to Chen's research, Chevalier and Mayzlin noticed that the online book sales figures have a positive relationship with respective the average stars ratings. The books with higher star ratings tend to have better sales performance than those lower ratings [4]. With such contradictive results from 2 different publications, there must be some other contributing factors that lead to the ups and downs of the sales figures.

There is another category, the time series information that frequently missed out or not available while performing logistic regression predictions or vice versa occurred as the predictions are only focused on time series data, excluding most of the other attributes needed for predictions. In this research, the past rankings of the book collected across many weeks are being included with many other relevant attributes to run the predictions. By merging the time series data with the other numerical attribute, it raises the difficulty level for

[1][1] https://www.statista.com/statistics/288746/global-book-market-by-region/

[2] https://www.statista.com/statistics/249036/number-of-self-published-books-in-the-us-by-format/

predicting the book rankings using the conventional algorithms available.

A. Research Questions

In this research, the study is to develop a comprehensive prediction model that is able to capture both the sales trend and ranking for both printed and digital copies from the many different attributes collected from the Amazon site. Typically, in order to solve such a problem, many conventional types of regression machine learning algorithms like linear regression, random forest, and gradient boost will come into most people's minds as the key approach. Nevertheless, in most of the companies, they do not have sufficient historical data to build a model with good accuracy for the demand and sales forecasting. Furthermore, without the previous historical records of the book sales, implementing the conventional machine learning algorithms becomes more challenging.

In the year 1991, Specht developed and introduced a general regression neural network that provides estimates of continuous variables and converges to the underlying linear or nonlinear regression surface [5]. Subsequently, many companies and researchers started to venture into deep learning algorithms for predicting and forecasting. Similarly, for this research, due to the intricacy of the data with a complex mixture of nominal, ordinal and time-series data, deep learning frameworks similar to Generative Adversarial Network (GAN) are selected in conjunction with other artificial neural networks models as forecasting techniques.

Due to inconsistency of the review studies, plus lack of features in the traditional machine learning algorithms, deep learning frameworks like GAN is able to provide better flexibilities to handle the mass amount of endogenous and exogenous variables of the books. GAN can even perform the sales trend prediction without the demand and many historical sales records. In the end, the research is to develop a comprehensive framework of deep learning that able to complete the tasks in understanding and predicting the hardcopy and softcopy books sales trends with various features.

II. Literature Review

A. On-Line-Analytical Processing and Association Rule Mining Frameworks

In the ubiquitous mobile connected society, there is a tremendous increase in the Social Network Services (SNS) data. The demand for processing mass amounts of data is rising rapidly. At the same time, the volume of collected data made it even more challenging to uncover useful and meaningful information. In order to analyse the SNS data, there are several steps need to follow through. Generally, after data is being collected, the noise in the text needs to be cleaned using Natural Language Processing (NLP) process. The detected text is then documented into matrices forms using the Latent Dirichlet Allocation (LDA) algorithm [6].

On-Line-Analytical Processing (OLAP) and Association Rule Mining (ARM) has been used as a rule-based topic trend analysis (See Fig. 1). OLAP is used to create hierarchical table formation while ARM used to extract the relevant keyword. Working hand-in-hand, they are used to identify previously unknown information and special events. Park and his team (2017) used OLAP and ARM to analyse the social trend and identify similar discussion topics from different users and insights. They showed the feasibility of a combination of the two different data mining techniques [6]. However, challenges still remained for a better understanding of topic trends. Since the frequencies of each topic are classified as measure values in the fact table, in order to handle other types of measured values such as the relative ratio of topics, structure and unstructured data, another deep learning framework is still required.

B. Knowledge base Neural Network Frameworks

In the rise of intensified competition to capture readers' attention, it has then become very important to understand and model online popularity dynamics. Many researchers have been exploring feature-based methods such as random forest and regressions in tackling the task. Since the data is rich in contents and context, Dou and his team used heuristically link online items with existing knowledge base entities to improve the popularity prediction [7]. Fig. 2 shows the schematic diagram of the proposed model. The team has utilized the time series and context data collected in order to generate a robust prediction model.

There are 3 key important issues that need to be considered for the context information popularity prediction. They are types of general contexts used, unified and compact way of representation, and lastly the integration and utilization of the context. To address the issues, a knowledge base neural network is embedded in the Long-Short Term Memory (LSTM) networks for predictions. The prediction gained further improvement when Dou and his team integrated knowledge base neighbours that are used to help to group similar popularity dynamics [7]. The experiment results showed that both the popularity dynamics of the knowledge base neighbors and embedding of the target item improve the prediction results. However, not all the entities can find corresponding knowledge base entries, other methods can be explored to enhance the prediction performance of nonlinked items.

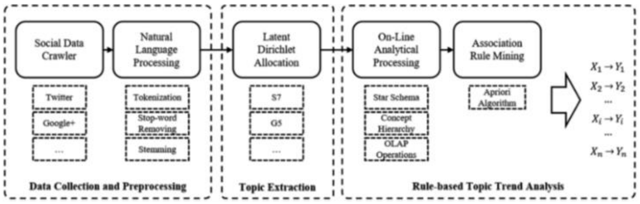


Fig. 1. The Overall Architecture of the Proposed Method [6].

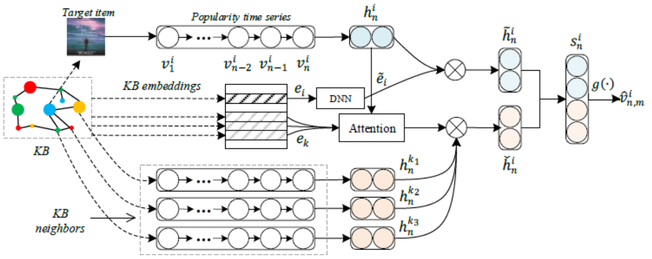


Fig. 2. The Overall Schematic Diagram of the Proposed Model [7].

C. Neural Graph Collaborative Filtering Frameworks

Collaborative filtering (CF) has been widely used in estimating user adoption rate on an item based on the pass interaction behaviours. There are two key components in the learnable CF models. Learning vector representations (aka. Embeddings) that transform the users and items to vectorize representation. The other component is the interaction modeling or matrix factorization (MF) that reconstructs the historical interactions on the embeddings. Due to the lack of explicit encoding of the collaborative signal in the CF, Wang's team developed Neural Graph Collaborative Filtering (NGCF) to make up for the deficiency of suboptimal embeddings [8]. See Fig. 3 for the architecture of the NGCF model developed by Wang's and his team.

By adding an embedding propagation layer, the collaborative signals between the users and items can be harvested for analysis. However, the NGCF still needs further improvement by adding connectivities of different items' orders. The attention mechanism to learn variable weights for neigbours during the embedding propagation also can be enhanced by introducing other models like adversarial learning [8].

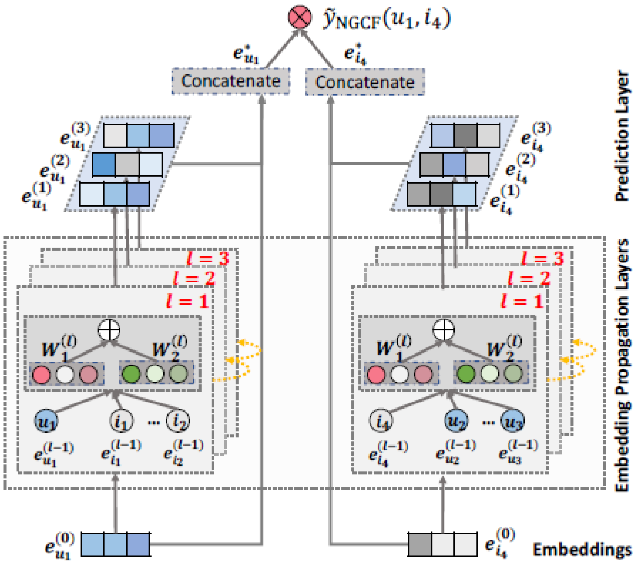


Fig. 3. An Illustration of NGCF Model Architecture (the Arrowed Lines Present the Flow of Information). The Representations of useru1 (left) and Item i4 (Right) are Refined with Multiple Embedding Propagation Layers, whose Outputs are Concatenated to make the Final Prediction [8].

III. Related Work

Recently, there is a lot of attention to the Generative Adversarial Network (GAN). GAN has been introduced by Goodfellow and his team to simultaneously training both the generator model and the discriminative model. GAN has been widely used on a large structure like images with multi-dimension and big output space [9]. Little attention is being drawn to use GAN in solving a single dimension data like classification and regression related problems that only generate small output space. Until recent, Aggarwal and his team started to introduce Conditional Generative Adversarial Network (CGAN) [10] as another comparative model in regression prediction. On the other side, Autoregressive Integrated Moving Average (ARIMA) is one of the popular models when comes to the time series prediction. In Zhang's research, he and his team have selected the GAN model with Long-Short Term Memory (LSTM) network as the generator and Multi-Layer Perceptron (MLP) as the discriminator for predicting the stock price [11].

Studies have been made on using GAN on regression and time series prediction. Both types of research using GAN for prediction are yielding positive results compared against many other deep learning models. As for our research, the data we collected is made up of a mixture of both forms of information. Hence, the GAN framework is being selected and used predicting the book rankings based on all the books' features presented and the past rankings collected.

IV. Using the Template

The methodology in data science requires a structural system of methods to derive particular models targeting a specified area or study. Cross-Industry Standard Process for Data mining (CRISP-DM) methodology is being applied to track and monitor the different milestones of the project. The 6 keys stages are business understand, data understanding, data preprocessing, modeling, evaluation, and deployment [12]. With the CRISP-DM functional template, this will ensure proper procedures being follow-through during the research in order to generate good functional models in predicting the book sales trend.

There are three datasets being selected and used in the research. All the thrrr datasets that were used in this research are considered as secondary data. They are easily accessible from the open dataset site like Kaggle. The first dataset are originally gathered from the Amazon sites. The first dataset contains "ASIN" (a 10 characters long unique Amazon identifier), and other key attributes like the title, authors and publishers of the books. The group and format contained in the dataset helped to distinguish physical and digital versions of the particular ASIN code. The second dataset is focused on Kindle edition. Besides the basic data, the Kindle dataset has the rating, price, number of pages for the books, and some other text attributes which are the languages, lending, customer reviews, short descriptions of the books being published and etc. that the first set lack off.

From both different datasets, they consist of three different types of data. They are made up of quantitative and qualitative data that comprised of numerical, cardinal and text format. Lastly, the third set is made up of collective rankings of a particular ASIN from 1st January 2017 until 29 June 2018. There are totals of 118,200 JSON files being collected. The captured rankings from each file are named in ASIN code, relative to the first dataset. Each individual JSON file represents the book's ranking across the 77 weeks, collected as frequent as an hour to 24 hours interval and the rankings recorded are stamped in binary date-time format. There are some books without ranking initially during the early weeks as they probably not published yet during that period.

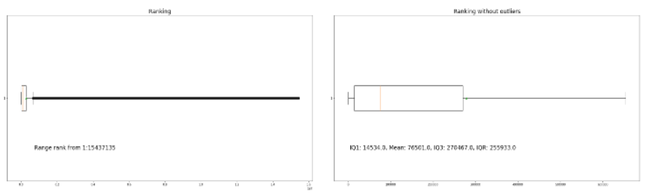


Fig. 4. Boxplots of the Book Rankings with and without Outliers.

Fig. 4 shows the boxplot for all the book rankings collected from the combined dataset. The figure on the left shows the boxplot with the outliers while the figure on the left is the boxplot exclude the outliers. The books' ranks range from the best rank of 1 until the worse rank of over 15 million rankings during the collection period over 77 weeks. From the plot below, most of the books' rankings fall below 700,000 with the average ranking at 76,501 which is only about one-tenth of the majority. Therefore, in our dataset, we noticed that most of the books are performing rather well in their rankings.

V. Setup and Preprocessing

Preprocessing the collected data is one of the important stages before constructing the model. To generate a good model, the quality of data needs to be considered. Not all the single data collected is instantaneously suitable to train and to build the model. Whatever input fits into the model will greatly impact the performance of the deep learning framework and later further affect the output. Hence, the data that is going to feed into the modeling process has to be carefully selected, the Not a Number (commonly appear as NaN) has to be replaced and the entry errors are required to be removed from the list before allowing them to the built the algorithm.

A. *Combine and Setup*

Firstly, the JSON files are being processed by merging all the individual files into one comma-separated value (CSV) files for processing in the later stage. The binary date format in the JSON files are being converted to the readable date-time format as well as the ranking value captured on that particular date and time are stored together in one single file. After the conversion of the files, we proceed to extract the weekly highest ranking and the last ranking achieved for all the books as the output value for training later. As for the other two datasets, they contain rows of books and columns of attributes collected for the book. Removing of missing information and entry errors of the datasets are performed to ensure that the wrong information is not being selected during the modeling process. Lastly, using the unique identifier–the ASIN from all three sets of files, we merged the entire data from the three files into one big dataset.

The 77 weekly highest rankings for the books extracted from the individual lists are also being transposed and inserted as 77 separate columns. The last of the week ranking will then become the targeted sales ranking for the entire research. By merging all the disconnected data into one single set, it will help us to have a better understanding of the relationships between all the variables contents, and enable an effective model building later in the process. Fig. 5 shows the overall rankings being captured across 77 weeks for 2 different books. Fig. 6 shows the mined result from weekly top ranking for the same 2 books. Compare Fig. 5 and Fig. 6, the original graphs' patterns of the overall books' rankings are kept the same even after extracting the top weekly ranking. This will greatly help to bring consistency and uniformity for the model building on the other different books.

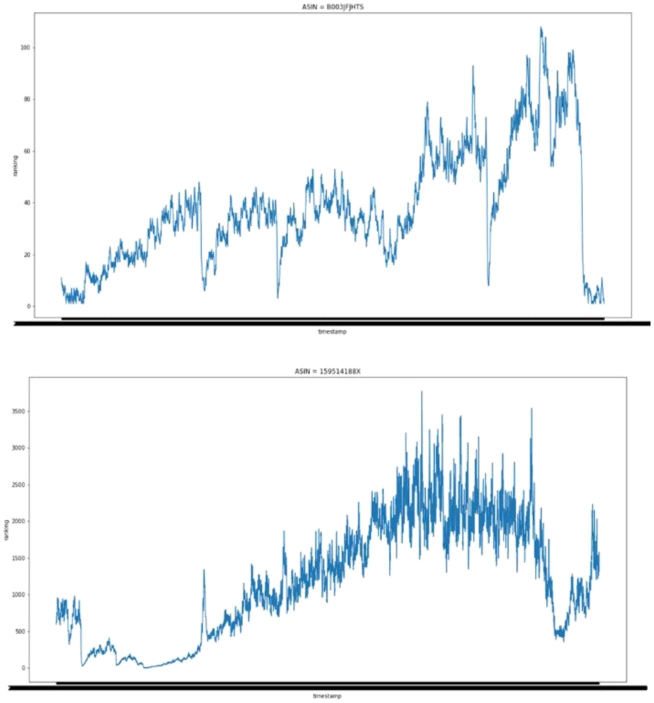


Fig. 5. Snapshots of 2 Books Full Ranking Across 77 Weeks.

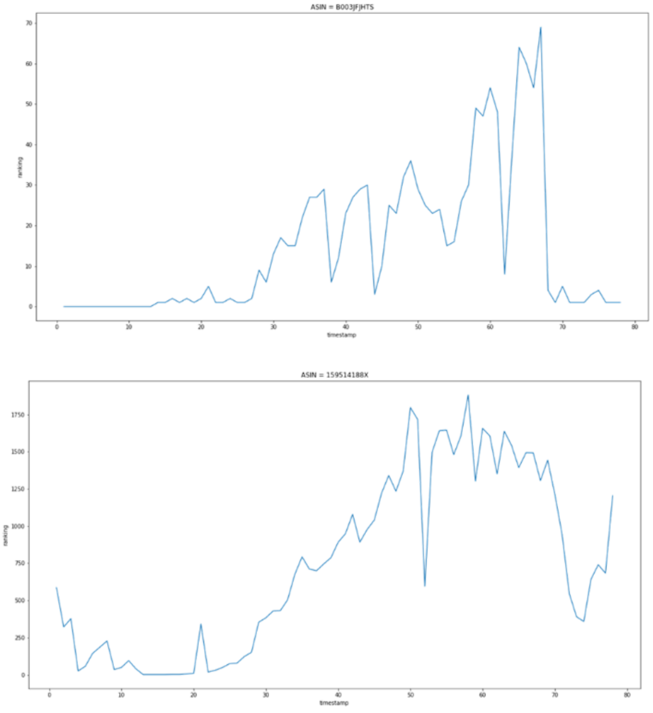


Fig. 6. Snapshots of 2 Books Weekly Top Ranking Across 77 Weeks.

B. Clean and Transform

Once all the data is contained into 1 big set, the data underwent another round of preprocessing by removing duplicates, and all other columns that are not required during the modeling process. Nominal and categorical columns are also being replaced with numerical values for subsequent analysis. As an example, in the 'Lending' feature, the 'Enabled' is replaced with '1' and 'Not Enabled' is replaced with '0' and the four different 'Format' of the book published is replaced with an integer '1~4'.

After the comprehensive constructions and setup are completed, all the preparation works accomplished in the preceding process is to support the main goal, which is to get the data ready for analysis and model building. After merging, the single dataset comprised 108 different columns, and each column representing a parameter. However, for building models and fitting into the mathematical regression, not all 108 features are useful and applicable. Over at this stage, it is necessary to execute another level of data cleaning again, just to make sure the parameters selected are truly meant for training the model. Those unique features that do not belong to any categorical structure will be dropped from the dataset. As an example, the title, authors, publishers, and URL of the books are being removed from the study. In the end, once the cleaning and removing of the unwanted columns are completed, we are left with 90 useful parameters that can be rescaled and transformed for modeling.

To execute the deep learning framework modeling process, the values contained in the dataset need to be transformed into the accepted format. All the features data required to be rescaled, using a min-max scaling formula. It is important to scale the values in the dataset as it allows the relative differences among the values to be treated as equal terms. Plus, it helps to increase proficiency in the arithmetic operations. Especially during the model building process, the transformed values can be used unambiguously by the deep learning framework [13]. Each individual feature is transformed using the formula (see Equation (1)) below, with the range from zero to one.

$$x' = \frac{x - min(x)}{max(x) - min(x)} \qquad (1)$$

where x is the original value and the x' is the new scaled value. In this research, total ranking values were scaled by multiplying by 0.0000000775 and adding 0.

C. Split and Divide

In order to better understand the model performance, it is important to split the data into two different training and validation sets. The prediction results from the validation set allow the user to access the model accuracy after training the model [14]. After all the merging, cleaning and converting on the values in every column, we obtained 1932 rows and 90 columns of good information that can be used for modeling. For the training and validation of the models, we split the total into a ratio of 80:20. We have 1546 sets for training the model and 386 sets to be used during the validation process. From the 1546 sets, we need to have another separate train and test set when building the model. Hence, from the Train_Test_Split function in Python library, we set aside another 30% from the 1546 sets as test dataset. In the end, we have 1082 sets for training; 464 sets for testing and 386 sets keep aside for evaluation after the modeling. To achieve consistency during the entire study, the random state for all the settings is set to state zero.

Cross validate is another commonly use data splitting method in the modeling process. The sample-set is randomly divided into *k* different equal size subsamples where k can be any integer number and often called a number of folds. The *k*-1 subsamples are used to train the model and the remaining one sample is used to validate the model. The entire process is then repeated with k numbers of times and performance on each fold is recorded for evaluation [14]. In our research, we divided our samples into *k*=5 different folds in the modeling process. And the highest, lowest and average scores of the cross-validate results are recorded for further evaluation later in the process. Fig. 7 illustrates the 5 fold cross-validation technique.

D. Correlation

Continue after the preprocessing, we followed by performing a Pearson correlation on all the features in the well-cleaned dataset. This is to evaluate the statistical relationship between the variables. The Pearson correlation coefficient *r* [15] tells the strength and direction of the linear relationship between the 2 variables. The correlation between the two variables can be denoted as r_{xy} and computed as Equation (2):

$$r_{xy} = \frac{cov(x,y)}{\sqrt{var(x)} * \sqrt{var(y)}} \qquad (2)$$

where the *cov* is the sample covariance of *x* and *y*, *var* is the sample variance. The two values gain a stronger positive relationship as the *rxy* moves closer to +1. As the *x* value increases, the *y* value will also increase along with *x* for a positive relationship, vice versa, as the results move closer to -1. When the *r* value moves toward the direction of 0, the relationship between both values grows weaker. Lastly, if *r*=0, both values show no relationship at all.

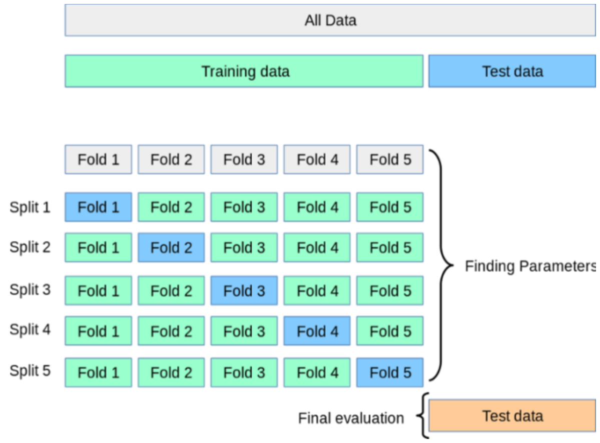


Fig. 7. Illustration of Cross-Validation Technique[3].

[3] https://scikit-learn.org/stable/modules/cross_validation.html

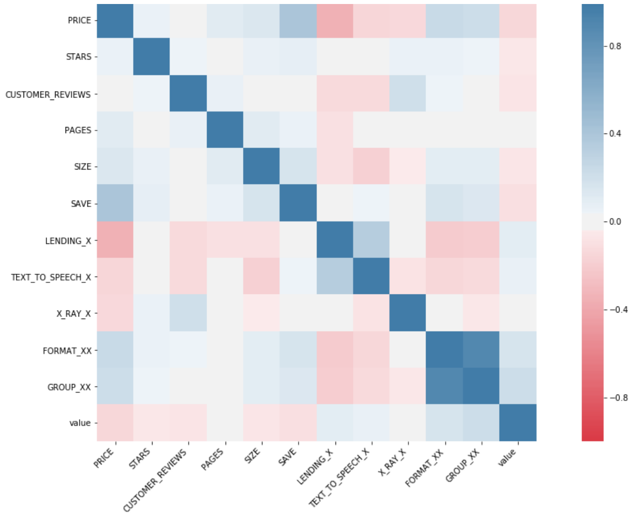


Fig. 8. Snapshot of Pearson's Correlation Heatmap with Values.

In the correlation research, all the top weekly rankings are being excluded from the study. The aim is to study the book's features in response to the final ranking value. From Fig. 8 below, we can see that all the features do not possess a strong relationship with the ranking. There is also a mixture of positive and negative correlations with the last ranking value. Important note for this research, the correlation heatmap shows that the price has a weak negative correlation with the ranking value. Though, the ranking of the book sales is inverse with the numerical number. The lower the numerical value appears in the ranking, the better the book performs. Hence, a clear example is as the price of the book goes up, the book ranking will likely drop because the ranking number grows bigger.

VI. Modeling

There are 3 major types of data analytics for businesses and researchers in understanding and deriving useful information from data. They are descriptive analytics, predictive analytics, and prescriptive analytics. Depending on the values and information that each individual intends to extract out from the data, it requires a different set of analytical techniques. For our research, the focus is to predict the book sales ranking base on the book features and historical ranking collected over a period of time. Predictive analytics [16] is more suitable for our research. It involves a variety of statistical techniques to make predictions about the future. Therefore, predictive models in machine learning are to be selected and applied to generate the desired outcome that fulfills the purpose of the entire study.

In the data mining cycle, the modeling phase is the heart of the process. Just like the heart, it pumps and supplies blood with nutrients to the entire body, the created and selected model is vital to assist businesses and researches in providing accurate and desired results. During the modeling phase, various modeling techniques are chosen and trained once the data, features and models' parameters are properly setups. Generated models are tested, assessed and possibly revised again on parameter settings in order to obtain a perfect outcome. In this paper, Multilayer Perceptron (MLP), Deep Belief Network, Single and two-dimensional Convolution Neural Networks (CNN) are the few deep learning algorithms being selected as a study to compare with the GAN framework. These few artificial neural network architectures comprise of many nodes and several networks connected by one layer with another in sequence to produce the desired results. Research has been conducted using the mentioned neural networks above to understand the performance of each different type of deep learning algorithms in predicting the book sales ranking.

A. *Multilayer Perceptron (MLP)*

The first neural network algorithm that we built is the Multilayer Perceptron (MLP). MLP is one of the most common neural networks and its architecture is also the fundamental architecture for the majority of the neural networks. It is also simple to build and widely used by many researchers. It comprises an input layer and an output layer. In between, it can have many hidden layers connected between the input and output. In the layer itself, it can have one or more artificial neurons called perceptrons. Each perceptron carries weight with activation function to produce a value for the next layer. The output from each node can be represented as Equation (3):

$$h_{w,b}(x) = g(\boldsymbol{w}.\boldsymbol{x} + b) \qquad (3)$$

where *g(x)* is the activation function, *w* is the weight leading to the node and b as the bias [17]. The multilayer perceptron architecture is shown in Fig. 9.

In this study, we constructed a 4 layers MLP. The first layer is the input layer. As we have a total of 90 different columns in our dataset with 89 variables and 1 output, the input dimension for the first layer in the MLP is set to 89 with 360 nodes, a rectified linear unit (ReLU) as the activation function. Subsequently, the other 2 hidden layers are set with 540 nodes and 180 nodes respectively with the same ReLU activation function. The fourth output layer is set to 1 node that will be the output results from the algorithm. As for the output layer, the activation function is set as linear instead as the results will be in a linear regression form. Adam optimizer with the default learning rate of 0.001 is selected when constructing the MLP compiler and train with 1000 epochs.

B. *Deep Belief Network (DBN)*

Similarly, the deep belief network also comprises of 3 different layers. First is the visible layers where the input values are inserted. The next hidden layers are built with Restricted Boltzmann Machines (RBMs) [18] and last is the single output layer that can be in the form of classifications or mathematical regression. The mutual graph of the visible units that represent observations are connected to binary, stochastic hidden units using undirected weighted connections are called RBMs. They are restricted because there are no visible or hidden connections between them. The model gets refined and improved as the hidden layers distribution of the model keeps replacing whenever a better model that is learned by treating the hidden activity vectors produced from the training data as the training data for another RBM. The RBMs have an efficient training procedure which makes them suitable as building blocks for DBNs.

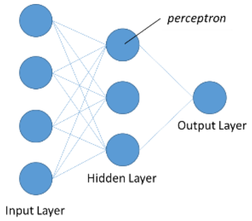


Fig. 9. Multilayer Perceptron Architecture.

In the DBN structure, the θ is the weight of the model, h is the vector of the hidden layer with the distribution of p(h|θ) and the probability vector as p(v). The formula is written as Equation (4) below [18]:

$$p(v) = \sum_h p(h|\theta)\, p(v|h,\theta) \qquad (4)$$

Fig. 10 shows the deep belief network architecture. For our DBN model, we created it similar to the MLP with 4 layers. There are 1 visible layer, 2 hidden layers and 1 final layer as the output layer. The number of nodes available for each layer is set to the same number of nodes as the MLP in the ratio of 360:540:180 and the output layer is activated with a linear regression model. In all the RBM layers' settings, the learning rate is set to 0.1, 5 iterations over the training dataset during the training process and 10 mini-batch sizes.

C. Convolutional Neural Network (CNN)

As the data size gets larger and more dimensions are being introduced to the dataset, especially in the larger images and video contents, the classic neural networks take up a lot of memory space and require very huge computing power to process them. The Convolutional Neural Network (CNN) [19] is then introduced to handle the bigger appetite on data management and data analysis. From the word itself, the neural network uses the convolution technique instead of the matrix multiplication on its layer. The kernel that is much smaller than the input size is put through the activation function to form the output feature map. Pooling function is another feature in the CNN that use to down-sampling the input in order to further increase the receptive field of the outputs. Single or multiple iterations of the convolution process are performed until the parameter gets flatten into a single dimension layer. Then the following process to get the output is similar to MLP once it is flattened [20]. Compare to the traditional neural network consist of 3 layers, the input, the output, and the hidden layer; the algorithm's parameters got reduced and the complexity got simplified by adding the convolution layer and sub-sampling layer [21].

As a whole, the output for CNN is (see Equation (5)):

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right) \qquad (5)$$

where Mj is the selection of the input parameters. i, j, k, l representing the input map, output map, kernel size, and the convolution layer. Lastly, figure b is the additive bias from each output map [19]. The convolution neural network framework is displayed in Fig. 11.

Two different types of convolution neural network (CNN) algorithms have been produced in predicting the book sales rankings. We created a single dimension CNN (1DCNN) and two dimensions CNN (2DCNN) models in this research. For 1DCNN, it is used for input signal patterns like voice and time-series data; and for 2DCNN, it is generally meant to process input signal like images. As for our dataset, the structure is closer to 1 dimension input with a single row and a mixture of numerical and time-series information. For 1DCNN input, we reshape the train and test set by introducing a single channel filter to the dataset with the shape of (row, columns, channel=1). Nevertheless, we can still perform 2DCNN on our dataset by reshaping the data into a 2-dimensional array. We introduced the second dimension and a single channel filter as well to the same train and test dataset into the shape of (row, columns, additional dimension=1, channel=1).

For both 1D and 2D CNN setup, we have two convolution layers, and batch normalization function right after the convolve layers to regulate and normalize the input layers[4]. After the convolution processes, we flatten the signal and add a fully connected layer before the signal is passed to the output. Between each layer, we introduced a 50% dropout regularization technique to prevent overfitting in the neural networks [22]. As the research is focusing on regression output, the linear activation function with $y = ax$ can be used for continuous output. Hence all the layers from input until the output are set with a linear activation function. The filter size for all layers except the last output layers is set to 90 for 1DCNN and 2DCNN. The second input that is set to the convolution layer in the neural network is the kernel size. For our 1DCNN model, the convolution layer kernel size is set to 2 while the 2DCNN model kernel size is set to 1. Both algorithms are trained with 1000 epochs. Adam optimizer with the default learning rate of 0.001 is selected when constructing the 1DCNN and 2DCNN compilers.

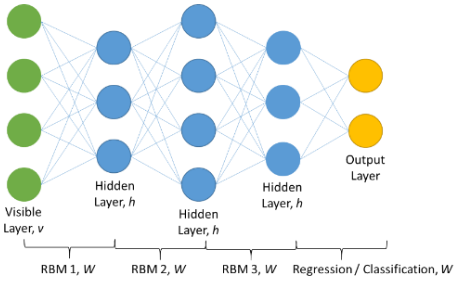


Fig. 10. Deep Belief Network Architecture.

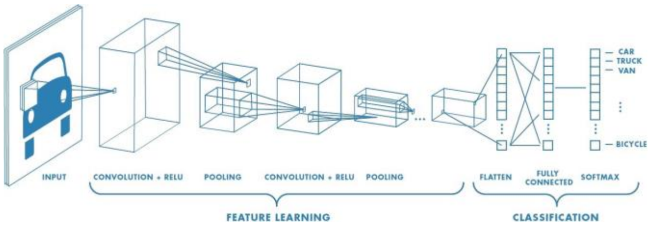


Fig. 11. Convolution Neural Network Frameworks[5].

[4] https://towardsdatascience.com/batch-normalization-in-neural-networks-1ac91516821c

[5] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

D. Generative Adversarial Network (GAN)

GAN is an interesting and unique deep learning framework. It has certainly gained a lot of attention and popularity. GAN functions by integrating two different neural networks, the Generator (G), and the Discriminator (D) to compete and work together simultaneously, just like the zero-sum game from the game theory [23]. In order to cheat the discriminator, the Generator's role is to generate data that mimic the real data by taking in random noise. On the other hand, the Discriminator’s role is to distinguish real and fake data, input from both real datasets and generator sets. Both algorithms are trained together until it reaches a stage that the generator is capable of generating the fake data that the discriminator unable to classify as fake input [11]. In summary, the Discriminator function is to maximize the probability of identifying correct input while the Generator is continuously trained to minimize the probability of letting the Discriminator identified as the generated output as fake input [20].

Fundamentally, the generator and the discriminator are running on two different neural network algorithms as compared to the conventional deep learning framework that performed base on a single neural network. The GAN algorithm starts by defining a prior on the input noise variables *pz(z)* that will be taken in by the generator that represented as neural network *G(z; θg).* The *G* is a differentiable function represented by a neural network with parameters *θg*. As the second definition for the discriminator neural network *D(x; θd),* that provides a singular scalar score The *D(x)* denoted the probability from the sample data *x*. The *D* then trains to maximize the probability of assigning the correct label taking the input from both sample data and generator’s output. Whilst, the *G* is trained to minimize the rejection from *log(1-D(G(z)))* [9]. The GAN’s model equation and diagram are shown in Equation (6) and Fig. 12.

$$min_G max_D V(D,G) = E_{x\sim P_{data}(x)}[\log D(x)] + E_{Z\sim P_z(z)}\left[\log\left(1 - D\big(G(z)\big)\right)\right] \tag{6}$$

For typical GAN would need to be trained with ten to hundreds of thousand iterations to get the optimum results. Conditional GAN or known as CGAN framework is used to assist in the entire GAN prediction process. Just like the GAN, CGAN also has both the generator and discriminator networks. On top of the fundamental *x* and the noise *z*, both generator and discriminator are conditioned with the third variable *y*. The *y* is the information that can be in any form like classification labels or continuous values. It acts as an additional input to both the generator and discriminator for conditioning purposes. *y* joint in the *z* together as an input *p(z|y)* for the generator and presented together with the x as an input *p(x|y)* for the discriminator. This helps to provide boundaries for the expected outputs and speed up the entire training process in the GAN network by giving the generator and discriminator this direction [24]. To illustrate further, the equation for the CGAN and framework are as below (see Equation (7) and Fig. 13):

$$min_G max_D V(D,G) = E_{x\sim P_{data}(x)}[\log D(x|y)] + E_{Z\sim P_z(z)}\left[\log\left(1 - D\big(G(z|y)\big)\right)\right] \tag{7}$$

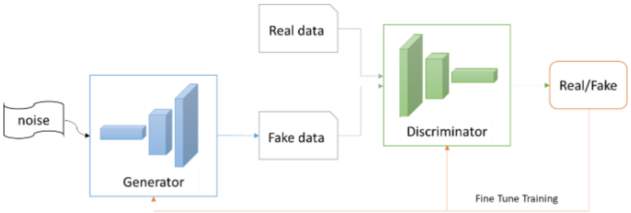


Fig. 12. General Generative Adversarial Network Framework.

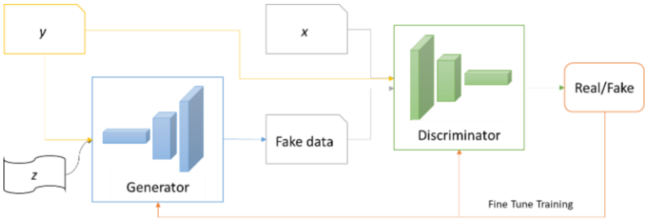


Fig. 13. Conditional General Generative Adversarial Network Framework.

As for our GAN framework, after receiving the random noise input, the generator requires to generate fake data that has the same row and column with the real data. We built a generator network consist of five layers of MLP framework that will take it the noise z and y values to generate out floats in the array format of 89 columns and 1 rows. From the input layer until the final output layer of the generator network, the number of nodes ratio with respect to the number of attributes of the dataset is set to 1:3:2:1:1 in sequence. At every neural network layer, LeakyReLU activation function is added for weight rectification on the nodes. The constant multiplier, *α* with the value of 0.2 is being set on every LeakyReLU function in the generator network. Batch normalization function is also being inserted after the activation function with the momentum value of 0.8 helps to reduce the noise in the gradient. *tanh* activation function is selected for the last output layer in the generator network. The generated output format is then reshaped to make sure the result is identical to the real input *x* that will be feed to the discriminator network later for identification.

In the discriminator network of our GAN design, it is tasked to handle the real and fake *x* input from both real data and generated data by the generator network. Five layers of MLP framework is also being modeled in the discriminator network to handle data array with 89 columns, 1 row. Similarly, the node ratio for the discriminator network at every layer is also set to 1:3:2:1 accordingly except the last output layer is just a single node. LeakyReLU activation functions with *α* of 0.2 are being inserted between the layers. To reduce the overfitting, a 50% dropout rate is set for every discriminator layer. Sigmoid activation is set to the single node at the last output layer for true false identification for the discriminator.

Once set, the GAN is trained with 5000 iterations on the training dataset and the discriminator weight is saved for the prediction later. While training the GAN, we included the condition value y, which is the final ranking value of the training dataset to the generator network and discriminator network in order to hasten and smoothen the training process. Subsequently, the discriminator weight is load again and train with another 1000 epochs for prediction of the book sales ranking.

VII. Evaluation

To understand the reliability of the models built based on the training dataset, the models need to be further evaluated by feeding them with the testing dataset. There are many different types of evaluation techniques that can be applied to understand the performance of the models. Different types of models require specific evaluation techniques to measure their performance and reliability base on their respective outputs. For example, the confusion matrix is well known to evaluate classification type of output while the mean square error and mean absolute error is commonly used in the regression. Therefore, applying the correct evaluation metrics, analysts and business owners can understand how the models behaved before selecting a suitable model for real-world deployment. For our predictive modeling research, the Mean Absolute Error (MAE), the Mean Square Error (MSE) and Root Mean Squared Error (RMSE) are selected to evaluate the deep learning models computed books' rankings against the collected rankings.

A. *Mean Absolute Error (MAE)*

MAE is a quality measurement metric that widely used in regression evaluation. It is used to measure the absolute difference between the actual values and the estimated values. The MAE formula is defined as Equation (8):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \tag{8}$$

where *n* is the number of the sample size, y_i is the original observed value and $\hat{y}_i$ is the predicted value. In this context, the *n* is the size of the testing and predicting dataset and *y* is the final captured book's ranking across the study. The model's predicted ranking is represented by the $\hat{y}$. When using the MAE for evaluation, the smaller the number, the better the predicted values as they are closer to the expected values.

B. *Mean Square Error (MSE) and Root Mean Squared Error (RMSE)*

Similar to MAE, MSE is another type of quality measurement metric that popular in the regression evaluation as well. Instead of absolute difference, it measures the average squared distance between the actual values and the estimated values. However, both MAE and MSE have a slight difference in the meaning of the value calculated. The MAE ignores the direction or the negative values from the calculation. Whereas the MSE squared the differences between observed and expected value. Hence, the MSE carries more weight to a bigger loss in the calculation, in which larger errors are particularly undesirable. MSE formula is defined as Equation (9):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2 \tag{9}$$

As for the RMSE formula, it is just the square root for the MSE. For MSE and RMSE measurements, similar to MAE scores, the smaller the MSE value, the better the predictive model generates results.

C. *Results Comparison*

Besides the 30% testing dataset that split out from the training set, we have another 386 prediction dataset that is not part of any training and testing set which is meant for evaluating the algorithms on untapped datasets outside the training and testing sets. The prediction dataset was randomly separated during the preprocessing is to ensure that none of the data selected during the modeling phase is being recycled again for training. This is to show the reliability and performance of the model in predicting the unknown dataset. The tables below shows the MAE, MSE and RMSE scores for the 5 different types of deep learning networks generated on predicting the book sales ranking against the original value. The percentage scores in both tables show the delta *Δ* of improvement for the models comparing the test results and prediction results. The formula of the *Δ* improvement is as Equation (10):

$$\Delta\% = \frac{y_{test} - y_{predict}}{y_{test}} \tag{10}$$

In the first glance from Table I, the smallest figure tabulated starts with 5 digits and the largest integer can go larger than 350,000. It seems to have a rather big MAE score in the books' rankings predictions. However, when we compared with the books' rankings ranges in 10 of millions, the worst performance neural network, 1DCNN scores around 350,000 is only about 3.5% from the total ranking value. Therefore, the evaluation scores are considered acceptable. From the 5 neural networks, MLP overall scores the best MAE for predicting the book rankings for both test and predict dataset prepared. Whereas the GAN frameworks that leverage on MLP architecture for the generator and discriminator networks perform second best and it is just a few thousand scores off from the MLP. Nevertheless, when comes to the comparison between the test set and predicted set, the MLP, 2DCNN, and GAN saw a positive difference. Among the 3, the GAN framework has the best improvement of about 5.5%. It is more than 5% better as compared to MLP which does not see much improvement between the test set and predict set. If more datasets and training can be provided to GAN, it can definitely generate better predictions for the books' rankings.

Table II shows the MSE and RMSE results scored by each neural network in predicting the books' ranking for test and predict dataset. Due to the MSE squared function, plus the large ranking values and the big ranking ranges possessed within the dataset, the scores grow so huge that the worst performance 1DCNN results reach beyond 11 digits. Hence the RMSE helps to reduce the dimension so that it is much easier to read and decipher. Among all five different frameworks, GAN has the lowest RMSE values for both test and predict sets. Within the GAN framework, the scores also reduced from test RMSE to predict RMSE with an overall improvement of 14.25%. Although it seems 2DCNN ranked second at the improvement scores, the 2DCNN RMSE values for both the test and predict are more than double the value of the GAN RMSE scores. MLP algorithm shows the best reduction of the RMSE value at 16.77% from the test to predict the dataset. Nonetheless, our GAN generator and discriminator networks are using the MLP framework to perform hand in hand together for generating the prediction on the books' ranking. Similarly, as the MAE results, if we can perform more iterations of training to the GAN framework, we can see even better RMSE scores as well.

TABLE. I. MEAN ABSOLUTE ERROR (MAE) RESULTS

Neural Network	**Test MAE**	**Predict MAE**	**Δ Improved**
MLP	**80836**	**80815**	0.03%
DBN	94681	109936	-16.11%
1DCNN	352150	355656	-1.00%
2DCNN	181876	174085	4.28%
GAN	89723	84787	**5.50%**

TABLE. II. MEAN SQUARED ERROR (MSE) AND ROOT MEAN SQUARED ERROR (RMSE) RESULTS

Neural Network	**Test MSE**	**Test RMSE**	**Predict MSE**	**Predict RMSE**	**Δ Improved**
MLP	2.73e10	165108	1.89e10	137415	**16.77%**
DBN	2.25e10	149909	4.08e10	201872	-34.66%
1DCNN	1.67e11	409053	1.73e11	416005	-1.70%
2DCNN	9.02e10	300352	6.44e10	253713	15.53%
GAN	1.98e10	**140611**	1.45e10	**120576**	14.25%

MAE and RMSE are often used together as accuracy indicators for continuous variables. By having both the indicators tabulated together, we can derive another level of information. From the definition, RMSE will never be smaller than MAE, as RMSE ≥ MAE. Both values can be in the range from 0 to ∞. By subtracting RMSE and MAE, we are able to understand the variation of errors in the forecasted results. The larger the remaining value from the subtraction, the greater the variance the individual error can be found in the sample set. When both errors are having the same magnitude, then RMSE is equal to MAE [25], RMSE = MAE.

Since we have RMSE and MAE value calculated, we can understand how well the deep learning frameworks performed in forecasting the books' rankings that we expect. Table III below shows the comparison results for the 5 algorithms after we subtract the MAE with RMSE. Among all, GAN has the least difference from both test and predict dataset. It shows that the variance in GAN individual error is the smallest among all the neural networks. Out of surprise, the predict set number is 30% lower as compared to the test set. This shows that GAN predictions on the books' rankings versus the actual rankings are even closer to each other.

The tables below show the cross-validated training results for all the models. In our research, the training and testing datasets are randomly split into five different sets for modeling. MAE and RMSE results are shown in Tables IV and V, respectively after 5 rounds of cross-validations. Table VI shows the values of RMSE–MAE on all 5 algorithms. This time around, in the cross-validation training, GAN performed the best in all areas. On average, GAN scores the lowest in both MAE and RMSE results.

TABLE. III. RMSE SUBTRACT MSE SCORES

Neural Network	**Test RMSE - MAE**	**Predict RMSE - MAE**
MLP	84272	56600
DBN	55228	91936
1DCNN	56902	60349
2DCNN	118475	79628
GAN	**50887**	**35789**

TABLE. IV. CROSS VALIDATED MAE RESULTS

MAE Results	***k* = 1**	***k* = 2**	***k* = 3**	***k* = 4**	***k* = 5**	**Average**
MLP	731664	778127	918803	623871	827868	776067
DBN	105952	136709	123384	102635	112538	116243
1DCNN	754233	794243	930235	689700	644626	762608
2DCNN	754233	794243	930234	689700	644626	762607
GAN	85278	107280	88790	75405	82809	87912

TABLE. V. CROSS VALIDATED RMSE RESULTS

RMSE Results	***k* = 1**	***k* = 2**	***k* = 3**	***k* = 4**	***k* = 5**	**Average**
MLP	2.04e6	1.83e6	2.16e6	1.43e6	1.82e6	1.86e6
DBN	2.73e5	4.37e5	3.08e5	1.78e5	2.66e5	2.92e5
1DCNN	1.82e6	1.97e6	2.19e6	1.59e6	1.51e6	1.81e6
2DCNN	1.82e6	1.97e6	2.19e6	1.59e6	1.51e6	1.81e6
GAN	1.37e5	1.51e5	1.52e5	1.24e5	1.26e5	1.38e5

TABLE. VI. RMSE SUBTRACT MSE (CROSS-VALIDATED)

RMSE - MAE	***k* = 1**	***k* = 2**	***k* = 3**	***k* = 4**	***k* = 5**	**Average**
MLP	1307696	1054275	1241158	810764	988374	1080453
DBN	167086	300770	184280	74945	153155	176047
1D CNN	1061617	1173695	1261903	899627	863283	1052025
2D CNN	1061617	1173695	1261904	899627	863283	1052025
GAN	51941	44020	63187	48897	43551	50319

VIII. CONCLUSION

Generative Adversarial Network, GAN has gained a lot of attention in researches and a lot of momentums in publications. Ever since it was introduced in the year 2014, the numbers of publishing GAN related papers and journals are increasing over the past few years. In addition, many sub categorical types of GAN have been invented for example Deep Convolutional Generative Adversarial Networks (DCGAN), Wasserstein Generative Adversarial Networks (WGAN) and Conditional Generative Adversarial Networks (CGAN) that helps the GAN training to be more stable and easier [20]. However, the majority of them are focused on multidimensional unsupervised learning, especially in computer vision processing. Until recently, CGAN starts to capture attention in regression and time series prediction. Our research also shows that CGAN is capable of performing the regression predictions integrated with time series information effectively, for example, the books' rankings prediction that contained the books' features and their past rankings.

Initially, from the original books' ranking dataset, we have more than 100,000 books' rankings being captured over the 77 weeks. Secondly, in the original books' features dataset, we have collected nearly 50,000 individual rows of books' attributes. In the best-case scenario, we should have almost 50,000 different books that consist of their own attributes and rankings that collected over one year. After we performed the necessary cleaning and combined all the dataset into one single meaningful dataset base on similarity features, we are left with about 1900 books. As we further split the dataset for training, testing and predicting, the training set is left with no more than 1100 books. This shows that a lot of the book titles are not being captured in the feature set that ends up lead to very few datasets for training our algorithms.

The other challenging part that we faced is the books' rankings range. The gap within the books' rankings is too wide. The best and worst performance book ranking can vary from top 1 to bottom 15 million in rank values. To make things even worse, based on the boxplot, the books' rankings are not evenly spread as well. The other finding we noticed while performing the Pearson correlation coefficient matrix is the book's attributes do not have a strong relationship with the final ranking values. Many of the correlation coefficient, r fall below ±0.1 and near to zero. According to The Basic Practice of Statistics, for any r value fall below ±0.3 indicates a weak relation between the two attributes [26]. The best positive and negatively correlated values are not more than 0.23 and -0.15. These indicate that the books' features collected have a weak relationship with the books' rankings. Even if we include the weak attributes in building the deep learning frameworks, they do not bring significant weights in predicting the final ranking values.

With only 1100 rows of books that possessed weak features, and having a very big spread of rankings values that appeared between the 1100 books, it is not the models to blame on underperforming. These also lead to huge numbers with many digits are appearing in the MAE and RMSE scores on all the deep learning frameworks in our research. With just a few bad predictions and cause a huge difference between the predicted ranking and expected values, the calculated loss functions will spike upward significantly.

In order to strengthen and benefit the model training, the first approach is to increase the number of datasets available for training. It is very important as the books' rankings have a large variation. Bringing up the numbers in the training dataset will help to narrow the gap between one book with another. With more datasets, the books' features and the rankings can become more distinctive as well. Secondly, it is to collect and identify more books' features that possess either a strong positive or a strong negative correlation with the ranking. Currently, the book's features are weak and the prediction of the book's final ranking figure relies heavily on the past rankings collected over a year. By inserting a strong correlation coefficient book's attributes, hopefully, the book's attributes and historical rankings can achieve a better balance in the weight of the mathematical functions in the deep learning framework.

For future research, we would suggest continuing to explore the supervised and semi-supervised learning with the GAN framework. There are many other areas that we can continue to study and leverage on the GAN to build a linear, nonlinear, or logistic regression type of mathematical algorithms. To have a better understanding of different GAN behaviours in regression predictions, moving forward, we probably can explore by including more types of GAN frameworks in our studies and compare the performance among themselves.

For both the generator and discriminator networks that appear in the GAN framework, both the neural networks are working and competing at the same time as the zero-sum game kind of relationship. Due to this, GAN typically requires an extensive amount of training, large computing power and long hour of processing in order to achieve the desired results. Henceforth, there is still a big room of improvement available for GAN to move forward and evolved to the next level with better capabilities and higher efficiencies.

REFERENCES

[1] Flood, A., 2011. How self-publishing came of age. The Guardian.

[2] Cadavid, J.P.U., Lamouri, S., Grabot, B., 2018. Trends in Machine Learning Applied to Demand & Sales Forecasting: A Review 11.

[3] Chen, P.-Y., Wu, S., Yoon, J., 2004. The Impact of Online Recommendations and Consumer Feedback on Sales 15.

[4] Chevalier, J.A., Mayzlin, D., 2006. The Effect of Word of Mouth on Sales: Online Book Reviews. J. Mark. Res. 43, 345–354.

[5] Saha, S., 2018. A Comprehensive Guide to Convolutional Neural Specht, D.F., 1991. A general regression neural network. IEEE Trans. Neural Netw. 2, 568–576. https://doi.org/10.1109/72.97934.

[6] Jeon, Y., Cho, C., Seo, J., Kwon, K., Park, H., Chung, I.-J., 2017. Rule-Based Topic Trend Analysis by Using Data Mining Techniques, in: Park, J.J., Chen, S.-C., Raymond Choo, K.-K. (Eds.), Advanced Multimedia and Ubiquitous Engineering. Springer Singapore, Singapore, pp. 466–473. https://doi.org/10.1007/978-981-10-5041-1_75.

[7] Dou, H., Zhao, W.X., Zhao, Y., Dong, D., Wen, J.-R., Chang, E.Y., 2018. Predicting the Popularity of Online Content with Knowledge-enhanced Neural Networks 8.

[8] Wang, X., He, X., Wang, M., Feng, F., Chua, T.-S., 2019. Neural Graph Collaborative Filtering. ArXiv190508108 Cs. https://doi.org/10.1145/3331184.3331267.

[9] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets 9.

[10] Aggarwal, K., Kirchmeyer, M., Yadav, P., Keerthi, S.S., Gallinari, P., 2019. Conditional Generative Adversarial Networks for Regression. ArXiv190512868 Cs Stat. (10).

[11] Zhang, K., Zhong, G., Dong, J., Wang, S., Wang, Y., 2019. Stock Market Prediction Based on Generative Adversarial Network. Procedia Comput. Sci. 147, 400–406. https://doi.org/10.1016/j.procs.2019.01.256.

[12] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. Step-by-step data mining guide. SPSS Inc 76.

[13] Harwell, M.R., Gatti, G.G., 2001. Rescaling Ordinal Data to Interval Data in Educational Research. Rev. Educ. Res. 71, 105–131. https://doi.org/10.3102/00346543071001105.

[14] Xu, Y., Goodacre, R., 2018. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. J. Anal. Test. 2, 249–262. https://doi.org/10.1007/s41664-018-0068-2.

[15] Yeager, K., 2020. LibGuides: SPSS Tutorials: Pearson Correlation [WWW Document]. URL https://libguides.library.kent.edu/SPSS/PearsonCorr (accessed 1.8.20).

[16] Nyce, C., 2007. Predictive Analytics White Paper. Am. Inst. CPCU 24.

[17] Patterson, J., Gibson, A., 2018. Getting started with deep learning [Book]. O'Reilly Media, Inc.

[18] Mohamed, A., Dahl, G., Hinton, G., 2009. Deep Belief Networks for phone recognition 9.

[19] Bouvrie, J., 2006. Notes on Convolutional Neural Networks 8.

[20] Neff, T., 2018. Data Augmentation in Deep Learning using Generative Adversarial Networks 113.

[21] Wu, Y., Yang, F., Liu, Y., Zha, X., Yuan, S., 2018. A Comparison of 1-D and 2-D Deep Convolutional Neural Networks in ECG Classification 4.

[22] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting 30.

[23] Oliehoek, F.A., Savani, R., Gallego-Posada, J., van der Pol, E., de Jong, E.D., Gross, R., 2017. GANGs: Generative Adversarial Network Games. ArXiv171200679 Cs Stat.

[24] Mirza, M., Osindero, S., 2014. Conditional Generative Adversarial Nets 7.

[25] Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247–1250. https://doi.org/10.5194 /gmd-7-1247-2014.

[26] Mindrila, D., Balentyne, P., Ed, M., 2012. Scatterplots and Correlation. Basic Pract. Stat. 6th Ed, Chapter 4 14.

Sea Breeze Damage Estimation Method using Sentinel of Remote Sensing Satellite Data

Kohei Arai
Faculty of Science and Engineering
Saga University, Saga City, Japan

***Abstract*—Sea breeze damage estimation method using Sentinel of remote sensing satellite data is proposed. There are two kinds of sea breeze damage. Namely, one is vegetation degradation due to sea salt from sea breeze and the other one is leaf lodging due to strong winds from sea breeze. Kyushu, Japan had severe storm due to the typhoon #17 during from 21 September to 23 September 2019. Optical sensor and Synthetic Aperture Radar: SAR onboard remote sensing satellite are used for disaster relief. NDVI and false colored imagery data derived from the Sentinel-1 and 2 data are used for disaster relief. Through experiments, it is found that sea salt damage on rice paddy fields in particular can be relieved by NDVI and false colored imagery data while rice lodging can also be relieved by SAR data.**

***Keywords*—Sentinel; disaster relief; satellite remote sensing; flooding; oil spill; synthetic aperture radar; optical sensor; vegetation index**

I. Introduction

A large and strong typhoon #17 traveled north from Okinawa in the East China Sea, and approached Kyushu on September 22 in 2019 and advanced from the vicinity of the Tsushima Strait to the Sea of Japan.

According to the Saga Local Meteorological Observatory, if Typhoon #17 moved near the center of the forecast circle, Saga Prefecture entered a strong wind area with a wind speed of 15 meters or more by the early of the 22nd of September. On the evening of the same day, it entered a storm zone with a wind speed of 25 meters or more and approached most early in the night. As the typhoon approached, warm and humid air flowed in and the air condition became extremely unstable. There was a very heavy rainfall, and there was also a risk of lightning, tornadoes and other severe gusts. The expected rainfall was 50 mm per hour in many places and 200 mm in 24 hours.

According to the Agricultural Administration Bureau, the cultivated area of rice in the prefecture is 24,100 hectares, the yield has dropped to 71,800 tons, 56% of last year, and more than 30% of non-standard rice is generated. Yield per 10 Ares was 298 kg, down 234 kg from last year. The reason was that the number of spikelets and firs decreased in early July and late July due to low temperatures and lack of sunshine, but this was due to the salt damage caused by Typhoon #17 in September and the damage caused by insect pests, brown planthoppers.

The crop index 58 was even lower than 86 in Kyushu as a whole. In each district, Matsuura is 79 and Saga is 53. According to the Bureau of Agriculture, typhoon #17 caused winds to blow north from the Ariake Sea, causing salt damage to spread mainly along the coast.

There are two major influencing disaster induced by typhoon, heavy rain and severe wind. Heavy rain is major cause of flooding, landslide, etc. while severe wind is major cause of storm surge, sea breeze, sea salt damage, lodging and so on. These disaster can be detected with remote sensing satellite based mission instruments, visible to near infrared radiometer, short wave infrared radiometer, thermal infrared radiometer, microwave radiometer, and Synthetic Aperture Radar: SAR and so on.

Previously, the method for estimation of sea salt amount attached to rice leaves with remote sensing satellite data is proposed [1]. Validation and confirmation of the proposed method, however, are not enough. In this regards, additional confirmation on effectiveness of the remote sensing satellite data for estimation of damage grade due to sea salt derived from typhoon sea breeze is conducted in this paper.

In order to relief disaster occurred areas, Sentinel-1 and 2 satellites based Visible to Near InfraRed: VNIR radiometer and SAR are used. Through comparison of the VNIR and SAR imagery data derived from Sentinel-1 and 2 satellite based instruments between before and after disaster occurrence, these disaster can be detected, in general. It is relatively easy to detect landslide, flooding, etc. because ground cover targets may change in particular vitality of vegetation while storm surge, sea breeze, sea salt damage and lodging are not so easy to detect. In particular, rice lodging causes a small vegetation vitality changes while backscattering coefficient of the vegetated areas may change due to rice lodging. Therefore, a method for sea salt damage area estimation is proposed by using VNIR imagery data together with a method for rice lodging damage area estimation by using SAR imagery data in this paper.

The next section describes related research works followed by the research back ground of this study. Then, experimental results are described followed by conclusions together with some discussions.

II. Related Research Works

There are some related studies on disaster relief and mitigation research works, Method for estimation of damage grade and damaged paddy field areas sue to salt containing sea breeze with typhoon using remote sensing imagery data is proposed and validated [1]. Four dimensional GIS and its application to disaster donitoring with satellite remote sensing data is proposed in the Conference on GIS and Application of Remote Sensing to Disaster Management [2], [3].

Meanwhile, URL search engine with text search tools for disaster mitigation is proposed in the Asian Disaster Reduction Center R&D Project Workshop [4]. On the other hand, visualization of disaster information derived from Earth observation data is proposed [5]. Also, ICT technology for disaster mitigation, tsunami warning system is proposed in the 1st International Workshop on Knowledge Cluster Systems [6]. In the same time, cellular automata approach for disaster propagation prediction and required data system in GIS representations is proposed in the 1st ICSU/WDS Conference - Global Data for Global Science [7].

Meantime, cell based GIS as Cellular Automata: CA for disaster spreading prediction and required data systems is proposed in the CODATA Data Science Journal [8] together with disaster relief with satellite based synthetic aperture radar data [9]. Meanwhile, Sentinel 1A SAR data analysis for disaster mitigation in Kyushu is presented [10]. On the other hand, Sensor network for landslide monitoring with laser ranging system avoiding rainfall influence on laser ranging by means of time diversity and satellite imagery data based landslide disaster relief is proposed and validated [11].

Quite recently, flooding and oil spill disaster relief using Sentinel of remote sensing satellite data is reported [12],

III. REASEARCH BACKGROUND

On the morning of September 22, the typhoon #17 hits Japanese vicinity. Fig. 1 shows the trajectory of the typhoon #17. Starting from southern sea area of Okinawa, to the center of the Japan Sea, typhoon #17 traveled with strong wind and heavy rain. At around 9 a.m. on September 23, the typhoon was changed to temperate cyclone.

Fig. 2(a), (b), and (c) show atmospheric pressure maps and geostationary meteorological satellite MTSAT (Japanese Meteorological Satellite in the geostationary orbit) images observed at 3:00 (UTC) on 21, 22, and 23 September 2019, respectively.

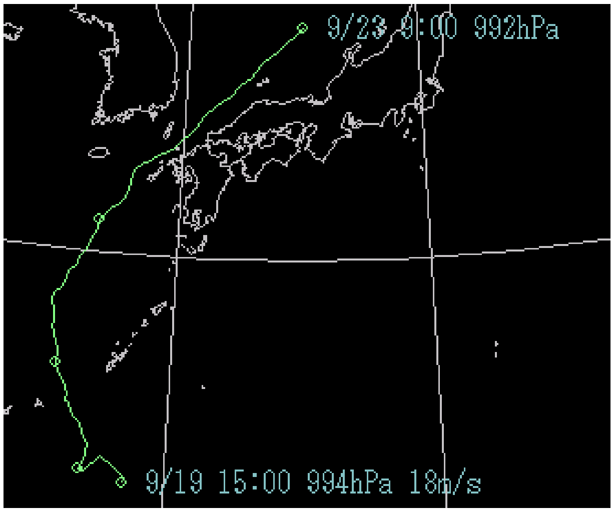
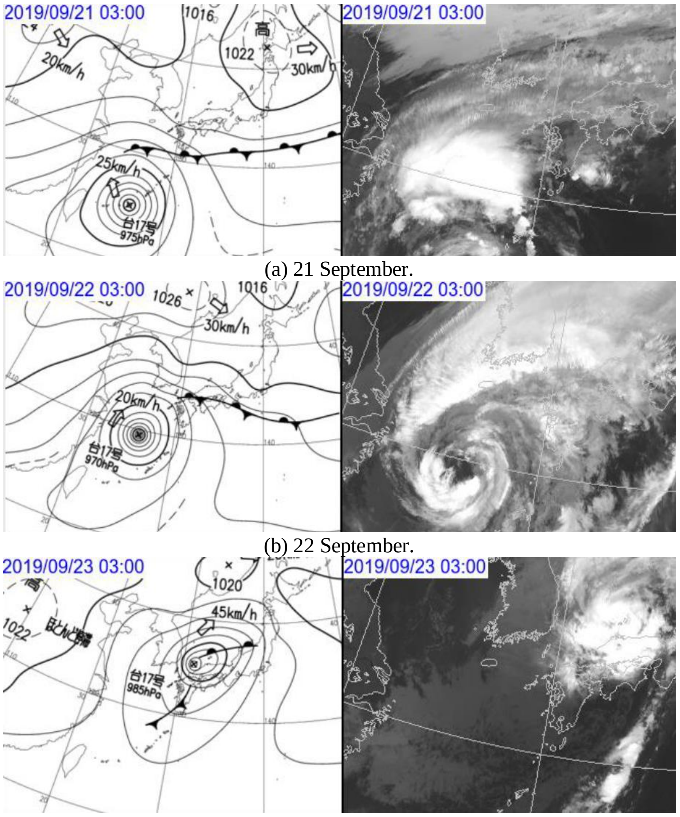


Fig. 1. Trajectory of the Typhoon #17.



(a) 21 September.

(b) 22 September.

(c) 23 September.

Fig. 2. Atmospheric Pressure Maps and Geostationary Meteorological Satellite Images Observed at 3:00 (UTC) on 21, 22, and 23 September 2019.

In particular at 12:00 (JST), noon on 22 September, Kyushu Japan had a heavy rain and a severely strong wind as shown in Fig. 3. Elapsed time of the maximum wind speed and the momentary maximum wind speed are shown in Fig. 4(a) and (b), respectively.

Although rainfall rate of the typhoon #17 is not so heavy, wind speed of the typhoon #17 is extremely high. Due to the severely strong wind shown in Fig. 4, sea salt damage and rice lodging are occurred in Kyushu, Japan.

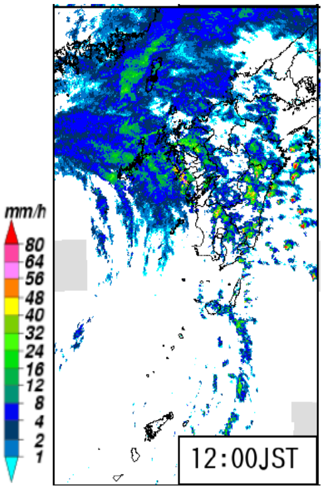
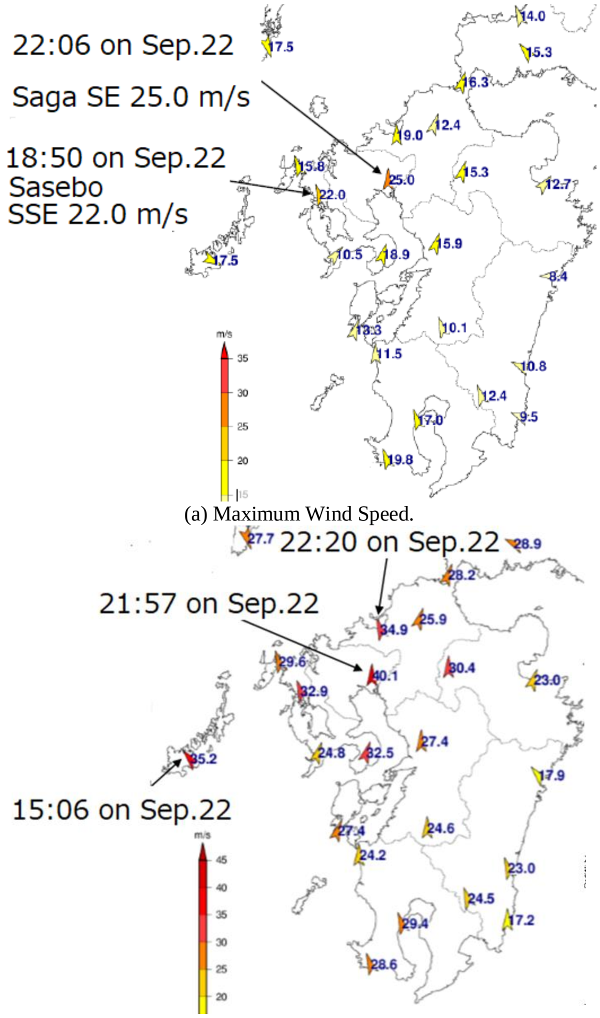


Fig. 3. Heavy Rain and a Severely Strong Wind at 12:00 (JST), Noon on 22 September, Kyushu Japan.



(a) Maximum Wind Speed.

(b) Momentary Maximum Wind Speed.

Fig. 4. Elapsed Time of the Maximum Wind Speed and the Momentary Maximum Wind Speed.

IV. Proposed Method

As shown in previous paper, there is the relation between sea salt amount attached to the rice leaves in concern and Normalized Difference Vegetation Index: NDVI derived from remote sensing satellite (Sentinel-2) based visible to near infrared radiometer data as shown in Fig. 5. Therefore, it is possible to estimate vegetation damage by sea salt due to sea breeze by using this relation. Sea breeze damage grade is estimated with degradation of NDVI through a comparison between NDVI before typhoon and after typhoon.

On the other hand, it is hard to estimate vegetation damage by leaf lodging by strong wind due to sea breeze. It, however, is possible to estimate leaf lodging by using back scattering coefficient derived from synthetic aperture radar data which is derived from Sentinel-1A and 1B.

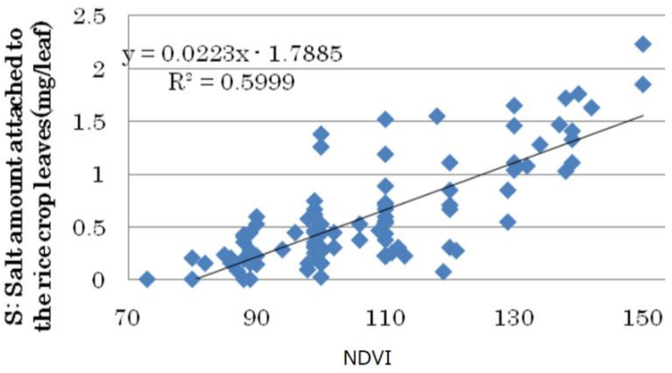


Fig. 5. Relation between Sea Salt Amount Attached to the Rice Leaves in Concern and NDVI Derived from Remote Sensing Satellite based Visible to near Infrared Radiometer Data.

V. Experiments

A. Intensive Study Area

Intensive study area of Saga prefecture and northern Kyushu area are shown in Fig. 6(a) and (b), respectively. 25.0 m/s of maximum wind speed from the south west direction was observed in Saga prefecture at 22:06 (JST) on 22 September 2019 while 40.1 m/s of momentary maximum wind speed from the south was observed at 21:57 (JST) on that day.

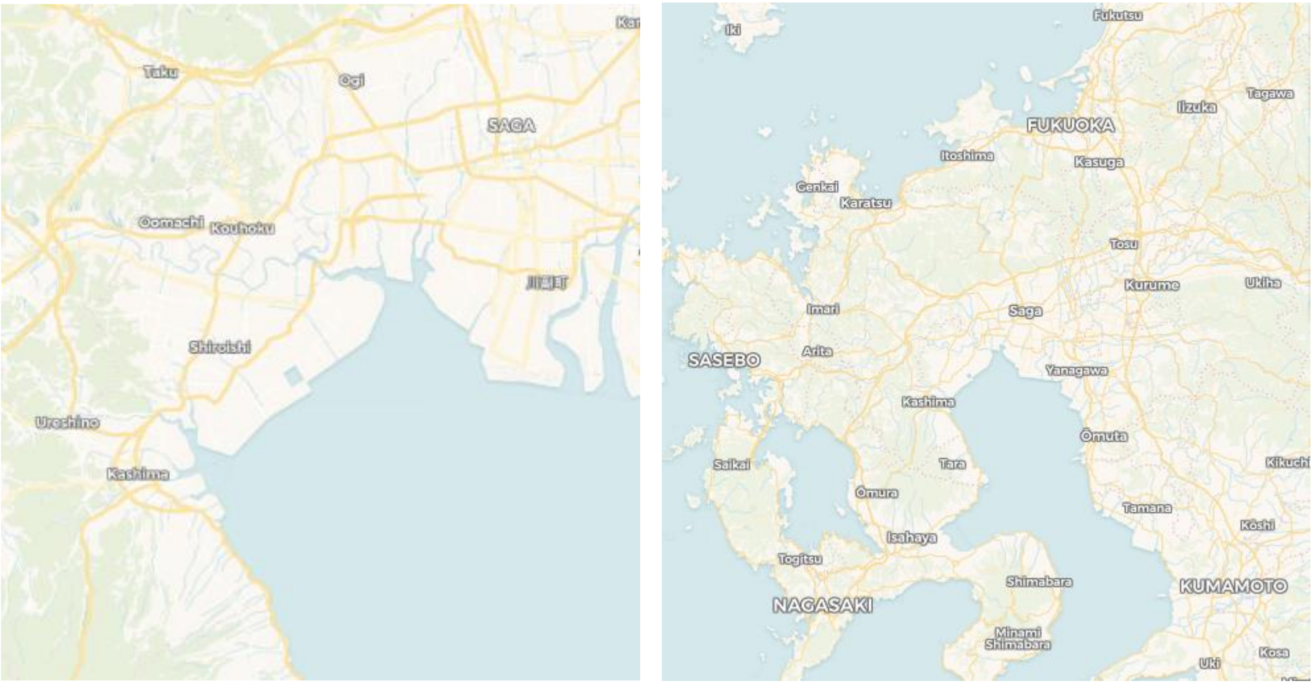
As shown in Fig. 6(a), rice paddy fields are situated in the coastal areas in particular, and entire Saga prefecture. In particular, sea breeze, sea salt damage are occurred in the coastal areas and the areas along with rivers also rice lodging is occurred in the rice paddy fields situated in all over the Saga prefecture. Therefore, the intensive study area is selected.

B. Data used

Sentinel-1 of SAR data is used for detection of oil spill and collapsed area detection. There are two Sentinel-1 satellite, 1A and 1B. Both of repetition cycle is 12 days. Therefore, it is possible to observe the earth surface every 6 days. Also, there are two polarization of available SAR data, VV and VH (V and H stands for vertical and horizontal polarization so that VV means emit V polarization of Electromagnetic Wave: EM (C band) and receive V polarization of EM return echo from the earth surface. Spatial resolution of SAR on the ground is 5 m. Table I shows major specification of Sentinel-1 of SAR.

Meanwhile, Sentinel-2 carries 10 m resolution of visible to short wave infrared radiometer. Table II shows major specification of optical sensor onboard Sentinel-2 Band 12 is Short Wave Infrared SWIR band while band 8 is Near Infrared: NIR band. Also, band 4 is red color band so that Normalized Deviation of Vegetation Index: NDVI and be retrieved with the following equation,

$$NDVI = (B8 – B4) / (B8 + B4) \quad (1)$$



(a) Intensive Study Area (Saga).

(b) Northern Kyushu.

Fig. 6. Location of Saga Prefecture.

TABLE. I. MAJOR SPECIFICATION OF SENTINEL-1 OF SAR

Stripmap	80 km	5 m × 5 m	HH-HV, VV-VH

TABLE. II. MAJOR SPECIFICATION OF OPTICAL SENSOR ONBOARD SENTINEL-2

B1	443 nm	60 m
B2	490 nm	10 m
B3	560 nm	10 m
B4	665 nm	10 m
B5	705 nm	20 m
B6	740 nm	20 m
B7	775 nm	20 m
B8	842 nm	10 m
B8a	865 nm	20 m
B9	940 nm	60 m
B10	1375 nm	60 m
B11	1610 nm	20 m
B12	2190 nm	20 m

Also, SWIR color composite image can be derived from Band 12, Band 8A and Band 4 while false color composite image can be derived from Band 8, 4, and 3 where Band 3 is green color band.

C. Experimental Results

Fig. 7(a) and (b) show the false color and NDVI images of the intensive study area acquired at 02:08:20UTC on 14 September 2019 (just before the typhoon #17).

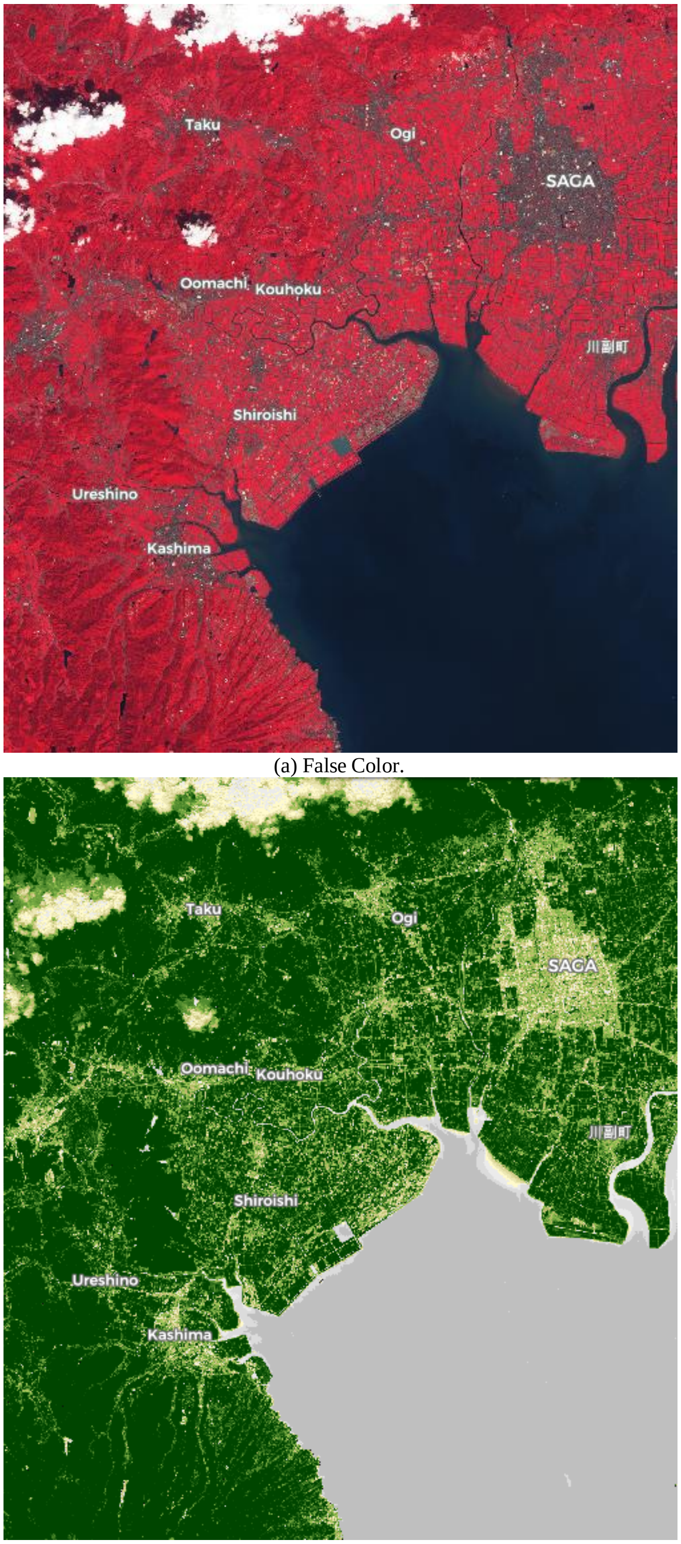
Although Sentinel-2 observed the intensive study area on 14 of September, the data acquired on 17, 19, 22, 27, 29 of September are covered with clouds. After the typhoon #17, partially cloudy scene was acquired on 4, 9 and 14 of September. Good Sentinel-2 data were not acquired on 7 and 12 of September. On the other hand, Fig. 8 shows the false color and NDVI images of the intensive study area acquired on 4 of October (just after the typhoon #17). In the figure of NDVI, color scale is as shown in Fig. 9.

Meanwhile, Fig. 10 shows the false color and NDVI images of the intensive study area acquired on 10 of October.

Mean and standard deviation of false colored image of the paddy fields in the intensive study area acquired on 14 September are 175.15 and 42.73, respectively. Meanwhile, those for 10 October are 149.14 and 45.56, respectively. Therefore, vitality of the rice leaves is degraded by 17.43 % due to the typhoon #17. In other word, 17.43 % of paddy fields in the intensive study areas are degraded due to sea salt damage and sea breeze as well as rice lodging. Fig. 11 shows subtracted image between the false images between 14 September and 10 October. These red colored areas are damaged areas due to the typhoon #17.

Table III shows trend of the percentage ratio of the averaged NDVI of the rice paddy fields in the intensive study area derived from Sentinel-2 of satellite imagery data. The ratio is defined as the ratio between the averaged NDVI of September 14 and the other NDVI. The ratio of September 14 (just before the typhoon #17) is much greater than the others.

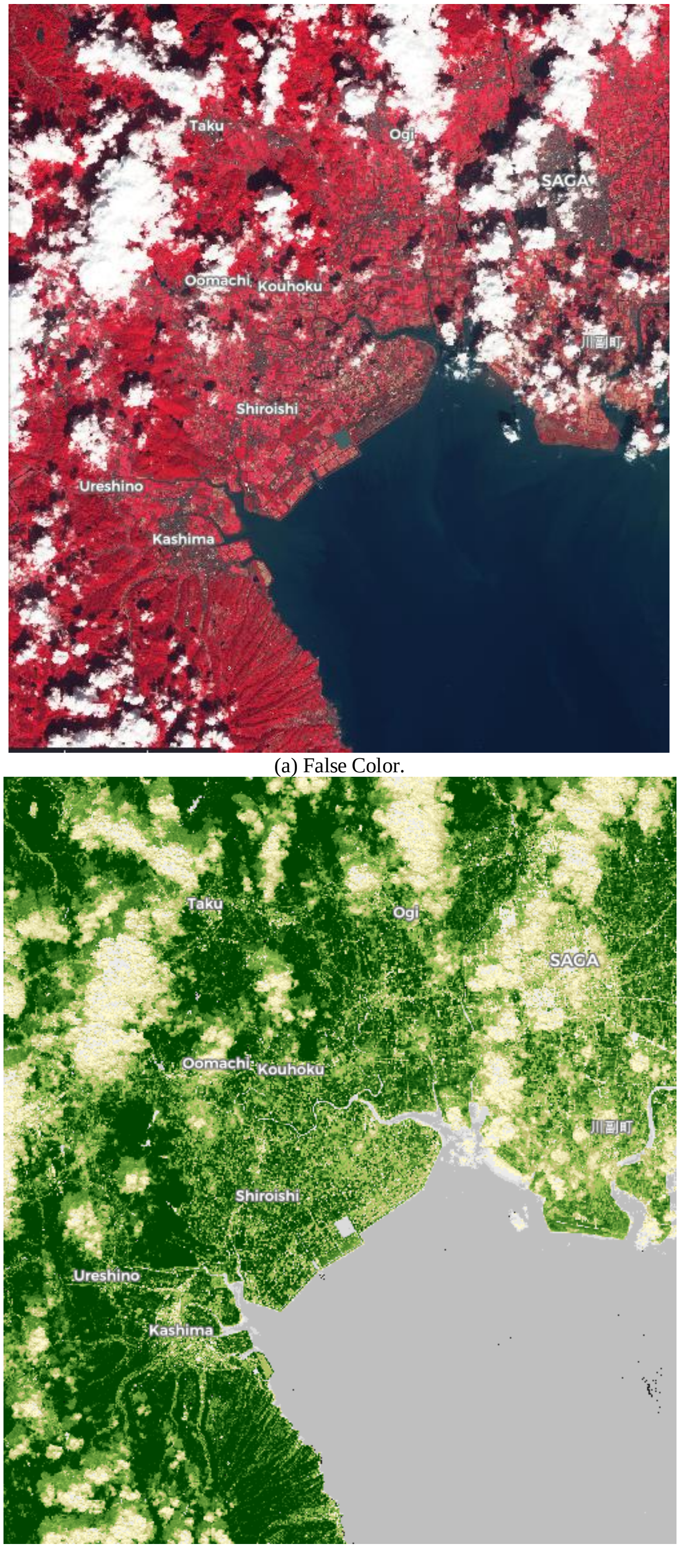
The ratio of October 4 (just after the typhoon #17) is getting down shapely. Then it is going down gradually on October 9 and 14. This is the well-known feature of the sea salt damage of rice leaves. Namely, NDVI is getting down in accordance with elapsed time duration due to sea salts.



(a) False Color.

(b) NDVI.

Fig. 7. False Color and NDVI Images of the Intensive Study Area Acquired at 02:08:20UTC on 14 September 2019.



(a) False Color.

(b) NDVI.

Fig. 8. False Color and NDVI Images of the Intensive Study Area Acquired on 4 of October.

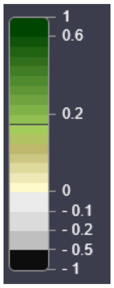
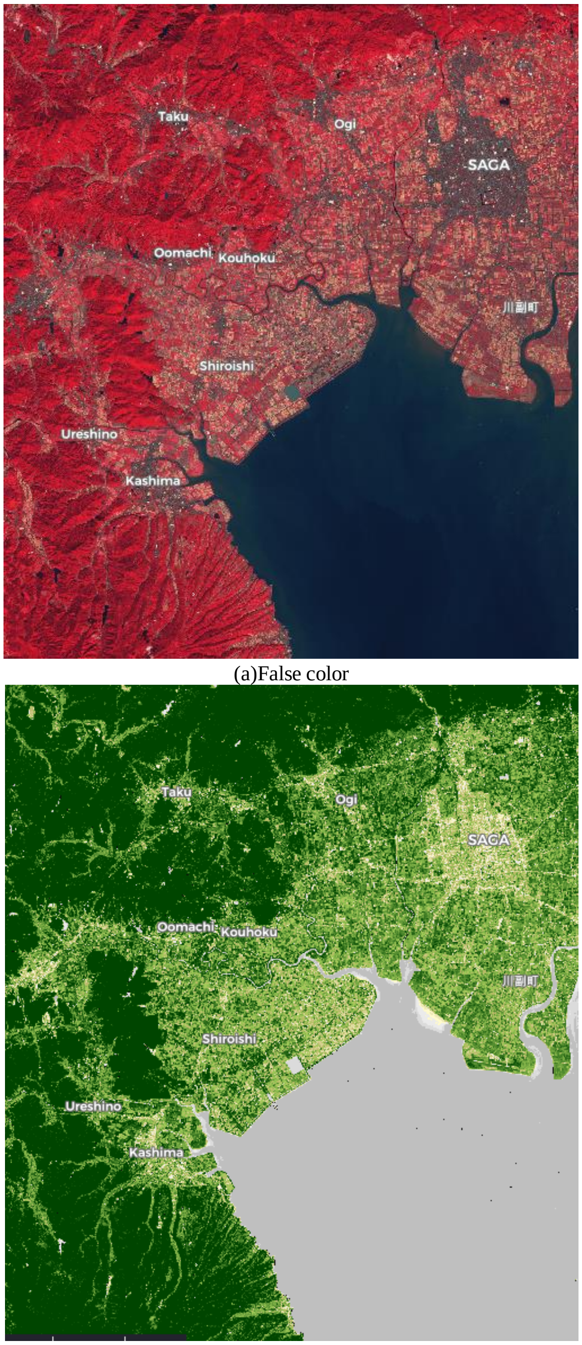


Fig. 9. NDVI Color Scale.



(a)False color

(b)NDVI

Fig. 10. False Color and NDVI Images of the Intensive Study Area Acquired on 10 of October.

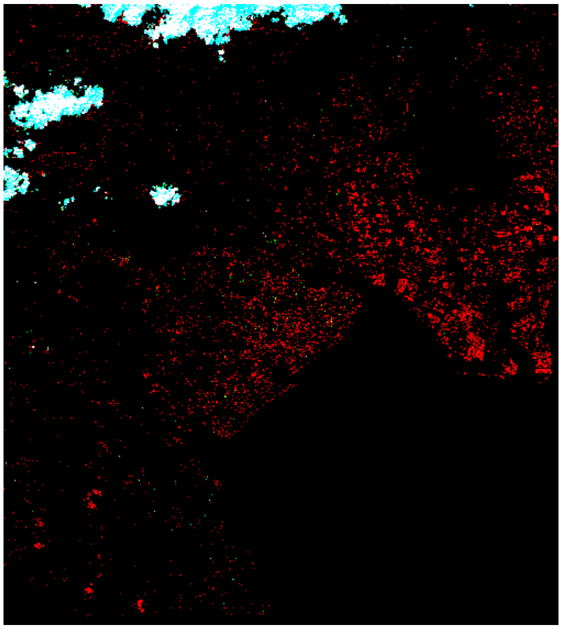


Fig. 11. Subtracted Image between the False Images between 14 September and 10 October.

TABLE. III. TREND OF THE PERCENTAGE RATIO OF THE AVERAGED NDVI OF THE RICE PADDY FIELDS IN THE INTENSIVE STUDY AREA DERIVED FROM SENTINEL-2 OF SATELLITE IMAGERY DATA

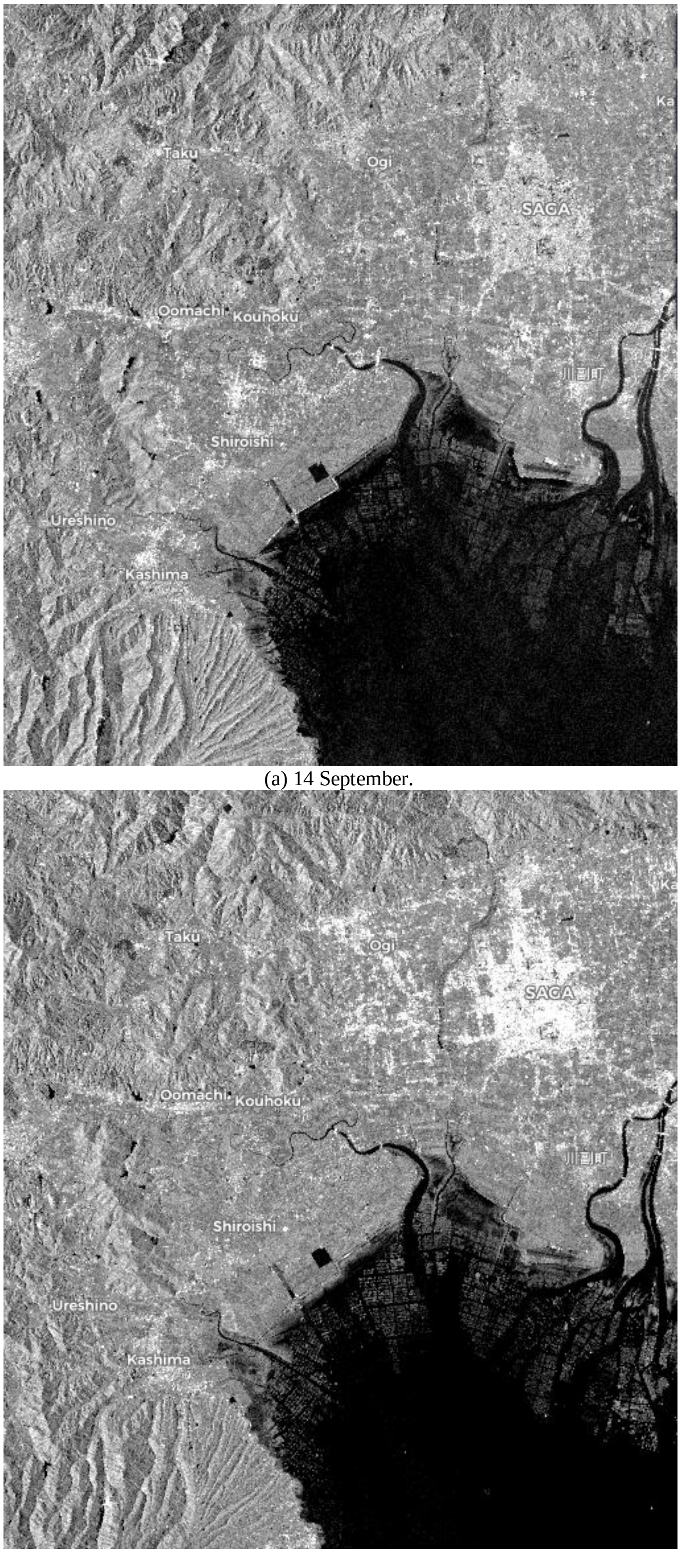
September/14	-
October/4	30.99%
October/9	34.13%
October/14	45.63%

Damaged rice paddy fields are mostly situated in the coastal areas and are partially situated in the entire the intensive study area as well.

In particular, rice lodging can be detected with SAR data because backscattering coefficient of rice lodging areas is decreased theoretically. Fig. 12(a) and (b) shows ortho rectified VV sigma note (back scattered cross section of the earth surface) in unit of decibel of the areas which are acquired on 20 September (just before the typhoon #17) and 1 October 2019 (just after the typhoon #17), respectively. Also, Fig. 13 shows the subtracted image between ortho rectified VV sigma note (back scattered cross section of the earth surface) in unit of decibel of the areas which are acquired on 20 September and 1 October.

Mean and standard deviation of the rice paddy fields in the intensive study area are 151.45 and 25.99 for 20 September and are 150.48 and 21.85 for 1 October, respectively. Therefore, 0.64 % of back scattering coefficient decreasing is observed between before and after the typhoon #17.

Due to the fact that the wind direction in the intensive study are is from south, it is a blowing wind. Therefore, rice lodging is not occurred. It would be significant when the wind direction is from north because it means that a blown down wind. Therefore, rice lodging would occur. It is obvious that 0.64 % is not significant. Therefore, it is true that the rice lodging is not significant at all.



(a) 14 September.

(b) 10 October.

Fig. 12. Ortho Rectified VV Sigma Note (Back Scattered Cross Section of the Earth Surface) in unit of Decibel of the Areas.

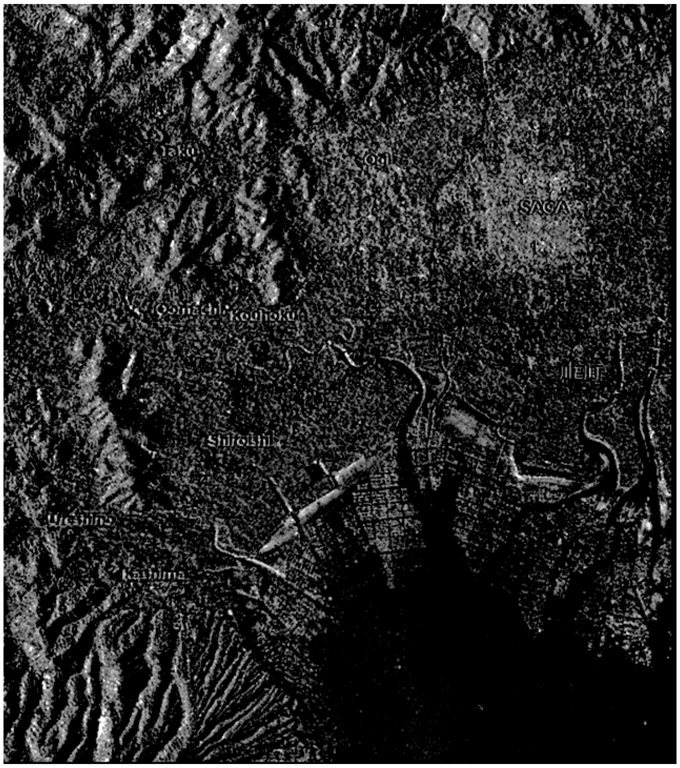


Fig. 13. Subtracted Image between Ortho Rectified VV Sigma Note (Back Scattered Cross Section of the Earth Surface) in unit of Decibel of the Areas which are Acquired on 20 September and 1 October.

VI. Conclusion

Sea breeze and sea salt damage disaster relief using Sentinel of remote sensing satellite data is conducted. Kyushu, Japan had severe storm due to the typhoon #17 during from 21 September to 23 September 2019. Optical sensor and Synthetic Aperture Radar: SAR onboard remote sensing satellite is used for disaster relief. NDVI and false colored imagery data derived from the Sentinel-1 and 2 data are used for disaster relief. Merits and demerits of the optical sensor and SAR instrument are compared from the disaster relief of point of view. It is found that sea salt damage on rice paddy fields in particular can be relieved by NDVI and false colored imagery data while rice lodging can also be relieved by SAR data.

It is found that vitality of the rice leaves is degraded by 17.43 % due to the typhoon #17. In other word, 17.43 % of paddy fields in the intensive study areas are degraded due to sea salt damage and sea breeze as well as rice lodging. It is also found that 0.64 % of back scattering coefficient decreasing is observed between before and after the typhoon #17. It is obvious that 0.64 % is not significant. Therefore, it is true that the rice lodging is not significant at all.

VII. Future Research Works

Further experimental studies are required for the validation of the proposed method. Also, applicability of the proposed method has to be confirmed through further experiments.

ACKNOWLEDGMENT

The author would like to thank Professor Dr. Hiroshi Okumura for his valuable discussions.

REFERENCES

[1] Kohei Arai, Method for estimation of damage grade and damaged paddy field areas sue to salt containing sea breeze with typhoon using remote sensing imagery data, International Journal of Applied Science, 2, 3, 84-92, 2011.

[2] 135. Kohei Arai, Four Dimensional GIS and Its Application to Disaster Monitoring with Satellite Remote Sensing Data, Proceedings of the Conference on GIS and Application of Remote Sensing to Disaster Management, 132-137(1997).

[3] 151. Kohei Arai, The Conference on GIS and Application of Remote Sensing to Disaster Management Four Dimensional GIS and Its Application to Disaster Monitoring with Satellite Remote Sensing Data, Proceedings of the Conference on GIS and Application of Remote Sensing to Disaster Management, 132-137 Greenbelt, Maryland, U.S.A., 1997.

[4] Kohei Arai, URL search engine with text search tools for disaster mitigation, Proceedings of the Asian Disaster Reduction Center R&D Project Workshop, Mar.3, (2000).

[5] Kohei Arai, Visualization of disaster information derived from Earth observation data, Proceedings of the Asian Disaster Reduction Center R&D Project Workshop, Aug.31, (2000).

[6] 266. Kohei Arai, ICT technology for disaster mitigation,-Tsunami warning system-, Proceedings of the 1st International Workshop on Knowledge Cluster Systems, 2007.

[7] 318. Kohei Arai, Cellular automata approach for disaster propagation prediction and required data system in GIS representations, Proceedings of the 1st ICSU/WDS Conference - Global Data for Global Science, 2011.

[8] Kohei Arai, Cell based GIS as Cellular Automata for disaster spreading prediction and required data systems, CODATA Data Science Journal, 137-141, 2012.

[9] Kohei Arai, Hiroshi Okumura, Shogo Kajiki, Disaster relief with satellite based synthetic aperture radar data, Proceedings of the SAI Future Technology Conference 2017, No.521, 1026-1029, at Vancouver, 2017.

[10] Kohei Arai, Sentinel 1A SAR Data Analysis for Disaster Mitigation in Kyushu, Kyushu Brunch of the Japanese Society on Remote Sensing, Special Lecture for Young Engineers on Remote Sensing, Nagasaki University, 2018.

[11] Kohei Arai, Sensor network for landslide monitoring with laser ranging system avoiding rainfall influence on laser ranging by means of time diversity and satellite imagery data based landslide disaster relief, International Journal of Applied Sciences, 3, 1, 1-12, 2012.

[12] Kohei Arai, Flooding and oil spill disaster relief using Sentinel of remote sensing satellite data, International Journal of Advanced Computer Science and Applications IJACSA, 10, 12, 290-297, 2019.

AUTHOR'S PROFILE

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

JEPPY: An Interactive Pedagogical Agent to Aid Novice Programmers in Correcting Syntax Errors

Julieto E. Perez[1], Dante D. Dinawanao[2], Emily S. Tabanao[3]
Department of Computer Science, College of Computer Studies
MSU–Iligan Institute of Technology, Iligan City Philippines

***Abstract*—Programming is a complicated task and correcting syntax error is just one among the many tasks that makes it difficult. Error messages produced by the compiler allow novice learners to know their errors. However, these messages are puzzling, and most of the times misleading due to cascading of errors, which can be detrimental to running a syntax-error free program. In most laboratory setting, it is the role of the teachers to assist their students while doing activities. However, in our experienced, considering the large number of students in a class, it may seem difficult for teachers to assist their students one-by-one given the time constraints. In this paper, the design and implementation of an interactive pedagogical agent named JEPPY is presented. It is intended to assist novice learners learning to program using C++ as a programming language. In order to see on how students struggle or progress in dealing with errors, the proponents implemented the Error Quotient (EQ) developed by Jadud. The principles of the cognitive requirements of an agent-based learning environment were followed. The agent was put into test by novice learners in a laboratory setting. Logs of interaction between the embodied agent and the participants were recorded, aside from the compile errors and edit actions. These mechanisms show us some insight on the interaction behavior of learner to the agent.**

Keywords—Pedagogical agent; error quotient; syntax-error correction; compile errors; human computer interaction

I. Introduction

Computer programming is a complicated task. According to Jenkins [1], Programming "is a complicated business" which requires the mastery of several skills such as problem solving, abstraction, mathematical logic and testing, debugging and so forth. In addition to this, in case of novice programmers, knowledge was found to be limited and shallow hence they lack the ability to write syntactically-correct programs. Over the years, several studies have been conducted to look at how compiler errors have affected the learning curve of students learning to program, particularly, novice programmers. Becker [2] showed that compiler errors can be frustrating and students in his study described them as "barriers to progress". In addition, Denny, Luxton-Reilly and Tempero [3] showed that students have difficulties locating and correcting syntax errors using average compiler. Moreover, Kummerfeld and Kay [4] concluded that even the more experienced students took significant time to correct some syntax errors. Studies have been conducted to understand interaction of learners to the compilers. Separate studies of Jadud [5] and Becker [6] showed a metric in quantifying these repeated errors. Jadud [5] called the interaction of the learners to the compiler as "compilation behaviour" and called the metric as the Error Quotient. To support learners in dealing cryptic messages produced by compiler, Ahmed, Kumar, Karkare, Kar, and Gulwani [7] developed a system called TRACER (Targeted RepAir of Compilation Errors) that perform repairs on compilation errors. In a study of Becker, Goslin, and Glanville [8], an enhancement to JAVA compile error messages was made and employed for intervention. Comparison between control and intervention groups showed that enhancing compiler messages is of advantage.

For environment of practice by novice programmers such as that in a laboratory setting, the current methods and tools employed focused on identifying behaviors using the online protocols and browser to inform teachers who among their students struggles and then provide manual intervention if necessary. However, considering the number of students in a classroom it is not realistic that the teacher can always assist the entire class for every laboratory session given some time constraints. This issue motivated the researchers in this study to come up another approach to augment the problem.

This study made an attempt in employing an embodied agent and see its potential use to aid novice programmers in their battle over syntax errors. This can help mentors attend to several other skills to consider in teaching programming, rather than focusing on assisting compile errors correction. However, skills like problem solving and logical reasoning were not yet addressed in this study and learning on that aspects requires different measures to help novice learners.

In this paper, the proponents presented the design and implementation of an interactive pedagogical agent which will be used as a tool to assist novice programmers in the daunting task of correcting syntax error produced by the compiler. Moreover, the proponents look into the interaction of the learner to the agent along with their interaction to the compiler. This can give us insights on the improvement of the agent and to the target benefit at which the agent was employed.

II. Review of Related Work

There were studies that looked at how novice learners interact with the compiler while practicing programming. Jadud [9] define novice compilation behavior as the study of students' interaction with their compiler while learning to program. In his study, Jadud [3] developed a quantification of the student's compilation behavior based on grounded theory. He called it the error quotient or EQ. Every record in the data logs represents one compilation event. Stored in each record is the error message if there was an error at the time of compilation, the location of the error in the file which is reported by the compiler as a line number, and the source code. An EQ score is of the range 0 to 1.0, where 0 is a perfect score. An EQ score of 0 does not mean that the student made no syntax errors in their programming process. What it means is that at no point did the student encounter the same syntax error consecutively. Whereas a session scoring 1.0 means that every compilation resulted to the same syntax error all the time.

Agapito and Rodrigo [9] looked into students' compilation behaviors as they wrote their programs in C++ by analyzing automatically collected online protocols. Students' data were analyzed by computing for their Error Quotient. Results confirmed that freshmen programmers do experience difficulties and that the Error Quotient is a practical tool that can be used to characterize their compilation behaviors.

Many of the programming environment or IDEs used today have embedded capabilities or features added to help programmers do their job easily instead of just writing it using plain text editors. This same IDE is also used by novice programmers in their first programming experience using specific language. Many works reported development of automated syntax error correction. However, the approach does not care whether learners have assisted their own mistake.

Some works produced feedback through an interface where support is provided. Carter [10] developed an intelligent tutor for debugging called ITS-Debug. This is achieved by developing a system with four standard modules (Domain, Student, Pedagogical, Communication) of Intelligent Tutoring System. A web-based system was developed wherein students learn debugging through different phases. Students were able to edit the code, compile and run the code, and receive assistance on a host of syntax, runtime, and logical defects that might be present in the exercise or that they may inadvertently create themselves. In the study of Kummerfeld and Kay [4], a web-based reference guide was developed which catalogues some common C++ compiler generated errors.

So far, the work of Edwards, Rajagopal, Kandru [11] reported the use of embodied characters that assist learners in dealing syntax. The proponents developed an emotionally-intelligent pedagogical agents to deliver effective and efficient feedback to students about their programming assignments and also act as a teaching assistant for any general programming related queries. The main objective of their study is to communicate clearly the feedback about student programs while motivating them to perform better. This is so far, the work that was closely related in this study.

Veletsianos and Russell [12] defined pedagogical agents as anthropomorphous virtual characters employed in online learning environments to serve various instructional goals. Pedagogical agents were employed by Carlotto and Jacques [13], Kim [14], Liew, Zin, and Sahari [15], and Kim, Thayne, and Wei [16] in a form of an animated characters, virtual or digital characters. It was used as a chatbot as reported by Savin-Baden, Tombs, and Bhakta [17], an influencer such as of Kim and Baylor [18], or a tutor Kim [14]. They can also simulate conversations and nonverbal behavior according to Liew and Tan [19]. In the work of Schroeder, Romine, and Craig [20], pedagogical agent was employed to enhance student learning. Johnson and Lester [21] cited a nonverbal feedback capability of pedagogical agents. The nonverbal cues can take various forms including nodding or shaking the head, facial expressions such as smiling or surprise. This paper employs the use of nonverbal cues for the embodied agent and used the agent as an assistant.

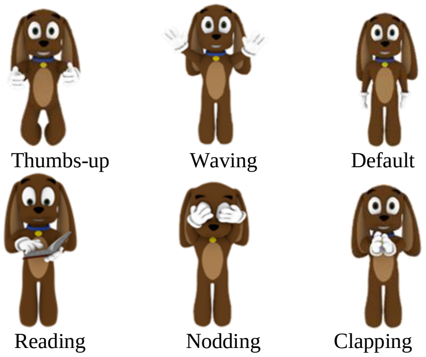
III. Methodology

A. Defining Agent Design Requirements

According to Baylor [22], the prime cognitive consideration in the design of agent-based learning environment is the management of control. The first dimension of control involves instantiating the instructional purpose of the environment on a constructivist (high learner control) to instructivist's (high program/agent control) continuum. A critical issue from a constructivist approach to agent-based learning environments is in moderating between the agents taking over thinking for the student with the agent training the student to think more effectively. In the constructivist approach, the agent is a medium that does not teach the student directly. In this study, the presentation of knowledge about errors comes in a form of recall and example. Note that in an error message, the compiler may refer to some token in the code. Meaning, different token may appear even for the same error type. For example, the error message "expected ';' before 'int'" contains the token ';' and 'int' enclosed within single quote. In recall, the content presented by the agent will not specifically tell the student the specific solution but instead present the similar or general case. For instance, for the error mentioned above, the agent would say "Remember that in C++ every statement must end with a semi- colon. In an example, the agent would present an example statement with a semi-colon at the end. This is how the proponents push the student to do the thinking. The second dimension of control entails managing feedback, and several issues need to be considered: type, timing, amount, explicitness, and learner control of agent feedback. An important consideration in terms of feedback is that the pedagogical agent should not provide too many insights and thereby annoy the student. In the current design of intervention, the agent will depend on the current computed value of the error quotient. This means that whenever the student is stuck in a specific error, the agent will intercept every compilation. Although by default help should be minimal, part of our intention is to give us insight on the interaction of the learner to the agent in the environment. So, the proponents allow the agent to be proactively intervening as long as the EQ limit of 0.3 or greater was reached. Third consideration is when agent versus learner control is further defined through the desired relationship of the learner to agent. Some examples of instantiating the learner-agent relationship include the agent as learning companion, agent as mentor, multiple pedagogical agents, agent as personal assistant, or agent as resource. In this paper, the proponents define the role of the agent to be an assistant that informs the learner on their mistake. The feedback flows from agent looking at the error quotient and appears when EQ is greater than the value 0.3. The agent looks only on the first error message per compilation since the first error most of the time is the cause of cascading error and if not eliminated will cause the student to get stuck on that error. This is also consistent to the existing computation of error quotient in which only one error was considered for computation in every compilation. Fourth, to be instructionally effective, the agent must assert enough control so that the learner develops confidence in the agent in terms of believability, competence, and trust. The critical issue that concerns believability is the message that the agent will provide. Incorrect message to provide will decrease trust and competence. The persona and behavior of the agent was also considered to make the agent believable.

B. The Agent Persona

Fig. 1 shows the Agents' gestures. Sequencing these gestures make up some form of behavior. The choice of the interface is a cartoon character and was named JEPPY. The proponents choose not a very serious character to capture the attention of the serious learners. The behavior space includes deictic and affective gestures as shown in the Fig. 1. The following gestures were combined to form actions that make the embodied agent more life-like.



Thumbs-up Waving Default

Reading Nodding Clapping

Fig. 1. Deictic and Affective Gestures of JEPPY.

C. Testing

To test and validate the functionality of the components, the proponent put JEPPY in to a test with participants in an actual laboratory session. Participants were students taking up introductory programming course in a State University. Before the participants continue in the task, they were given questionnaire to verify whether they are really novice programmers. This is because the agent is intended for novice programmers only. There are 18 participants which where identified to be novice programmers. They were given a source code which contains cascading errors. Meaning, one error may come after another after correcting the first one. They were all given the freedom and time to finish the problem without asking help from other participants or instructor around.

D. The Architectural Design

The implementation follows the typical architecture of a pedagogical agent but was contextualized according to purpose of used. Fig. 2 shows the architecture of the agent in this study. The pedagogical module was implemented as plugin in Code::Blocks. The errors produced by the compiler were preprocessed to include only necessary information. The event logger was responsible on logging the preprocessed compiler errors, the edits done in the code, the interaction of the learner with the agent and the calculated value of error quotient. These data logged by the logger were inserted in the SQLite database. The communication module implemented using Java comprises the interface and inference controller. The interface is where the learner interacts with the agent. The embodiments are gif files which are retrieved depending on the interaction and current state of the learner.

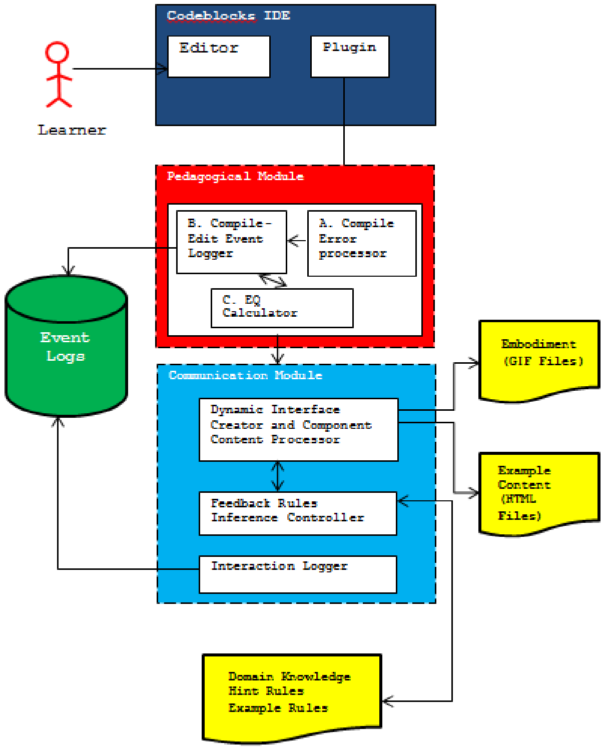


Fig. 2. Architectural Design.

Recalls which can be interchangeably call as hint and examples were written as an html files, which can then be viewed in the interface. The interface contains a balloon tip which is an open source program written in java. Html files which are retrieved from the domain module were displayed inside the balloon tip. The inference controller is responsible for retrieving knowledge during intervention. This part of the implementation connects the pedagogical module and the domain module. The knowledge on the errors was written in CLIPS as rules in an if-then format.

IV. Results and Discussions

A. The Implemented Agent

In the first compilation, the agent would appear and introduce itself to the student. Starting from the first compilation also, the logger is activated. So, every time the learner edits some lines in the code it will be recorded line per line. For every compilation starting from the third compilation, the two pairs of events can be created. At this point the Error Quotient can be calculated. When EQ is more than the threshold value, the agent will capture the first error, preprocessed it and retrieve message from the rules in the domain module that matches the error, and then display the help message through the embodied agent.

Table I shows an example EQ computation extracted from the compile-edit log. As per algorithm, the task is to compare two successive compilations. For instance, from Table I, looking at compilation number 2 and 3, both compilations ended with error, so a penalty of 2 was added. Since both compilations have same error type (expected token before token), a penalty of 3 was added. However, both compilations do not have same error location and line edit made, so no penalty was added. The total score for this pair (compilation 1 and 2) is 5. The total score was divided to 11, which is the highest possible score, and is now the normalized value 0.5556. The final error quotient for this pair is the average of the sum of all the normalized score in each pair, in the given example, it is 0.3889.

The implemented agent was shown in Fig. 3 to Fig 9. Fig. 3 shows the appearance of the agent when it offers help from the learner. As one can see, the agent does not provide directly the help on the error identified. Instead, an option was given to see whether help is needed, or the learner already knows the error. When help is used, the agent will then provide the help as shown in Fig. 4. Fig. 5 shows the case wherein the error occurred again, and the agent will offer another help. Fig. 6 is the screenshot of the agent portraying like reading some notes when telling student to use example.

When help is used again, help will be provided in a form example as shown in Fig. 7. Fig. 8 and Fig. 9 are the affective gestures of JEPPY when it is sad and glad, respectively.

B. Result of Interaction based on the Logs

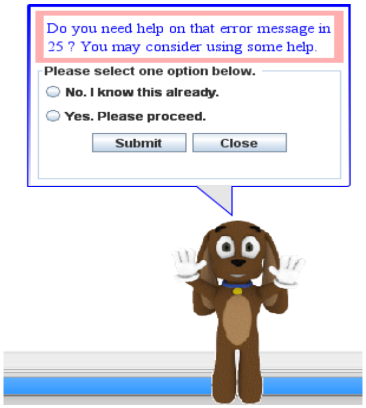
One critical part among the components is the correct message or support that the agent will provide. The interaction log provides a way for us to see whether correct help is given to an error message. Recorded in every row was the error message which is a result in preprocessing stage during compilation. Also, in the same row, was the help coming from the domain knowledge which is a result of the inference engine.

Aside from validating the functionality of the components through the logs, it also gives us some observations on the interaction of the learner to the agent. Out of 538 times that the agent appears, only 159 or 29.55% of the time the agent was used. It can also be observed that there are 119 or 22.12% of the time the agent was closed when help is asked. The proponents can also see instances wherein there is no interaction in an intervention, meaning the agent was ignored and after 20 seconds without any interaction it pops out. There are 260 or 48.33% of the time that the agent was ignored. The large number of time that the agent was ignored by students is maybe because they were so engaged in attempting to correct error by themselves. As mentioned by Jadud [3], students took significant time editing and compiling their code, and after several attempts without success, they may fall into frustration. But here, with the presence of JEPPY, we can be able to prevent such case. We can see that in the sequence of usage. From 159 interventions, 106 or 66.7% were hint usage and 53 or 33.33% were example usage. Even the students are proactively debugging these errors by themselves and do not use help even when they need it, based on the logs, out of the 106 hints usage, 70 or 66.04% of the time wherein errors were encountered are corrected after using hint. When error was not eradicated, the agent can reinforce this by offering an example. We see that there are 11 instances in the total usage wherein hint is immediately followed by example and the error was corrected after it. There is a total of 81 or 76.42% of errors corrected after using the support provided. In case of example usage only, meaning not preceded with hint, there is 56.60% of the total usage wherein the error was corrected right after.

Although the figures presented are not at large, the potential of JEPPY can be seen in helping the novice learners in dealing syntax error, of course, with further improvement.

TABLE. I. SAMPLE ERROR QUOTIENT CALCULATION

Compi lation no.	Error message	Error message type	Error locati o -n	Both event - s end with error	Same error type	Same error locati o n	Same edit locati - on	Pair no	score	Norma -lized score	Sum of normali -zed score	Error quotient
1	‘ans’ was not declared in this scope	error was not declared in this scope	55									
2	Expected ‘while’ before ‘cin’	expected token before token	29	2	0	0	0	1	2	0.2222	0.2222	
3	Expected ‘;’ before ‘endl’	expected token before token	50	2	3	0	0	2	5	0.5556	0.7778	0.3889
4	Expected ‘}’ before ‘else’	expected token before token	52	2	3	0	1	3	6	0.6667	1.4444	0.4815



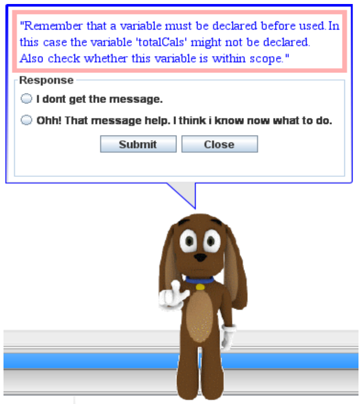
Do you need help on that error message in 25 ? You may consider using some help.

Please select one option below.

○ No. I know this already.

○ Yes. Please proceed.

Fig. 3. JEPPY Offering Help through Hint.



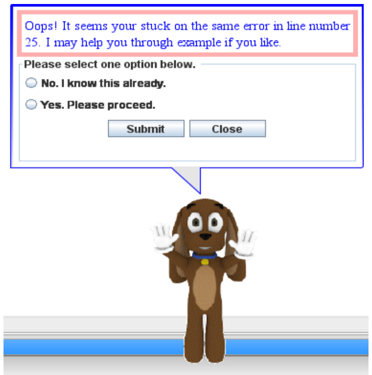
"Remember that a variable must be declared before used. In this case the variable 'totalCals' might not be declared. Also check whether this variable is within scope."

Response

○ I dont get the message.

○ Ohh! That message help. I think i know now what to do.

Fig. 4. JEPPY Showing Hint.



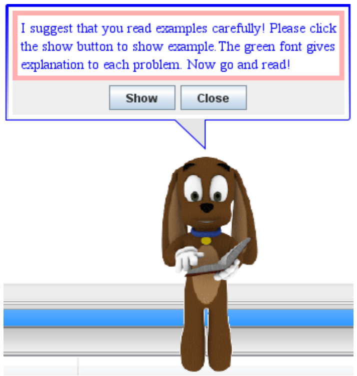
Oops! It seems your stuck on the same error in line number 25. I may help you through example if you like.

Please select one option below.

○ No. I know this already.

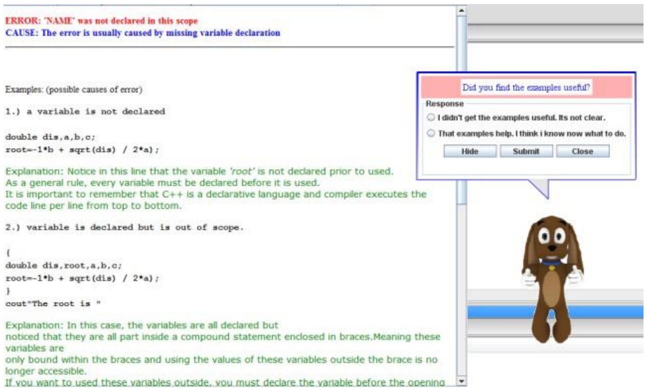
○ Yes. Please proceed.

Fig. 5. JEPPY Offering Help through an Example.



I suggest that you read examples carefully! Please click the show button to show example. The green font gives explanation to each problem. Now go and read!

Fig. 6. JEPPY when Instructing to Read Help Carefully.



ERROR: 'NAME' was not declared in this scope
CAUSE: The error is usually caused by missing variable declaration

Examples: (possible causes of error)

```
1.) a variable is not declared

double dis,a,b,c;
root=-1*b + sqrt(dis) / 2*a);
```

Explanation: Notice in this line that the variable *'root'* is not declared prior to used.
As a general rule, every variable must be declared before it is used.
It is important to remember that C++ is a declarative language and compiler executes the code line per line from top to bottom.

```
2.) variable is declared but is out of scope.

{
double dis,root,a,b,c;
root=-1*b + sqrt(dis) / 2*a);
}
cout"The root is "
```

Explanation: In this case, the variables are all declared but
noticed that they are all part inside a compound statement enclosed in braces.Meaning these variables are
only bound within the braces and using the values of these variables outside the brace is no longer accessible.
If you want to used these variables outside, you must declare the variable before the opening

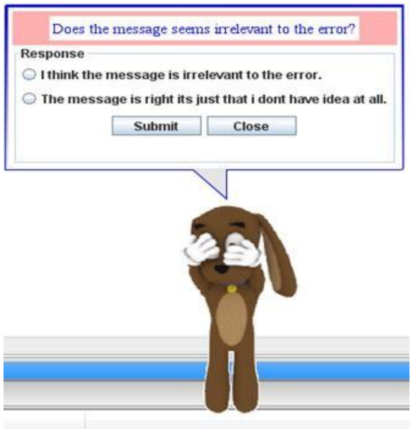
Did you find the examples useful?

Response

○ I didn't get the examples useful. Its not clear.

○ That examples help. I think i know now what to do.

Fig. 7. JEPPY Showing Example to an Error.



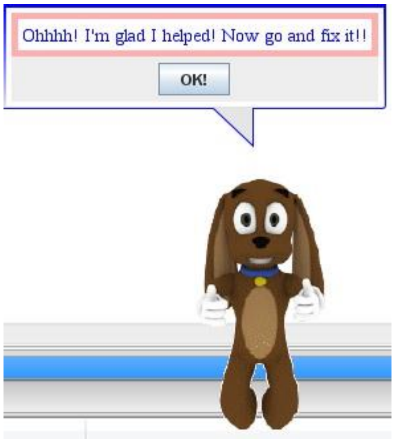
Does the message seems irrelevant to the error?

Response

○ I think the message is irrelevant to the error.

○ The message is right its just that i dont have idea at all.

Fig. 8. JEPPY when Help was Ignored or not used.



Ohhhh! I'm glad I helped! Now go and fix it!!

Fig. 9. JEPPY when Help was used and Error was Corrected.

To see whether content in the support is helpful, part of the interaction by the agent is to ask the learner whether the message is clear or helpful. There are 66 or 41.51% instances wherein student responded on the question whether hint is clear or understandable. From the total responses, all 66 of it responded that the message is clear. For the example usage, however, there is only one response which said that the message is clear.

The summary of our logs had given us insight in terms of interaction. In the design, the agent was intended to be proactive by having smaller threshold value of Error Quotient. But our logs tell us that the agent must be designed to carefully select timing in intervention, otherwise, the learner might get annoyed. Probably models on interaction along with EQ should be developed for timing in intervention. Nevertheless, when help is being used, the agent can be of help before the learner falls into frustration. However, it should be noted that our logging mechanism was not intended to deeply look on the efficacy of learning. The logs enable us to verify and validate the functionality of every component and give us opportunity to gain insight for further improvement of the agent.

V. Conclusion and Future Work

In this paper, the researchers presented the design and implementation of an interactive pedagogical agent. It was successfully embedded as a plugin in an Integrated Development Environment named Code::Blocks. The said environment for developing real-world applications was also used by the students in our institution. However, it was not developed to care on the problem encountered by Novice programmers such as syntax error correction. Hence, through this work, the researchers were able to address one of the many problems a Novice programmer may encounter.

Although our domain is specific to C++ as programming language, the modular fashion of the architectural design on the components can be easily expanded. For instance, rules containing errors and their corresponding help or corrections can be added without any changes in the rule engine as long as it conforms to the pattern. Currently, the study does not include yet the evaluation on the learning gain. It can be seen, however, that by using the computed EQ, one can determine how well a student is progressing with or without JEPPY. This can be done with a large number of participants and an ample time. The current work done focuses on the design and implementation of the agent and the EQ.

References

[1] T. Jenkins, “On the difficulty of learning to program,” 3rd Annual LTSN -ICS Conference. University of Ulster, LTSN Centre for Information and Computer Sciences, 2002.

[2] B.A. Becker, “An effective approach to enhancing compiler error messages,” In Proceedings of the 47th ACM Technical Symposium on ComputingScience Education, 126–131, 2016.

[3] P. Denny, A. Luxton-Reilly, E. Tempero, J. Hendrick, “Understanding the syntax barrier for novices,” ITiCSE In Proceedings of the 16th Annual Joint Conference on Innovation and Technology in Computer Science Education, 208–212, 2011.

[4] S. K. Kummerfeld and J. Kay, “The neglected battle fields of syntax errors,” In Proc. Fifth Australasian Computing Education Conference, 105-111, 2003.

[5] M.C. Jadud, “Methods and tools for exploring novice compilation behaviour,” Proceedings of the 2006 international workshop on Computing education research, pp. 73-84, 2006.

[6] B. A. Becker. “A new metric to quantify repeated compiler errors for novice programmers,” In Proceedings of the 21st ACM Conference on Innovationand Technology in Computer Science Education, pp. 296–301, 2016.

[7] U. Z. Ahmed, P. Kumar, A. Karkare, P. Kar, and S. Gulwani, “Compilation error repair: for the student programs, from the student programs,” In Proceedings of the 40th International Conference on SoftwareEngineering: Software Engineering Education and Training, 78–87, 2018.

[8] B. A. Becker, K. Goslin, and G. Glanville, “The effects of enhanced compiler error messages on a syntax error debugging test,” In Proceedings of the 49th ACM Technical Symposium on Computer Science Education, 2018.

[9] J. L. Agapito, M. M. T. Rodrigo, “An analysis of novice programmers' compilation behaviors in c++,” Philippine Information Technology Journal, 2012.

[10] E. A. Carter, "An intelligent debugging tutor for novice computer science students," Theses and Dissertations, 2014.

[11] S. H. Edwards, M. B.M. M. Rajagopal, N. Kandru, “Pedagogical agent as a teaching assistant for programming assignments: (abstract only),” Proceedings of the 49th ACM Technical Symposium on Computer Science Education, p.1079, February 2018.

[12] G. Veletsianos, G. Russell, “Pedagogical agents,” handbook of research on educational communications and technology, 4th Edition, pp. 759-769, 2014.

[13] T. Carlotto and P. A. Jaques, “The effects of animated pedagogical agents in an english-as-a-foreign-language learning environment,” International Journal of Human Computer Studies, vol 95, pp.15–26, 2016.

[14] Y. Kim, “The role of agent age and gender for middle-grade girls,” Computers in the Schools, vol 33, pp. 59–70, 2016.

[15] W. T. Liew, N. A. M. Zin, and N. Sahari, “Exploring the affective, motivational and cognitive effects of pedagogical agent enthusiasm in a multimedia learning environment,” Human-Centric Computing and Information Sciences, 7(9), 2017.

[16] Y. Kim, J. Thayne, and Q. Wei, “ An embodied agent helps anxious students in mathematics learning,” Educational Technology Research and Development, 65(1), 219–235, 2017.

[17] M. Savin-Baden, G. Tombs, and R. Bhakta, “Beyond robotic wastelands of time: abandoned pedagogical agents and new pedalled pedagogies,” E-Learning and Digital Media, 12(3-4), 295–314, 2015.

[18] Y. Kim, A. L. Baylor, Pedagogical agents as social models to influence learner attitudes,” Educational Technology, 47(1), 23–28, 2007.

[19] W. T. Liew, and S. M. Tan, “Virtual agents with personality: adaptation of learner-agent personality in a virtual learning environment,” 11th International Conference on Digital Information Management, pp. 157–162, 2016.

[20] L. N. Schroeder, W. D. Romine., and S. D. Craig, “Measuring pedagogical agent persona and the influence of agent persona on learning,” Computers and Education, 109, 176–186, 2017.

[21] W.L. Johnson, J.C Lester, “Face-to-face interaction with pedagogical agents, twenty years later” International Journal of Artificial Intelligence in Education, 26(1), 25–36, 2016.

[22] A. Baylor, "Cognitive requirements for agent-based learning environments,", Proceedings of the 2001 International Conference on Advanced Learning Technologies, pp. 462-463, 2001.

Intelligent Haptic Virtual Simulation for Suture Surgery

Mee Young Sung[1]*, Byeonghun Kang[2], Jungwook Kim[3], Taehoon Kim[4], Hyeonseok Song[5]
Department of Computer Science and Engineering
Incheon National University, Incheon, Republic of Korea

***Abstract*—The aim of this study is to develop an intelligent haptic virtual simulation for suture surgery, enabled with an AI assistant. This haptic VR suture simulation mainly composed of three parts: visuo-haptic rendering for surgery, replica training for surgery, and AI assistant for surgery. In the simulation, a trainee surgeon can practice the suturing procedure using tactile touches in the stereoscopic 3D virtual environment. The simulation adopts a "precise haptic collision detection method using subdivision surface and sphere clustering" developed in the previous studies of authors. In addition, an expert surgeon's suture operations can be replicated to the medical trainee's simulation to experience it as it is. This method has the advantage of reproducing the experience of the expert's ideal movements of surgical procedures. The data recording of the skilled surgeon's motions and operations are normalized into a form suitable for AI learning. The AI assistant distinguishes five typical types of suture surgery methods and learns the proper methods for various wounds using deep learning techniques, then it suggests the most appropriate suture method. The suture simulation can reduce the cost and time required for surgical training and eventually provide safe and accurate physical surgery.**

Keywords—Haptic; virtual simulation; suture surgery; artificial intelligence; assistant

I. Introduction

It is predicted that the advancement of AI (artificial intelligence) will lead to the disappearance of the medical profession, but rather the surgical AI techniques will become more popular. The recent technology trend is in the evolutionary combination of VR (virtual reality), AR (augmented reality), MR (mixed reality) and AI. Especially, AI winged haptic surgical simulation has the potential to revolutionize the healthcare industry.

One of the pioneers in the field of virtual surgical simulation is the MSim® [1] of Mimic Technologies. Other prominent products are LabSim®, EndoSim®, and TeamSim®[2] of Surgical Science. They provide a significant level of realistic virtual surgery simulations with haptic feedback. It would be also notable the CardinalSim [3] of Stanford University which is an immersive simulation environment developed for the preoperative rehearsal of complex surgical procedures. The recent trend is to combine artificial intelligence technology with these virtual surgical simulations to make them more convenient and useful. However, those commercial products are very expensive and require specific hardware.

The goal of this study is to develop a low cost and efficient suture surgery simulation system equipped with an intelligent assistant for surgeons to practice sufficiently prior to actual procedures. The system provides a realistic medical training environment for suturing operation in a 3D virtual reality environment. It also allows for experiencing the actual suturing procedure repeatedly using haptic devices. In addition to tactile sensation, the usage of a 3D VR headset gives the visual part feel realistic. Moreover, the trainee can acquire the surgical skills of the expert by reproducing the expert's surgical motions. If the training results are not satisfactory, the trainee can profit the calibration and the feedback from the surgical assistant AI.

This paper will describe the core technology for the development of an intelligent suture surgery training simulation using haptic VR technology. In order to develop a suture simulation system, precise graphics simulations that realistically reflect the physical phenomena applied to the virtual body are required. There is also a need for precise haptic simulations that provide realistic tactile sensation when the surgical tool makes contact with virtual tissue. To deliver smooth, sophisticated tactile feedback to users, an update rate of 1000 Hz or higher should be maintained.

Note that the 3D (3-dimensional) stereoscopic visualization can affect the performance of intuition, accuracy, and immersivenesss of the surgical training simulation [1]. In addition, it would be very effective if a trainee surgeon could acquire the skills of an expert surgeon by mimicking an expert's surgical behavior exactly. Moreover, if the trainee can receive personalized corrections and feedbacks from the AI assistant that helps to learn the suturing, the educational improvement effect will become amazing.

The rest of this paper proceeds with a short survey of related works. Then, a brief description of the haptic virtual simulation for suture surgery is presented. Next, the details of each part of the simulation are elaborated. Lastly, the conclusion and future work are followed.

II. Related Work

Haptics technology allows users to perceive tactile sensations regarding force and movement. Haptics plays an important role in simulations in virtual environments, such as

[1] MSim® [Internet]. 2018. Available from: https://mimicsimulation.com/msim/ [Accessed: 2020-01-25]

2 TeamSim® [Internet]. 2019. Available from: https://surgicalscience.com/ [Accessed: 2020-01-25]

3 CardinalSim® [Internet]. 2019. Available from: https://cardinalsim.stanford.edu/ [Accessed: 2020-01-25]

* Corresponding Author

in virtual surgical training or haptic gaming [2]. When a haptic device collides with a virtual object, the haptic device performs a haptic rendering process by calculating the reaction force of the moment [3]. Then it provides haptic feedback to users, thereby enabling users to feel the touch. Nevertheless, the challenge of accurate and realistic haptic feedback requirements for high-frequency updates (1000 Hz) is always an obstacle to overcome.

Recently, many studies have been conducted applying haptic virtual reality technology, however, the study of training systems using haptic virtual reality technology goes back more than 20 years [4]. Various core technologies for effective haptic virtual reality simulations are being studied in order to provide training systems for practicing before actual operations. One of the early studies is a simulation developed to perform an elaborate work using a haptic device. This study developed an educational program for the telemanipulation of carbon nanotubes using haptic feedback and a 3D display [5].

The 3D visualization method using a VR headset is expected to produce superior results in user experience than the 3D visualization method using a flat monitor which is commonly used. The HMDs (head-mounted display) are also advantageous for acquiring cognitive skills related to remembering and understanding spatial and visual information [6]. There is also a telehaptics application for haptic probing of remote objects [7].

In the field of medicine, diverse haptic virtual reality surgery simulations are being studied in order to provide medical students and surgeons practicing before the actual surgery. One of the remarkable early studies for surgery training simulations using haptic devices and virtual reality technology can be found in 2007. It is a simulation-based training in minimally invasive surgery (MIS) which allows the trainee touch, feel, and manipulate virtual tissues and organs while viewing images of tool-tissue interactions on a monitor as in real laparoscopic procedures [8]. Also, a dental anesthesia training simulation is developed based on anatomical data. In this study, haptic technology is used for dental treatment technology education [9]. One of the recent studies on haptic virtual simulations for surgery proposed a solution to incorporate a surgeon’s sensations (haptic factors, visual factors, and hearing factors) during practical surgery into the training system to achieve a higher level of immersion. This study demonstrated that haptically enabled training simulations combining these multisensory data would be able to provide a more immersive and effective training environment [10]. Relevant to cognitive science, a study on the HVDT (haptic visual discriminant test) [11] using haptic virtual reality technology can be noted. This work presents a possible integration of vision and haptic perception of physically or mentally disabled children [12].

There is no doubt that haptic virtual simulations have a great impact on preoperative medical training. However, even if presenting and dealing with 3D objects in a virtual environment represented by a 2D monitor, the position of the object on the 2D monitor can be different from what it is in real. The reasons are that there can be distortion on the 2D screen or a lack of perspective. Because of this problem, there is a limit to recognizing the position of objects through a simple 2D plane monitor. In order to practice the operation of the correct movement, intuitive, accurate, and immersive awareness of the operation position in the VR environment is needed [1].

III. Intelligent Haptic Surgical Simulation

In the simulation, instead of interacting only with the suture site, it provides a simulation that can enter the suture site from the actual human model. Therefore, the practitioner can perform the suture surgery experience more realistically. In this way, the simulation can be easily extended to other parts of the human body. Note that the simulation can allow trainees not only perform the suture surgery experience more realistically but also provide extensibility to surgical training of various parts of the human body. In addition, the simulation of the suture surgery can also provide the ability to change the surgical tool to the shape of the surgical tool used in actual suture surgery.

Fig. 1 presents some screen captures of the surgical simulation.

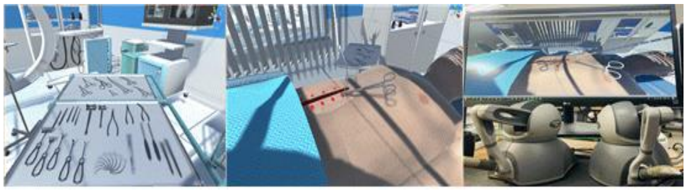


Fig. 1. Screen Captures of the Haptic Surgical Simulation.

The haptic simulation for suturing consists of the following three parts:

- Visuo-haptic rendering for surgery
- Replica training for surgery
- AI assistant for surgery

The design and implementation of each part of the haptic virtual simulation will be discussed in detail in the following sections.

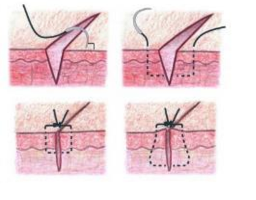
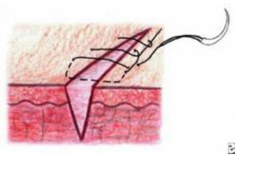
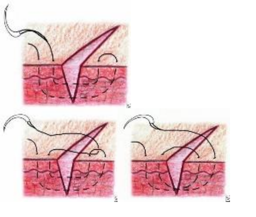
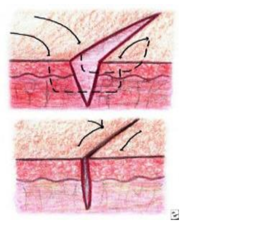
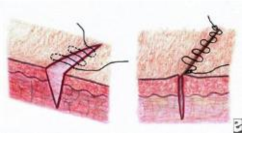
IV. Visio-Haptic Rendering for Surgery

The suturing is one of the basic techniques needed for patients with open wounds. There exist so many suturing methods. In this study, we distinguished five types of suture methods. They are:

- Simple interrupted suture,
- Running locked suture,
- Vertical mattress suture,
- Horizontal mattress suture,
- Running subcuticular sutures.

The types of suture methods and their characteristics are summarized in Table I.

TABLE. I. TYPES OF SUTURE METHODS[4]

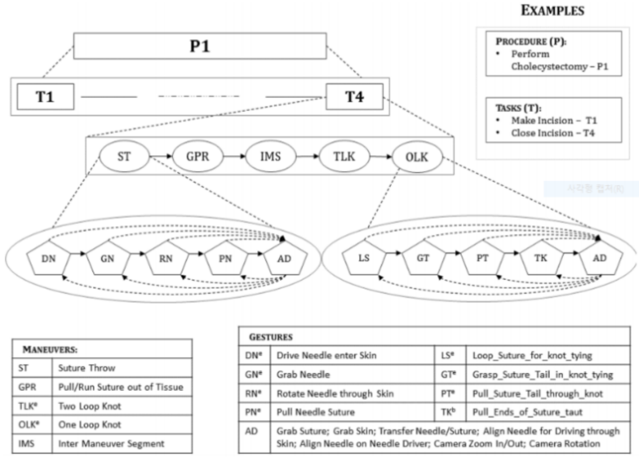
Suture Type	Image	Characteristics
Simple interrupted suture		The most commonly used and most versatile suture. The two sides of the stitch should be symmetrically placed in terms of depth and width.
Running locked suture		A simple running suture may be either locked or left unlocked, also known as the baseball.
Vertical mattress suture		This is a variation of a simple interrupted suture. The width of the stitch should be increased in proportion to the amount of tension on the wound.
Horizontal mattress suture		The suture is passed deep in the dermis to the opposite side of the suture line and exits the skin equidistant from the wound edge.
Running subcuticular sutures		It is a buried form of a running horizontal mattress suture. No suture marks are visible, and the suture may be left in place for several weeks.

A suture operation can be broken down into detailed steps. There is an excellent study on the analysis of a suturing and knot-tying task [13]. This work adopted for the design and implementation of the suturing simulation. Fig. 2 illustrates the process of suture surgery, step by step, and the process of each step. By learning the detailed process of suture surgery as shown in Fig. 1, it is possible to gain experience in the suture operation process of practitioners and induce effective learning effects.

It is notable that the most important implementation issue for visuo-haptic rendering is "precise collision detection". In order to prevent the progress of abnormal experiments due to the phenomenon of unrealistic penetration of rigid objects during the movement process, all of the objects must be set up to be a "collision object". In addition, the simulation adopts the "haptic collision detection method using subdivision surface and sphere clustering" developed in the author's previous studies [14], [15]. This algorithm can provide more precise haptic sensations. The implemented "precise collision detection" algorithm considers the following parameters:

1) The graphical tightness of bounding spheres.
2) The haptic collision detection measurements.
   a) The number of triangles.
   b) The average area of triangles.
   c) The number of bounding spheres.
   d) The collision detection time.
   e) The haptic update rate.

In order to give surgeons a realistic and immersive experience of surgical simulations, the Touch haptic devices (former PHANToM OMNI® haptic devices)[5] and Oculus Rift VR headsets[6] are used.



MANEUVERS:	
ST	Suture Throw
GPR	Pull/Run Suture out of Tissue
TLK^{e}	Two Loop Knot
OLK^{e}	One Loop Knot
IMS	Inter Maneuver Segment

GESTURES			
DN^{e}	Drive Needle enter Skin	LS^{e}	Loop_Suture_for_knot_tying
GN^{e}	Grab Needle	GT^{e}	Grasp_Suture_Tail_in_knot_tying
RN^{e}	Rotate Needle through Skin	PT^{e}	Pull_Suture_Tail_through_knot
PN^{e}	Pull Needle Suture	TK^{b}	Pull_Ends_of_Suture_taut
AD	Grab Suture; Grab Skin; Transfer Needle/Suture; Align Needle for Driving through Skin; Align Needle on Needle Driver; Camera Zoom In/Out; Camera Rotation		

Fig. 2. Hierarchical Semantic Decomposition of Surgical Activity. E Denotes that the Segment can be Performed using Either of the Robotic Arms, B Denotes that the Segment is Performed using both the Robotic Arms [13].

V. REPLICA TRAINING FOR SURGERY

This study considers the following two surgical imitation training:

- Replica training through file
- Replica training in real-time over the network

A. Replica Training through File

After recording the surgical procedure of the skilled medical expert in a specific file, it is restored to the simulation of the trainee so that it can be experienced as it is. This method allows the practitioner to learn the surgical motion step by step by reproducing the data recorded all the motions of the most ideal surgical simulation. This method has the advantage of allowing repeated training by imitating the expert's know-how.

B. Replica Training in Real-Time over the Network

This method allows the trainee surgeon to replicate and experience the surgical procedures of the expert surgeon in real-time. Professional work such as surgical operations can be

[4] Placement of Specific Suture Types. 2020. Available from: https://emedicine.medscape.com/article/1824895-technique#c3 [Accessed: 2020-01-25]

[5] Touch. 2020. Available from https://www.3dsystems.com/haptics-devices/touch [Accessed: 2020-01-25]

[6] Oculus Rift S. Available from https://www.oculus.com/rift-s/ [Accessed: 2020-01-25]

applied with the knowledge and sensations or know-how from the accumulated experience of a skilled person. Things are difficult to teach by word or writing and have characteristics that must be learned through hands-on experience. In this study, the surgical operations of a skilled expert can be imitated by medical trainees through network transmission. If the accumulated experience and know-how of skilled practitioners can be delivered through haptic feedback using haptic devices, the training effect of the practitioners will be much higher.

In order to implement real-time simulation via network transmission, client-server socket programming was used to improve communication between the trainee and the expert for the immediate copying of remote operations. The real-time synchronization through the network takes more time to be transmitted and received from a remote place. Therefore, the protocol selection at the transport layer should be made in consideration of the problem of latency of the response to haptic feedback. Protocols in the IP based transport layer include TCP and UDP. TCP is a connection-oriented protocol, which includes establishing a connection and performing flow control and retransmission protocols. The UDP protocol has a disadvantage that data loss is not known because retransmission and flow control is not performed. However, the TCP protocol is not suitable for applications where high speed is important. It is well-known that using UDP rather than TCP is suitable for real-time systems. Therefore, the real-time networked experience copying system uses UDP protocol instead of TCP because transmission speed is very important.

We developed a prototype system for replicating an expert surgeon's surgical sutures to a trainee surgeon. Fig. 3 presents an example of the replica training.

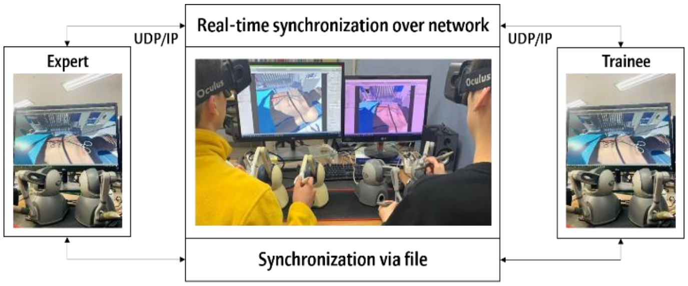


Fig. 3. Replica Training in Real-Time over the Network.

VI. AI Assistant for Surgery

We are developing a haptic VR simulation system that allows users to learn the suture technique appropriate for the characteristics of wounds. In order to make the simulation intelligent, we designed an AI (artificial intelligence) assistant which can recommend an adequate suture method for each specific wound. In addition, T system is designed to feedback on the results of previous training, and recommend an appropriate suture technique using AI technology.

In order to provide correct surgical information according to the level and characteristics of the trainee (left-handed, surgical-related physical characteristics, etc.) based on the proposed suture method through AI for a given wound in the surgical environment.

The method of suture depends on:

- the type of wound,
- the thickness of the skin,
- the degree of tension,
- the degree of suture scars,
- the degree of bleeding, etc.

After arranging data recorded by the skilled surgeon based on the above five criteria, a deep learning model was designed for distinguishing a variety of suture surgery methods for recommending an optimal suture method.

An evaluation standard is established to measure the trainee's proficiency based on:

- the accuracy judgment and
- the execution time of the suturing for evaluating the results of the trainee's suture.

The system measures the proficiency of the trainees according to the evaluation criteria, feeds back the parts deemed insufficient in the simulation and recommends additional suture training for further study.

Fig. 4 is the wound image processing model that predicts the kind of wound. A collection of wound data is normalized into a form suitable for the deep learning model for the classification. The wound image data are preprocessed through dimension reduction and color correction. The model also includes some traditional feature extraction and labeling by assigning a unique index to each label using the one-hot encoding.

The AI assistant recognizes the incision wound image and recommends the optimal suture for the desired closure of the incised wound. Among many machine learning techniques, the best performing machine learning technology for image recognition is CNN (convolution neural network). The CNN network is constructed by extracting features through convolutional computation in computer vision, applying activation functions to the configured feature maps, and constructing the active maps and passing the pooling layers through the active maps.

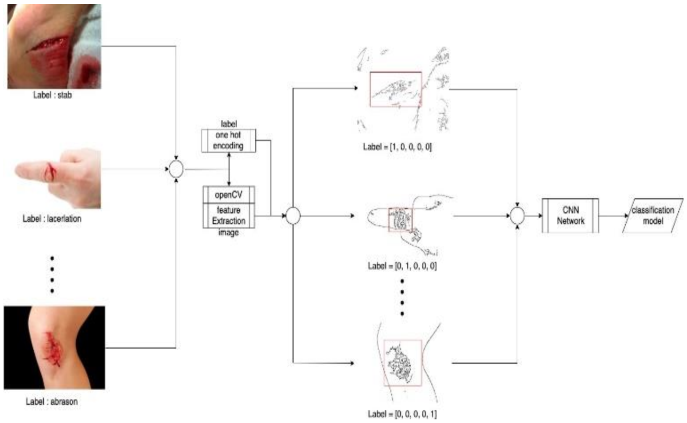


Fig. 4. Wound Image Processing Model.

An AI learning model was developed for the proper classification of wounds. Part of the data preprocessing and optimization for proper classification of wound images are the most important core goal of the research. In order to achieve this core goal, data analysis on the incision wound images should extract the characteristics of the wounds for classification. In addition, it is required to build an AI model for wound image training and control overfitting and hyperparameters that appear when the real data is trained and verified. Finally, the optimization process will generate a model that is best suited to accurately classify the cause of the wound for new wound image data input.

A brief description of the AI modelization is presented in the following three steps:

- Data collection and analysis for wound images
- Preprocessing for wound images
- AI model optimization for wound images

A. *Data Collection and Analysis for Wound Images*

The first step in developing an AI application is to collect good data. In addition, special attributes or feature information that can be extracted from the wound image will be secured through the cooperation of a university hospital.

The features and attributes of collected data are analyzed to be trained by the AI model. Attributes are categorized based on information on what kind of wound the data are:

- what the wound is,
- what the level of wound contamination is, and
- how the patient's age or health is.

The classification of detailed features and attributes of the data is directly related to increasing the accuracy of wound data classification by extracting and reflecting detailed features when the data is actually input.

B. *4.2 Preprocessing and Correction for Wound Images*

Based on the collected wound data analysis, preprocessing can be applied for the training data. The training data are preprocessed and further classified into detailed attributes in the appropriate form for the AI model.

Many studies are performed for recognizing wound images [16], [17], [18], [19], [20]. The correction of medical images is the most important process to improve the accuracy of the model.

C. *4.3 AI Model Optimization for Wound Images*

The most important part is to build an optimal AI model for classifying suture methods. The AI model should play an appropriate role in extracting features from the image data, and the final result label should be one of the types of suturing methods illustrated in Table I.

During the model construction process, the components such as hyperparameters and parameters that make up the AI model are adjusted and optimized to best fit the training data. In addition, the model optimization is performed by additionally applying a technique for eliminating the influences caused by similar images or a technique for preventing overfitting. Through this process, the AI model will be optimized for the wound images, which is the training data, and it can classify the type of wound data with high accuracy.

After the preprocessing, the wound image data are entered into the CNN model presented in Fig. 5.

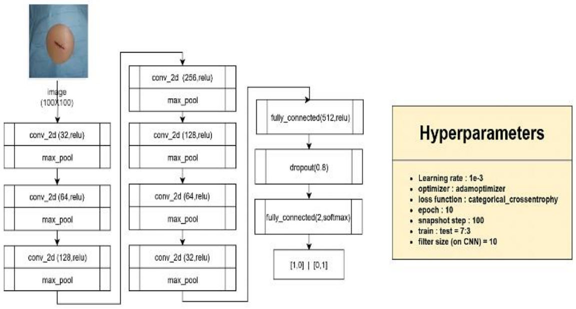


Fig. 5. AI Model and Hyperparameters for Wound Image Classification.

Fig. 5 summarizes the AI model and hyperparameters for wound image classification. The input images pass through two fully connected networks and performing seven convolution operations and max poolings in the wound image. The learning rate used in this model is 0.0001 and the optimizer is used as Adam optimizer. The loss function is categorical cross-entropy, 19 epochs are performed, and the ReLU (Rectified Linear Unit) function is used as an activation function. The filter size used in the convolution operation is 10×10.

VII. Discussion

Surgical simulations have been studied for over 20 years and now fully demonstrate their effectiveness. There is no doubt that by presenting a virtual surgical simulation as a stereoscopic 3D objects using a VR headset. the intuitive operation performance has been improved and especially operational accuracy has been greatly enhanced [1]. The most difficult part of virtual simulations is the graphical rendering of it. Precise rendering of graphics results in precise haptic rendering. Advances in graphics hardware enable realistic visual rendering. However, haptic rendering is still very different from realistic touch, and the precise and realistic haptic rendering has room for improvement.

Recently, mixed reality technology has been in the spotlight. I would like to emphasize that the simulation will become much more useful if it is converted to a mixed reality environment. In this case, the suture simulation can be useful not only for suture training but also for actual surgical procedures. In particular, the AI assistant for surgery can recognize the wound in real-time, suggest an appropriate suture method for the wound, and guide the suture procedure, etc.

VIII. Conclusion

In this study, a simulation for learning and evaluating suture surgery using haptic VR and AI technology is developed. In order to provide medical trainees with realistic visual and haptic effects, it is very important to obtain accurate physical properties of the body and the precise and stable presentation of touch. Therefore, a “precise collision detection algorithm using subdivision surface and sphere clustering” that was proposed in the author’s previous work is implemented.

Moreover, in the simulation, an expert surgeon's surgical procedure can be replicated to the trainee surgeon's simulation for mimicking the suturing. This method allows the trainee to learn the ideal motions of the expert surgeon.

The AI assistant for suturing learns the wound dataset and corresponding labels of typical suturing methods in consideration of the criteria of wounds, such as the type of wound, the thickness of the skin, the degree of tension, the degree of suture scars, and the degree of bleeding, etc. After arranging the wound data recorded by the skilled surgeon in those criteria, the AI model can learn the wound image data and their corresponding suture methods using deep learning techniques. Then, the AI assistant can recommend an optimal suture technique for a given wound.

It is expected that the simulation can allow for reducing the cost and time required for suture training and ultimately provide safe and accurate actual surgery. It is also advantageous that the learning curve of the trainee can be sharply increased by allowing them to experience the ideal surgical movements and procedures. In addition, the system can be improved for collaborative training, remote surgery, cooperative surgery, and remote cooperative surgery, etc. The results of this study can be used not only in the medical field, but also in all fields such as education, games, entertainment, and industry, etc.

Future studies will be focused on the evaluation of the developed suture simulation as well as the AI model for improving the efficiency of the suturing and on the provision of personalized feedback for enhancing the proficiency of the surgeon. In addition, the transformation of this intelligent haptic virtual simulation for suture surgery into a mixed reality environment will be conducted.

Acknowledgment

This work was supported by Incheon National University Research Grant in 2018 (No. 2018-0333).

The authors would like to thank Mr. Haesung Joo, a student of the Department of Computer Science and Engineering at Incheon National University, for the help in various experiments.

References

[1] T. Kim, C. Kim, H. Song, and M. Y. Sung. “Intuition, Accuracy, and Immersiveness Analysis of 3D Visualization Methods for Haptic Virtual Reality,” International Journal of Advanced Computer Science and Applications (IJACSA). vol. 10, no. 11, pp. 30–37, 2019. DOI: https://dx.doi.org/10.14569/IJACSA.2019.0101105.

[2] S. M. Kim, M. Y. Sung, “A Haptic Gaming System for Tactile Textures and 3D Shapes Discrimination,” International Journal of Multimedia & Ubiquitous Engineering, Vol. 9 Issue 9, pp. 319-334, September 2014.

[3] K. Salisbury and F. Barbagli, “Haptic rendering: introductory concepts,” Computer Graphics and Applications, IEEE vol. 24, no. 2, pp. 24-32, 2004.

[4] L. B. Rosenberg and D. Stredney, “A haptic interface for virtual simulation of endoscopic surgery,” Studies in Health Technology and Informatics. vol. 29, 1996, pp. 371–387. ISSN 0926-9630. PMID 10172846.

[5] Z. Gao and A. Lécuyer, “A vr simulator for training and prototyping of telemanipulation of nanotubes,” Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST 2008), 27-29 October 2008, Bordeaux, ACM, pp. 101–104. 2008.

[6] L. Jensen and F. Konradsen, “A review of the use of virtual reality head-mounted displays in education and training. Education and Information Technologies,” vol. 23, no. 4, pp. 1515–1529, 2018, DOI: https://doi.org/10.1007/s10639-017-9676-0.

[7] A. R. Choi, C. W. Kim, M. Gwak, and M. Y. Sung, “Haptic Interactions for Probing Real Objects in Remote Places,” International Journal of Applied Engineering Research (IJAER), vol. 12, no. 24, pp. 14948–14954. 2017. ISSN 0973-4562, eISSN 0973-9769.

[8] C. Basdogan, S. Mert, H. Matthias, and W. Stefan, “VR-based simulators for training in minimally invasive surgery,” IEEE Computer Graphics and Applications, vol. 27, no. 2, pp. 54–66, 2007. DOI: https://doi.org/10.1109/MCG.2007.51.

[9] M. Poyade, A. Lysakowski, and P. Anderson, “Development of a haptic training simulation for the administration of dental anaesthesia based upon accurate anatomical data,” Proceedings of the Conference and Exhibition of the Association of Virtual and Augmented Reality (EuroVR 2014), 8-10 December 2014, Bremen, The Eurographics Association, p. 143–147, 2014.

[10] Y. Tai, W. Lei, X. Minhui, Z. Hailing, L. Qiong, S. Junsheng, and N. Saeid, “A high-immersive medical training platform using direct intraoperative data,” IEEE access, vol. 6, 2018, pp. 69438-69452.

[11] L. M. Carron and P. W. Horn, “Haptic visual discrimination and intelligence,” Journal of Clinical Psychology, vol. 35, no. 1, pp. 117–120, 1979. DOI: https://doi.org/10.1002/1097-4679(197901)35:1%3C117::AID-JCLP2270350119%3E3.0.CO;2-Z.

[12] H. Y. Kim and M. Y. Sung, “Virtual Haptic Visual Discrimination Test,” Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 10, no. 1, pp. 5–11, 2018. ISSN: 2180-1843, eISSN: 2289-8131.

[13] S. S. Vedula, A. O. Malpani, L. Tao, G. Chen, Y. Gao, P. Poddar, N. Ahmidi, C. Paxton, R. Vidal, S. Khudanpur, G. D. Hager, and C. C. G. Chen. “Analysis of the Structure of Surgical Activity for a Suturing and Knot-Tying Task,” PLoS ONE, vol. 11, no. 3, March 7, 2016. DOI: https://doi.org/10.1371/journal.pone.0149174.

[14] A. R. Choi, S. M. Kim, and M. Y. Sung, “Controlling the Contact Levels of Detail for Fast and Precise Haptic Collision Detection,” Frontiers of Information Technology & Electronic Engineering (FITEE). vol. 18, no. 8, pp. 1117–1130, 2017. DOI: https://doi.org/10.1631/fitee.1500498.

[15] A. R. Choi and M. Y. Sung, “Performance improvement of haptic collision detection using subdivision surface and sphere clustering,” PLoS ONE, vol. 12, no. 9, 26 Septemeber 2017. DOI: https://doi.org/10.1371/journal.pone.0184334.

[16] H. Oduncu, A. Hoppe, M. Clark, R. J. Williams, and K. G. Harding, “Analysis of skin wound images using digital color image processing: a preliminary communication,” The International Journal of Lower Extremity Wounds, vol. 3, no. 3, pp. 151–156, 2004. https://doi.org/10.1177/1534734604268842.

[17] N. Engström, F. Hansson, L. Hellgren, T. Johansson, B. Nordin, J. Vincent, and A. Wahlberg, “Computerized wound image analysis,” Pathogenesis of Wound and Biomaterial-Associated Infections, Springer, London, 1990, pp. 189–192. DOI: https://doi.org/10.1007/978-1-4471-3454-1_24.

[18] F. J. Veredas, R. M. Luque-Baena, F. J. Martín-Santos, J. C. Morilla-Herrera, and L. Morente, “Wound image evaluation with machine learning,” Neurocomputing, vol. 164, pp. 112–122, September 2015. DOI: https://doi.org/10.1016/j.neucom.2014.12.091.

[19] F. L. Bowling, L. King, H. Fadavi, J. A. Paterson, K. Preece, R. W. Daniel, D. J. Matthews, and A. J. M. Boulton, “An assessment of the accuracy and usability of a novel optical wound measurement system,” Journal compilation Diabetes UK. Diabetic Medicine, vol. 26, pp. 93–96, 2009 DOI: https://doi.org/10.1111/j.1464-5491.2008.02611.x.

[20] H. Lu, B. L, J. Zhu, Y. Li, Y. L, X. Xu, L. He, X. Li, J. Li, and S. Serikawa1, “Wound intensity correction and segmentation with convolutional neural networks. Concurrency and Computation: Practice and Experience,” Concurrency. Published in Wiley Online Library (wileyonlinelibrary.com), vol. 29. issue 6, 2017. DOI: https://doi.org/10.1002/cpe.3927.

A Microservices based Approach for City Traffic Simulation

Toma Becea[1], Honoriu Vãlean[2]
Automation and Computer Science Faculty
Technical University of Cluj-Napoca
Cluj-Napoca, Romania

***Abstract*—The paper proposes a city traffic software simulation based on actors which run independently of one another and have specific characters in their behavior. To run indepentently actors are modeled as microservices and they are running within an orchestration framework. Their behavior is modeled as with specific algorithms for each of their type, embedded in each actor's type code. They may act based on the data about all the other actors, data which is gathered together by a single entity called city simulator. An orchestration model is proposed and all the actors use a communication protocol to offer data to the city simulator and request data from it.**

Keywords*—*Traffic simulation; microservices; distributed computing

I. Introduction

A solution which simulates car, pedestrian, etc. traffic in a given city may reap many benefits. It can help understand patterns of traffic and its flow. It can help understand the particularities and pecularities of a city's streets arrangements, together with their junctions. It can help identify bottlenecks. It can help find solutions to rush problems and explore them. But for those areas to be tackled, appropiate methods of simulating traffic must be found and explored. We are proposing a novel way of simulating traffic based on the microservices orchestration concept and using a discrete microscopic logic for the behavior of each actor.

The solution proposed is aiming to design a traffic simulation software which surpasses the computing limits of a single machine and of existing traffic simulation software and runs as a distributed system. In close interplay with the fundamental nature of distributed system (i.e. running a cohesive software on multiple machines) the solution defines entities or actors, modeled as separate microservices, such that the advantage is twofold. First, an entity being a single microservice (and a microservice containing only one entity) they will be easier and natural scheduled and run across more than one machine. Second, the independence of such an entity is also helping in personalize its behavior in randomness and character, offering different ways to model the traffic in a city and allowing more nuanced studies as opposed to macroscopic solutions. This solution can improve a system where only the global state is computed.

We define a **(city) actor** as being an independent entity which chose to move between two geographical points within a city. As a character, it can be a car, a pedestrian or a bike. We also define the **city simulator** as being a single entity (subject to distributed and load balacing services) which keeps data about the city (e.g. streets with city actors on them).

II. Related Works

Existing solutions have a general distinction of being scattered across a spectrum: macroscopic, mesoscopic or microscopic being milestones across it. This distinction is caused by how the traffic is modeled: on a macroscopic scale like streets or sections of a highway or on microscopic level, focusing on each car and its relation with neighbours. The relation between a car and its neighbours, can also be modeled as microscopic or nanoscopic.

Author in [1] describes an Agent Based Model (ABM) for improved traffic routing and achieve a system-optimal traffic flow. The similarity with the current paper is that the data is generating at the agent level and the decision remains at the same level, although there are two layers in total. First is the microscopic layer which consists of all the intelligent agents. The second layer is the macroscopic layer which facilitates the communications between agents. However, the main difference is that in [1] an agent is able communicate only with the surrounding ones through the use of cellular automata approach. Agents are communicating a handful of data with their co-participants within a certain range: position, velocity, route and type. Using the data they received they are using a transition decision-making model, based on cellular automata, to compute their next move.

In [2] the broad idea is to use a discrete event architecture, in which there are logical processes for executing and simulating a number of agents. For a computation intensive aim those processes can be clustered into agent clusters with dependencies between them, but the paper avoids presenting the necessary details on how the distributed mechanics would work. The particularity of this proposal is that it crosses the boundaries of a single computing machine and pave a way to distribute the load and information across machines, with certain limitations and challenges.

An approach which is focused on junction modeling is [3]. As opposed to the generality of various other related works, [3] is focused on two specific and real junctions. The simulation stems from SUMO, using real data acquired from cameras placed in junctions (and analyzed with image processing technologies) and is relaying traffic data to a Matlab instance. The simulation is aiming at evaluating metrics of the traffic which flows through junctions: arrival flow and queue length. This idea is closely related to an enhancement proposed in Section VI around junction modeling.

Although older, [4] deserves to be written about because the real data, from the highway portions of interest, is fed back into the simulation. Loop detectors, scattered across the highways are read and their readings are used to generate simulation models. The model is able to predict car densities across multiple road sections, an ability which the current paper is exploring and is basing its car actor's algorithms onto.

III. Implementation

The entities which are participating into a traffic simulation are called actors. They are two: the city actor and the city simulator. The supporting containers are not themselves part of traffic simulation but they do have an important and supporting role. All of them are modeled as microservices and a short introductory is needed although Section IV and paragraph III-E offers a broader perspective around them.

A. Microservices

Microservices are not a new concept. The idea behind them has existed since Linux kernel has started to be enriched with a concept called namespaces [5]. This allows one set of processes to see one set of resources while another set of processes see another set of resources, where resources might be, but not limited to, process IDs, file names and network resources. Those linux kernel abilities form the base of containers.

Thus, a container is a small set of processes which run in isolation. They allow packaging a linux distro, a set of libraries and a software development kit and on top of those the custom code of an application. Taken together, those form an image, which is essentially a tar gzipped file. Once the build process of an image is finished, it can be spinned up in one or more running containers. The custom code written and embedded in the image is running in parallel in each container.

To go to solution for building, manipulating and running images is Docker [6]. It allows easy software installation, it works cross platform and it offers a smooth experience most of the time.

B. City Simulator

The most common and easy solution to share data across all the city actors is to have a centralized store to keep it. City simulator acts as a centralized store for all other city actors. In the current implementation the city simulator keeps a set of data which can be described as a list of pairs, each pair having a line and a real number, called density. The line is a series of coordinates and in the proposed implementation their meaning is a street which a city actor is reportedly traveling on it.

The density is defined as the number of actors (cars) which are at a given moment present on a given segment of street. If the city simulator has no entry of a street segment then it will consider the density as being 0, i.e. there is no actor on that street. The density is modeled as an unsigned integer.

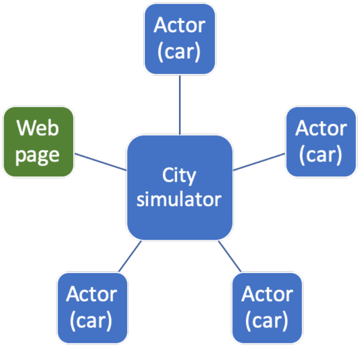


Fig. 1. The relations of city simulator with other entities

Fig. 1 depicts the city simulator in relation with the other entities. A notable exception is the web page. Its purpose is not to participate in the same information exchange the other entities have but to offer a visual interpretation about the way the city and the other actors are interacting with one another.

As various actors send data about their location, the city simulator needs to keep location and density data in its store. It does not know of Open Street Map maps and its corresponding map data sets but instead it requires any actor to send its current set of coordinates which it is crossing or which it left. If any other actor send the same set of coordinates, or a subset of it then the first set of coordinates will have its density incremented. Listing 1 shows the Go struct which is used by an actor to send its report to the city simulator and is used by the city simulator to decode a message received from an actor. Apart from this line there are two more details to complete the report picture. Listing 2 contains the type of communications between actors and the citysimulator.

```
// Report is the base type for reporting
// status and vectors to a city entity
type Report struct {
  CurrentLine  [][]float64
  ReportDetail int
}
```

Listing 1: Go struct for actor's report

As part of this proposal the city simulator is made to be a standalone entity (container) which communicates with actors. This might not be the case for other types of communications, as part of other architectures. One example can be the integration or the unifying of the city simulator with the city actor, both becoming one entity. In this case the communication between entities is subject to an entire panoply of choices.

C. Car Actor

The city actor represents a moving actor within a city. It can be a car which moves across the city, it can be a bike or it can be a pedestrian. Its naming suggests that it can be any entity or living being which moves within a city and interacts with the other entities or affects the other entities in some manner. A pedestrian would directly interact with other pedestrians but not with cars unless it crosses the street on red or on unmarked places. A pedestrian would indirectly affect other cars by willing to walk over a street crossing.

The implementation proposed here is aiming to model a single type of actor: a car which moves between two points across a city. Because the purpose of this paper is not to tackle maps representation and routing through a city (in itself, this area is way bigger than a mere technical paper) the city actor is using two notable services: Open Street Map [7] and GraphHopper [8]. Open Street Map is an open source licensed map of the entire world. Graphhoper is an open source service which offers directions APIs and route planning. It can use Open Street Map as an underlying map provider. They provide free tiers and paid subscription for accessing an API and compute various routes. However, a city actor is not using any public api but a special crafted container which contains GraphHopper and an open street map embedded into it. Thus, this container is running in parallel with the other city actors and provides them with the routing API they need.

A first way to introduce randomness into the entire simulation of a city is to choose a set of two coordinates inside the given city, a set for each city actor. Then the actor proceeds to ask GraphHopper service for a route between those two points. Because each city actor 'lives' only while its moving across its route and 'dies' as soon as it reached the finish point, the entire city simulation which takes place is made of independent, random and always new actors.

D. Interactions

An actor interacts, for the time being, only with the city simulator by using a Go struct and a Go enum. They can be seen in listings 1 and 2. The message sent from one side to another has a general structure called *Envelope*. It is meant to offer a top level message which can be serialized (or encoded, the way to do this in Golgang, if not gRPC, is the gob pacakge offered out of the box as a base package within the lanugage) and passed around, enabling any party involved to understand what this message is about and how to decode it.

Listing 3 show the top level message. It contains a message type which instructs the reader what kind of message it has to deal with and the possible types are the second part of listing 1: *SendReport*, *AskForLine*, *RespondWithLine*. Let's take them one by one. *SendReport* represents a message sent from an actor to the city simulator and the *Payload* contains the report seen in listing 2.

AskForLine is a message sent from an actor to the city simulator in which the actor asks about the density of any line. This way any actor can take conscious decisions on which route to go on, based on what lies ahead in terms of street densities. When a first route is chosen, between two desired points, the actor can ask about each line which is part of that route. The city simulator will respond with the known density of it. If the actor desires, it can try to find another route by asking GraphHopper service to compute a new route but with an additional rule: avoid a certain point (street, junction, etc.). *RespondWithLine* is the type of message which the city simulator sends back to an actor after it received an *AskForLine* message.

```
const (
    // ReportOnTheLine is the report sent by one agent to
    // notify the city that he is currently advancing
    // through one line.
    ReportOnTheLine = iota

    // ReportOffFromLine is the report sent by one agent to
    // notify the city that he has finished advancing through
    // one line and has departed from it.
    ReportOffFromLine = iota
)

const (
    // SendReport is a message passed from an actor to the city
    // indicating its status (e.g. location).
    SendReport = iota

    // AskForLine is a message passed from an actor to the city.
    // A response is awaited.
    AskForLine = iota

    // RespondWithLine is a message passed from the city to
    // an actor and it contains line data.
    RespondWithLine = iota
)
```

Listing 2: Go enumerations for messaging

```
// Envelope is the container for different messages sent back
// and forth between an actor and a city
type Envelope struct {
    MessageType int
    Payload     interface{}
}
```

Listing 3: Go top level struct (envelope)

E. Design Choices

As with every software project started from scratch there are a number of choices to make when choosing software stacks, programming languages, networking models, etc. This section is aiming to explore the rationales behind some of those decisions and how they influenced the building of the current prototype.

A first choice is the programming language to write both the city simulator and the city actor. Go programming langauge is born out of Google and it resembles the philosophy they were trying to embedded in it for taming the complexity of their systems [9]. Today it has gained a lot of popularity and tools like Kubernetes are wrritten entirely in Go, making it the default language of the cloud technologies. It is a "C-like" language and it has a familiar look but with few traits which makes it different. Out of those, few are notable, not only because they facilitate programming endeavors but also because both city simulator and city actor are modeled around them in their communications.

```
go func() {
  for {
    select {
    case <-ticker.C:
      advance(city, reportChan, lineChan)
    case _ = <-lineChan:
      //fmt.Println("Received answer with line", j)
    }
  }
}()
```

Listing 4: Go routine from city actor

Go routines are a lightweight thread of executing. Being lightweight it means they can be easily started, with no overhead and especially without the ceremony of dealing with threads which other popular languages (e.g. Java) have. Listing 4 shows such a routine. It is a part of logic where the city actor responds to either two events: the timer has expired and a decision needs to be made or an information about the line is currently traveling on has been received from the city simulator. Another Go trait is the channels, an indexed communication pipe where there are asynchronous writers and readers, used to decouple two routines. Listing 4 also shows couple of channels. Variables *ticker* and *lineChan* are two channels. They are declared using a type and after that they are used to write and read objects (or structs) of the respective type. The *select* keyword acts like a switch, not on variables and their values but on channels. It will execute the block of code for the channel which has something new to deliver. All those three concepts, *goroutines*, *channels* and *select* switches, combined together makes it easy to write code which takes advantage of multithreading paradigms and which has to run in a heavy networking environment.

IV. Orchestration

The term orchestration means the handling of containers and microservices in order to bring coherence into their interaction and an unifying experience to the end user of a service or of a product. They migth run on multiple nodes (computers or virtual machines) and at the same time they need to communicate with one another. They need to be updated in place and without service disruption. Whenever one of them crash they have to restart as quickly as possible. The services have to be able to discover themselves without dealing with intricancies of IP addresses, proxies, etc. Those are just few concerns around orchestration and the current paper is not aiming to provide a comprehensive view of what it means and what can be achieved with it but rather to set a basis of understanding enough context for the subject of simulating a city with its traffic.

Author in [10] offers a comprehensive view of the current state of orchestration of cloud providers. (to add details)

A. Current State

As noted above the standard in easiness of developing and working with containers is Docker [6]. While it has an offering of orchestrating containers, called *docker-compose*, which is simple to start with, its functionality is limited in comparison with other offerings, the most notable one being Kubernetes [11]. Those are not the single tools available and many more can be found but they offer a starting point (especially Docker) and Kubernetes, altough it has a steep learning curve, it does offer a comprehensive and complex panoply of details around orchestration.

B. Implementation

The city simulator is, currently, a single service which means it runs as a single instance as viewed by a city actor. The city actor container runs in multiple instances and it has the need to do so as part of the entire simulation workload. The other two instances which run in the simulation are the routing service, a Graphhopper instance, and the front end web server which displays a web page for offering visual clues about how the simulation is running. While they run as such there should be an easy way, without friction and additional compute logic to "discover" a certain service. A city actor will need to connect itself to the routing service and to the city simulator. At the same time the front end instance need to connect itself to the city simulator to source its data.

C. Docker Compose

The *docker-compose* tool gives the ability to manipulate more containers and services, with a simple file written in yaml format. Listing 5 shows the docker-compose file in its brevity and briefness. Let's dissect it. There are four services: *graphhopper*, *citysim*, *cityactor* and *cityfront*. Each of them need an image to run and this image can be specified in two ways: either as an image already compiled and hosted on an container registry or as a local folder which contains a file to build one (usually named *dockerfile*). Because Graphhoper, once compiled with desired maps and settings do not need any more development work, it is uploaded to a personal docker registry and taken from there, whenever needed. The other three services are the places where the most development efforts take place therefore they need to be compiled or recompiled each time the entire traffic simulation application starts.

The containers which are running may need to have different properties. First and most important is the container port which needs to be published. The port from inside the container, where a certain process expect a TCP communication is forwarded to the local host port and thus is accessible from outside the Docker internal network. Another detail of the container is its dependency. For example *cityactor* cannot run without textitGraphhopper because it doesn't have any place to obtain a route. Therefore it depends on *Graphhopper*, i.e. it waits for GraphHopper to start first. And on *citysim*, of course. The last bit of detail which can be seen here is the policy of restarting a container. By default the policy is set to "No" which means that the container will not be restarted if there is any failure within it and it stops. However, for simulation purposes, as we need a constant stream of city actors to swarm through the city, the policy is set to "Always" which will restart the container after if closes itself, i.e. it finishes its travel.

```yaml
services:
  graphhopper:
    image: tomabecea/graphhopper:latest
    ports:
      - "8989:8989"
  citysim:
    build: ./city/citysim
    ports:
      - "9000:9000"
  cityactor:
    build: ./city/cityactor
    depends_on:
      - "graphhopper"
    restart: always
  cityfront:
    build: ./cityfront
    ports:
      - "80:80"
    depends_on:
      - "citysim"
```

Listing 5: Docker-compose file

Finally, to run this docker compose there is a simple command to bring everything to life: ***docker-compose up***. This will take everything which is in the *docker-compose.yaml* file, will build the container if it has not been built before or if any source code file has been changed and then will run all of them in the order specified by the dependency graph.

The other detail of running this set of containers is to scale up a specific service. In our case we would like to have more than one city actor. Therefore the command to run is ***docker-compose up – scale cityactor=1001***. This way one thousand and one city actors will be created. In combination with the option of always restarting a container which exits, the entire simulation will run virtually forever.

D. Kubernetes

While the docker compose offers a basic set of functionalities to bootstrap our application, for a large number of city actors a single computer might not suffice. Here comes Kubernetes in play. Kubernetes [11] is an open-source system for automating deployment, scaling and management of containerized applications. It originates from Google and is their third approach on orchestrating microservices, as described in [12].

While Docker-Compose runs on a single machine, Kubernetes is made to run on multiple machines called nodes. It is by default enriched with certain abilities needed to run in this configuration. Made initially to run with LXC and Docker containers [13], it is now a more open system where one can choose the container runtime interface, the container network interface, storage interface, service meshes, etc. from a broad range of vendors. On top of those base offerings, which provides only the backbone of Kubernetes, a distributed application must be built. The desired architecture of a Kubernetes application may consists of many concepts and building bricks. Using [14] as a starting point we can see that there are the most simple and basic units, called pods, which may encompass one or more containers (usually there is one main container and the others, e.g. service mesh supporting side car, are called side cars). There are deployments which gather together pods and rollout or rollback control. There are services which makes pods universally available into the cluster and makes them discoverable, regardless of what node are they running onto and abstracting away failures and upgrades. And then there are replica sets, daemon sets, secrets, config maps and many other resources. On top of those one is able to deploy own custom resource definitions as well as custom controllers. Put together, all those concepts have a rather intimidating allure.

For running the entire traffic simulation onto a Kubernetes cluster we have the basic needs: all services are Docker images. Once each service is correctly described using a yaml format, each consisting of a service and deployment the entire application can be deployed via *kubectl* command. The scaling problem is solved using a similar approach with *docker-compose*. The command will be ***kubectl scale –replicas=1001 deployment/cityactor***. As before, with *docker-compose*, it is to be noted the simple approach towards scaling: a core tenet of distributed systems, a property which, as it stands for our use case is quite simple to model. But for other applications, a more sensible approach is needed, as noted by [15].

E. Networking Design

A distributed system has to carefully design the way in which its services communicates between them. As the nature of a distributed system is to run in multiple nodes or machines, it is obvious that the communication medium between them has to be one based on TCP/IP stack. Therefore the entire modeling of networking relies on this assumption.

A simple and basic idea, noted here only to help on constructing the final proposed solution, is to have each service always located at a certain IP address and a certain port within a network. This means, for example, the city simulator will always be located at 192.168.0.101:7450. The other services which need to be accessed will listen on similar IP addresses and sockets. While this offers a convenient way to have them unified, they are not suitable to run in other environments than a development PC. In this case *docker-compse* will run them such that they are accessible on local host address, i.e. 127.0.0.1:7450. As soon as there are multiple nodes involved this model is not suitable anymore.

Enter the DNS (Domain Name System), the backbone of the entire internet. In its simplest description, avoiding many inherent and nitpicking details, it is a dictionary which keeps track of every registered and easy to memorize name, e.g. *en.wikipedia.com* and the IP addresses where it is located. Any client which would like to access such an name will query first the DNS resolvers and then it will proceed to send a message to the obtained IP address. The most notable interaction of a user with the DNS is the address bar of a browser where the user inserts the name of the site they want to access and while writing the address suggestions are displayed by the browser with the help of recommandation engines [16].

When many resources are needed, to serve a great number of users or to support a great number of city actors, a certain service might be so busy with serving data that the compute and memory resources it needs are greater than the underlying hardware is able to support. Therefore it might be located at few addresses at the same time and any client which wants to access it should access the address where there are available compute and memory resources to serve its queries. Ideal would be to abstract or to decouple this information from the client and make it transparent for it. To do this the client needs to know only the service name and a certain ”networking” entity should route its request to the appropriate available service. Such a entity could be a DNS authoritative server which, based on the load, availability and latency of the services will route the request to the appropriate node [17].

Docker and Kubernetes are doing a similar job. They offer a DNS service and according to the load of each node where a service run, a request is rooted to a node which will be able to respond to it. This logic is completely decoupled from the clients. Listings 6 and 7 shows how both the city front and a city actor calls the city simulator. All what they do is to use the address, the service name, which was specified in the docker-compose file (listing 5). Docker or Kubernetes will do the actual job to route the request to the appropriate node of the service. One thing to note is the difference between the two. A city client will communicate via TCP/IP using the Go default serialization package while the city front use a WebSocket communication technology (add reference to the

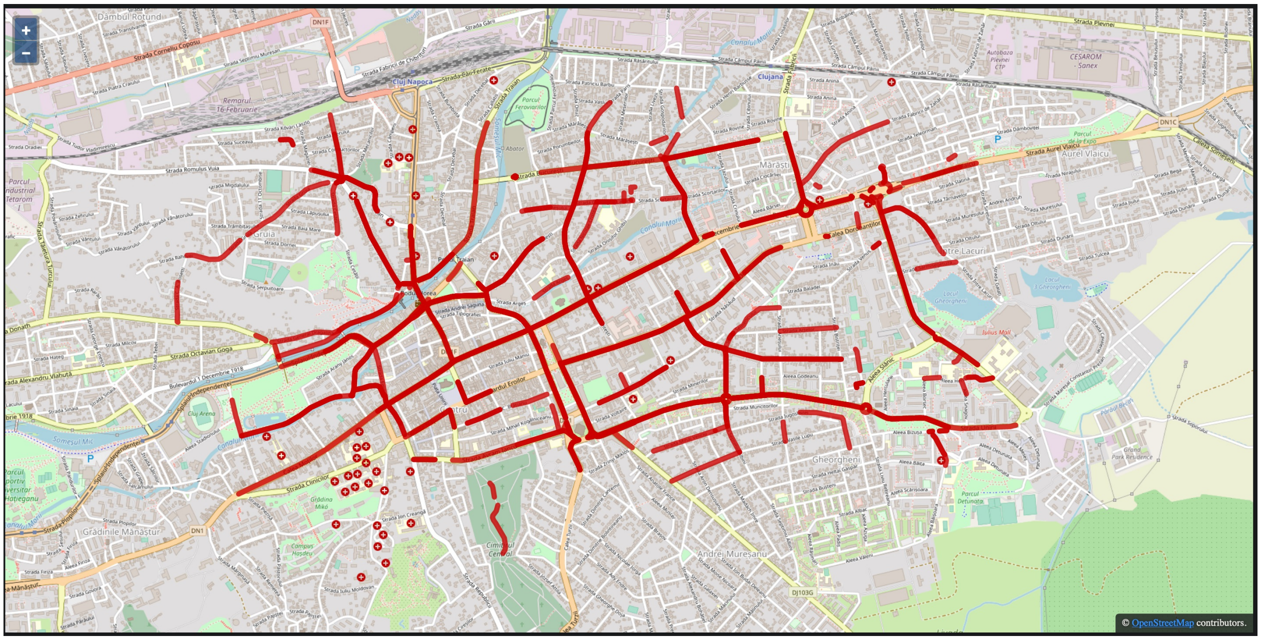


Fig. 2. Screenshot of the web interface showing the running simulation with a larger number of actors

design choices paragraph) but both of them are transparently routed by Docker.

```
conn, err := net.Dial("tcp", "citysim:7450")
if err != nil {
    fmt.Println("Error on dialing", err)
    break
}
defer conn.Close()
```

Listing 6: City actor dialing City simulator

```
var startWebsocket = function (callback) {

    var ws = new WebSocket("ws://citysim:9000/city")

    ws.onopen = function(evt) {
        console.log("OPEN");
        ws.send("just sent some messageeeee")
    }
    ws.onclose = function(evt) {
        console.log("CLOSE");
        ws = null;
    }
    ws.onmessage = function(evt) {
        callback(evt.data);
    }
    ws.onerror = function(evt) {
        console.log("ERROR: " + evt.data);
    }
}
```

Listing 7: City front dialing City simulator

V. Results and evaluation

Fig. 2 shows a screenshot of the web interface discussed in paragraph III-B, showing a running simulation. The web interface is a simple: a html/js only web page (vanilla js) which connects to the city simulator as can be seen in listing 7. The city simulator will send a notification to the web page server each time the density changes for a certain line. It can be that a new or first actor entered a certain line (a list of coordinates which represent a street) or it can be that one or the last actor left a line. The web page will draw any line with a density greater than 0 on top of the map.

The map is sourced from Open Street Map [7] and is showing the map zoomed to a specific city, same used by the city actors in their walkings. Thus, whenever a city actor reports that it is traveling across a line and this information arrives at the web interface through the city simulator, the respective line will be immediately drawn into the view.

A laptop used for simulating (MacBook Pro, 16 GB RAM, i5 3.1 GHz) is able to run easily 50 city actors. Above a certain threshold the bottleneck, surprisingly, is not the memory or the cpu but the way docker handles network interfaces. For a larger number of city actors a better tool must be used (see paragraph IV-D). Fig. 2 shows a simulation with 50 actors running at the same time throughout the city.

VI. Conclusions and Future Work

The concept of modeling or simulating traffic using actors or agents is widespread. The advantages it offers, compared to macroscopic simulation techniques, is that more nuanced data can be obtain, usually by simulating real events and situations more easier when the agents can interact and influence each other. On top of the actor (or agent, or entity) based model, the current paper, compared with the other noted beforehand, has shown the benefits of using today's concepts of distributed systems: containers and orchestration. Those ideas (or concepts or programming frameworks), not novel in their basic traits but novel in the ease of use, enable the scaling of simulation needs across many more physical machines. In parallel with this the code being written do not suffer from leaky abstractions of the distributed nature it runs in. It is kept slim and focused, while the frameworks used are taking care of all the other networking details.

Microservices concept has been presented and together with it the city simulator and the city actor types have been introduced. Their interactions, through a specific protocol, have been presented. Moreover, various other interactions, with supporting roles (routing engine and web page) have been presented: the routing engine can be called by any actor via a REST based API and the web page keeps an open WebSocket connection towards the city simulator to be notified of any change. Design choices, lead by the programming language (Go language) and container tool (Docker) have been discussed, together with their advantages. Networking details and design have been detailed, mostly lead by how the Docker and Linux containers landscape work. Kubernetes, as the de facto solution for microservices orchestration was also presented. Finally, a brief description provided the results of a simulation run on a single machine.

As with all technical projects, there are a number of enhancements which can be further developed around the concept of simulating a city traffic through actors running in microservices, as presented in this paper. This section will briefly propose few of those details in the following paragraphs.

As the number of the actors are growing, a first bottlneck in the current architecture will be the city simulator. It is a single entity and it represents a single source of truth for the actors. To keep it as single source of truth but at the same time to scale it horizontally means that we need to introduce few additional concepts. In [18] a first idea might be of using multiple cloud providers for taking advantage of their available compute resources and also for combining them. This would serve both to the city simulator as well to the swarm of city actors. The other part which remains is to keep the city simulator as a single source of truth by taking advantage of a distributed database. Something similar can be seen in [19] where a NoSQL database is used for horizontally scaling. To take advantage of the Kubernetes cluster there are similar horizontally scalable databases which can be built upon. One option is the distributed key-value store used by Kubernetes itself: Etcd database [20].

A second idea for enhancing the simulation is to make the actors to have a more granular and diverse logic of traversing the city. The solution proposed into this paper is a basic one: an actor has a route between two random points on the map and then it goes on to travel across that route. It can also be made to look for street densities in advance and act accordingly. If a street is too crowded the actor might choose to avoid it and compute a new route around it. A set of actors can be made to always run between same points (to simulate, for example, buses).

A third idea is to add more interactions between actors. It has been discussed that a first interaction is an indirect one: the number of actors on a certain street, a number called density. But there are many more points of interactions. Let's take junctions. The city simulator can model any junction and then it can permit actors to pass or not through it, while at the same time incorporating real data as shown in [3]. Then there are pedestrian crossings where different types of actors can meet each other. Or subways access, elevators for them, bus stations, etc.

A fourth idea is to have contained actors. A bus is an actor. A person is an actor. If a person is waiting on a bus station and then it takes a bus, there are two actors, but they are tied together, or contained and are not independent anymore until the actor-person choose to get out of the actor-bus. This means the densities across sidewalkings and bus stations need to be computed accordingly.

A fifth idea of enhancement is not tied to the way simulation is working but to how it is presented. The web page can display different colors based on the density of a line, offering a visual clue on it. Also it can easily show more information when the cursor is over a point of interest (a street, a junction, etc.) or it can speed up a past simulation and it can display an animation.

References

[1] R. Alqurashi and T. Altman, "Hierarchical agent-based modeling for improved traffic routing," *Applied Sciences*, vol. 9, p. 4376, 10 2019.

[2] A. Keler, J. Kaths, F. Chucholowski, M. Chucholowski, G. Grigoropoulos, M. Spangler, H. Kaths, and F. Busch, "A bicycle simulator for experiencing microscopic traffic flow simulation in urban environments," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Nov 2018, pp. 3020–3023.

[3] H. Sutarto, "Urban traffic simulation using sumo open source tools," 01 2016.

[4] J. Brügmann, M. Schreckenberg, and W. Luther, "Real-time traffic information system using microscopic traffic simulation," pp. 448–453, Sep. 2013.

[5] Linux namespaces, "Linux namespaces — Wikipedia, the free encyclopedia," 2002, [Online; accessed 29-October-2019]. [Online]. Available: https://en.wikipedia.org/wiki/Linux_namespaces

[6] Docker inc., "Docker," 2019. [Online]. Available: www.docker.com

[7] OpenStreetMap community, "Open street map," 2019. [Online]. Available: www.openstreetmap.org

[8] GraphHopper community, "Graphhopper," 2019. [Online]. Available: www.graphhopper.com

[9] A. A. Donovan and B. W. Kernighan, *The Go programming language*. Addison-Wesley Professional, 2015.

[10] C. Pahl, A. Brogi, J. Soldani, and P. Jamshidi, "Cloud container technologies: A state-of-the-art review," *IEEE Transactions on Cloud Computing*, vol. 7, no. 3, pp. 677–692, July 2019.

[11] Cloud Native Computing Foundation, "Kubernetes," 2019. [Online]. Available: www.kubernetes.io

[12] B. Burns, B. Grant, D. Oppenheimer, E. Brewer, and J. Wilkes, "Borg, omega, and kubernetes," 2016.

[13] D. Bernstein, "Containers and cloud: From lxc to docker to kubernetes," *IEEE Cloud Computing*, vol. 1, no. 3, pp. 81–84, Sep. 2014.

[14] K. Hightower, B. Burns, and J. Beda, *Kubernetes: up and running: dive into the future of infrastructure.* " O'Reilly Media, Inc.", 2017.

[15] P. Jogalekar and M. Woodside, "Evaluating the scalability of distributed systems," *IEEE Transactions on parallel and distributed systems*, vol. 11, no. 6, pp. 589–603, 2000.

[16] C. Risley, R. Lamb, and E. Guzovsky, "Domain name system lookup allowing intelligent correction of searches and presentation of auxiliary information," Dec. 18 2001, uS Patent 6,332,158.

[17] E. S.-J. Swildens, R. D. Day, and V. Garg, "Scalable domain name system with persistence and load balancing," Apr. 18 2006, uS Patent 7,032,010.

[18] C. Liu, M. Shie, Y. Lee, Y. Lin, and K. Lai, "Vertical/horizontal resource scaling mechanism for federated clouds," pp. 1–4, May 2014.

[19] A. Naskos, A. Gounaris, and I. Konstantinou, "Elton: A cloud resource scaling-out manager for nosql databases," pp. 1641–1644, April 2018.

[20] etcd community, "etcd," 2019. [Online]. Available: www.etcd.io

Genres and Actors/Actresses as Interpolated Tags for Improving Movie Recommender Systems

Nghia Duong-Trung[1], Quynh Nhut Nguyen[2], Dung Ngoc Le Ha[3], Xuan Son Ha[4], Tan Tai Phan[5], and Hiep Xuan Huynh[6]

[1, 3]Can Tho University of Technology, Can Tho city, Vietnam
[1, 4]FPT University, Can Tho city, Vietnam
[2]Vietnam Posts and Telecommunications Group (VNPT)
[5]National Chiao Tung University, Hsinchu, Taiwan
[6]Can Tho University, Can Tho city, Vietnam

***Abstract*—A movie recommender system has been proven to be a convincing implement on carrying out comprehensive and complicated recommendation which helps users find appropriate movies conveniently. It follows a mechanism that a user can be accurately recommended movies based on other similar interests, e.g. collaborative filtering, and the movies themselves, e.g. content-based filtering. Therefore, the systems should come with predetermined information either by users or by movies. One interesting research question should be asked: "what if this information is missing or not manually manipulated?" The problem has not been addressed in the literature, especially for the 100K and 1M variations of the MovieLens datasets. This paper exploits the movie recommender system based on movies' genres and actors/actresses themselves as the input tags or tag interpolation. We apply tag-based filtering and collaborative filtering that can effectively predict a list of movies that is similar to the movie that a user has been watched. Due to not depending on users' profiles, our approach has eliminated the effect of the cold-start problem. The experiment results obtained on MovieLens datasets indicate that the proposed model may contribute adequate performance regarding efficiency and reliability, and thus provide better-personalized movie recommendations. A movie recommender system has been deployed to demonstrate our work. The collected datasets have been published on our Github repository to encourage further reproducibility and improvement.**

***Keywords*—Movielens; movie recommender systems; tag interpolation; colloborative filtering**

I. Introduction

Recommender systems (RSs) have been developed to generate meaningful recommendations any products or items to a group of users that might get their attention. RSs [1], [2] are now widely use in research [3], industry [4], and education community [5], [6], where many approaches have been developed for improving recommendations. Many real world examples of recommendation operation can be found for books on Amazon [7], music on Spotify [8], activities on social media [9], [10], services on Twitter [11], [12], or movies on Netflix [13]. The design of these systems depends on the particular characteristics of the datasets, e.g. the ratings of 1 (most disliked) to 5 (most liked). Additionally, the systems might incorporate other information such as descriptions, multimedia contents, and demographic knowledge. Such data sources capture the interactions between items-items, users-users, and users-items. Recommender systems then analyze and learn the underlying patterns in these data sources to develop a correlation between users/items and or items/users which can be used to predict similar pairs. The architecture and evaluation of RSs are an active research area. The infinite solutions to RSs can be categorized into several categorizations. Content-based recommendation models recommend items that are similar to the items that a user has interacted in the past. The second approach is collaborative filtering that recommends items based on all users' past ratings collectively. Tag-aware recommendation [14], [15] approaches the interaction among items that are independent of the existence of users. Our contribution enhances the current research on the tag-based recommendation [16], [17]. Application of tag-based recommendation have been exploited in various domains from personalized social media services [18], e-learning environments [19], personalized location recommendation [20], image search [21], personalized news recommendation [22], personalized music recommendation [23], [24], and many others. A fourth category is a hybrid approach that combines two or more of the previously mentioned categories [25], [26].

Content-based recommendation [27], [28] is the proposal of items based on a comparison between the content of data items and/or user profiles. The content of each item is presented as a set of descriptions, lists of terms or tags, often words that appear in the textual form. The recommended items are primarily related to the items that are relatively rated as a recommendation. Content suggestions use different types of models to find similarities between sources to create the best proposal. The term collaborative filtering (CF) was introduced in a commercial recommender system that recommends newsgroups documents to users [29]. CF analyzes data interaction across users to find matching patterns resulting in other items recommendation [30]. The cold-start problem arises in CF systems where users exist and they have not rated several items before. The motivation of CF is to leverage social collaboration recommend the most similar items/products/services despite a large amount of data. Applications of CF have been developed in a wide range of domains from recommending books [31], musics [32], movies [33], advertisements [34] and other consumer products [35].

Twenty years of MovieLens datasets have witnessed a blossom of research that is garnering a remarkable significance with the advent of e-commerce and the whole industry. Variations of the dataset have been downloaded hundreds of thousands of times, reflecting their popularity and distinctive contribution in the field of recommendation systems and connected subjects. The samples take the form of *<user, item, rating, timestamp>* tuples where each tuple represents a personal preference for a movie at a particular time. A report made by their inventors shows that more than 7500 references to the keyword *movielens* have been made in Google Scholar [36]. A live research system[1] of the MovieLens datasets has been developed and maintained by experts to nurture the personalization and recommendation research. GroupLens research group[2] developed MovieLens as an online movie recommendation system that allows users to rate movies and integrates rating from different sources to collaboratively recommend to other people. Averaged 20-30 new users have signed every day for a long period. This system allows people to create profiles, rate movies, establish tastes and receive recommendations.

Our approach interpolates genres and actors/actresses as tags that predict similarity among movies and provide an appropriate suggestion. We investigate the two-way interactions between $\{item_i, [tags]_i\}$ and $\{item_j, [tags]_j\}$, where an item i is similar to an item j using the similarity score between their two tags [25], [37]. Practically, tags are collected from users' annotations during the involvement of a recommender system. However, what if the information is missing or does not exist in the first place? Table I presents a quantitative summary of the MovieLens datasets in which the first two variations of the datasets contain no tag information. In this paper, the authors consider another principle design of a movie recommender system: watching movies containing similar genres and actors/actresses (as other movies) lead to watching more same movie categories, which leads to an approach called tag interpolation-based recommendation. We have evaluated the proposed approach on Movielens' variations that contains no manual and/or collected tags from users. Instead, the tags come from movies' genres and actors/actresses. To the best of our knowledge, the research on a movie recommender system based on tags has never been done on the MovieLens 100K and MovieLens 1M variations.

TABLE I. THE QUANTITATIVE SUMMARY OF VARIATIONS OF THE MOVIELENS DATASETS.

Variation	# Users	# Movies	# Ratings	# Tags
MovieLens 100K	943	1,682	100,000	**0**
MovieLens 1M	6,040	3,706	1,000,209	**0**
MovieLens 10M	69,878	10,681	10,000,054	95,580
MovieLens 20M	138,493	27,278	20,000,263	465,564

II. RELATED WORK

One of the early attempts to develop a model and build a movie recommender system has been proposed by Azaria *et al.* [38]. In that paper, the authors introduce the profit and utility maximizer algorithm (PUMA) which mounts a black-boxed movie recommender system and predicts movies that will maximize the system's revenue. Another research direction focus on human emotions as the input for movie recommendation [39]. The approach accepts the user profile as part of the system. Deldjoo *et al.* introduce multimodal content-based movie recommender system [40] that is evaluated on the MovieLens 20M dataset. They exploit the effects of genres as the metadata feature. However, the tags have already provided in MovieLens 20M. The genre features have been further addressed by the same research team of Deldjoo [41]. Another interesting paper that focuses on tag-aware recommendation and the effects of tags over a recommender system is presented in [42]. In that paper, the authors investigated tags from genres and textual reviews on the tag-available dataset, e.g. MovieLens 10M. These models have one thing in common that they perform the recommendation task on tag-available datasets with extra agglomeration from other sources, e.g. genres, users' information, and textual reviews. Our approach differs from these previous ones by the fact that the tags are automatically interpolated from genres and actors/actresses, without any additional manual effort from the users' side and predetermined tags. Consequently, this work is the first to exploit tag interpolation in MovieLens 100K and MovieLens 1M datasets and can be furthered referred for tag-based recommendations.

III. EXPERIMENTS

A. Datasets

As mentioned in the previous section, the authors employ the MovieLens 100K and MovieLens 1M variations in the experiments because there are no pre-defined tags, but instead, tags are interpolated from the movies' genres and actors/actresses. The datasets [36] can be downloaded on the MovieLens 100K[3] and MovieLens 1M websites[4]. The MovieLens 100K dataset consists of 100,000 ratings (from 1 to 5) from 943 users on 1,682 movies. Each user has rated at least 20 movies. The MovieLens 1M dataset comprises 1,000,209 ratings (from 1 to 5) from 6,040 users on 3,706 movies. For each dataset, the training and test sets have been already split into five-fold cross-validation. The authors run the proposed model on all sets and take an average in the end.

Interpolated tags. We could identify 19 different genres in both MovieLens variations. These tags can be easily extracted from the *u.genre* file of each MovieLens dataset. Regarding actors/actresses, these tags are not included explicitly. Instead, the authors link the movies of MovieLens dataset with their corresponding web pages at Internet Movie Database (IMDb)[5] from the *u.item* file of each MovieLens dataset, and extract actors/actresses from the IMDB database. One movie in 100K variation contains at least 1 and at most 45 actors/actresses while one movie in MovieLens 1M consists of at least 2 and at most 235 actors/actresses. The number of 14291 and 46198 actors and actresses can be extracted from MovieLens 100K and 1M respectively. The summary of interpolated tags is presented in Table II. The top 10 most used interpolated tags are summarized in Tables (III and IV). The authors make the collected datasets available at our Github repository[6]. We encourage reproducibility, further comparison and improvement.

[1] https://movielens.org/

[2] https://grouplens.org/

[3] https://grouplens.org/datasets/movielens/100k/

[4] https://grouplens.org/datasets/movielens/1m/

[5] https://www.imdb.com/

[6] https://github.com/nnquynh4496/MovieLens-Tags-Interpolation

TABLE II. INTERPOLATED TAGS FROM MOVIELENS VARIATIONS.

Variation	# Genres	# Actors/actresses
MovieLens 100K	19	14,291
MovieLens 1M	19	46,198

TABLE III. THE NUMBER OF APPEARANCES THE MOST 10 POPULAR GENRES AND ACTOR/ACTRESSES AS INTERPOLATED TAGS IN MOVIELENS 100K DATASET.

Nr.	MovieLens 100K			
	Genres	Freq.	Actors/Actresses	Freq.
1	Drama	725	Samuel L. Jackson	21
2	Comedy	505	Robert De Niro	19
3	Action	251	Steve Buscemi	17
4	Thriller	251	Christopher Walken	14
5	Romance	247	Meg Ryan	14
6	Adventure	135	Christopher McDonald	14
7	Children's	122	Robert Duvall	14
8	Crime	109	Harrison Ford	13
9	Sci-Fi'	101	Gwyneth Paltrow	13
10	Horror	92	Anthony Hopkins	13

TABLE IV. THE NUMBER OF APPEARANCES THE MOST 10 POPULAR GENRES AND ACTOR/ACTRESSES AS INTERPOLATED TAGS IN MOVIELENS 1M DATASET.

Nr.	MovieLens 1M			
	Genres	Freq.	Actors/Actresses	Freq.
1	Drama	1633	Samuel L. Jackson	25
2	Comedy	1218	Joan Cusack	25
3	Action	508	M. Emmet Walsh	25
4	Thriller	495	James Stewart	24
5	Romance	482	Christopher McDonald	24
6	Horror	344	Dan Hedaya	24
7	Adventure	289	Robert Duvall	23
8	Sci-Fi'	276	Frank Welker	23
9	Crime	216	Whoopi Goldberg	22
10	War	159	Robert De Niro	21

B. Evaluation Metric

Root mean squared error (RMSE) and mean absolute error (MAE) are widely used to evaluate the performance of a recommender system given a rating prediction task. The errors quantify the difference between the true rating values and the predicted rating values made by the recommender system. In this work, the authors evaluate the performance of the system by RMSE. We denote r_{ij} and $\hat{r}_{ij}$ as the true rating value and the predicted rating value respectively. Then the RMSE e is calculated as follows:

$$e = \sqrt{\frac{1}{n}\sum_{i,j}(r_{ij} - \hat{r}_{ij})^2} \, , \qquad (1)$$

where the smaller the e is, the better the result is.

Furthermore, based on the computed rating scores of movies, the similarity between any two movies u and v is calculated by using cosine similarity $c_{i,j}$ as follows.

$$c_{u,v} = \cos(\overrightarrow{r_u}, \overrightarrow{r_v}) = \frac{\sum_{i=1}^{m} r_{u,i} r_{v,i}}{\sqrt{\sum_{i=1}^{m} r_{u,i}^2}\sqrt{\sum_{i=1}^{m} r_{v,i}^2}} \, , \qquad (2)$$

where m is the dimensional space of u and v. The list of recommended movies is sorted by values of $c_{u,v}$.

Equation (1) is used to evaluate the performance of the recommendation system. Based on the list of recommended movies created by Equation (2), the system compares the most similar movie's ratings with its prediction by Equation (1). However, in the system deployment presented in Section (IV), the list of recommended movies is more important to the users than the RMSE scores. Hence, the calculation of Equation (1) is ignored in real-time.

C. Experimental Results

The MovieLens 100K and MovieLens 1M datasets contain information on several meta-data such as genres and the names of actors/actresses. The information is considered as the feature of movie observations. We describe the experimental results in the following three scenarios.

1) Scenario 1: Tags are interpolated from the movies' genres.: In this scenario, the authors investigate the performance of the recommender system by utilizing genres as the tags. The experimental results are presented in Tables V and VIII for MovieLens 100K and MovieLens 1M respectively.

2) Scenario 2: Tags are interpolated from the movies' actors/actresses.: Scenario 2 is about the effect of actors/actresses tags to the system's performance. The experimental results are presented in Tables VI and IX for MovieLens 100K and MovieLens 1M respectively.

3) Scenario 3: Tags are interpolated from the movies' combination of genres and actors/actresses.: In the last experimental scenario, the authors combine the tags of both genres and actors/actresses. Tables VII and X presents the system's performance in this scenario.

TABLE V. THE SUMMARY OF PREDICTION RESULTS AND RUNNING TIMES ON THE MOVIELENS 100K. TAGS ARE THE MOVIES' GENRES. – MEANS NO MEASUREMENT.

No.	Train-test sets	RMSE	Running times (ms)
1	u1.base	–	851.67
	u1.test	1.1672	
2	u2.base	–	816.51
	u2.test	1.1373	
3	u3.base	–	900.46
	u3.test	1.1298	
4	u4.base	–	802.54
	u4.test	1.1455	
5	u5.base	–	782.53
	u5.test	1.1489	
Average score		**1.1457 ± 0.0126**	**830.74 ± 41.52**

TABLE VI. THE SUMMARY OF PREDICTION RESULTS AND RUNNING TIMES ON THE MOVIELENS 100K. TAGS ARE THE MOVIES' ACTORS/ACTRESSES. – MEANS NO MEASUREMENT.

No.	Train-test sets	RMSE	Running times (ms)
1	u1.base	–	47,561.84
	u1.test	1.0626	
2	u2.base	–	48,091.09
	u2.test	1.0472	
3	u3.base	–	52,400.74
	u3.test	1.0343	
4	u4.base	–	50,335.71
	u4.test	1.0960	
5	u5.base	–	48,721.73
	u5.test	1.0382	
Average score		**1.0556 ± 0.0250**	**49,422.22 ± 1,964.03**

TABLE VII. THE SUMMARY OF PREDICTION RESULTS AND RUNNING TIMES ON THE MOVIELENS 100K. TAGS ARE THE MOVIES' THE COMBINATION OF GENRES AND ACTORS/ACTRESSES. – MEANS NO MEASUREMENT.

No.	Train-test sets	RMSE	Running times (ms)
1	u1.base	–	47,561.84
	u1.test	1.0626	
2	u2.base	–	48,091.09
	u2.test	1.0472	
3	u3.base	–	52,400.74
	u3.test	1.0343	
4	u4.base	–	48,073.16
	u4.test	1.0960	
5	u5.base	–	51,526.38
	u5.test	1.0382	
Average score		**1.0556 ± 0.0250**	**49,530.64 ± 2,252.39**

TABLE VIII. THE SUMMARY OF PREDICTION RESULTS AND RUNNING TIMES ON THE MOVIELENS 1M. TAGS ARE THE MOVIES' GENRES. – MEANS NO MEASUREMENT.

No.	Train-test sets	RMSE	Running times (ms)
1	u1.base	–	22,400.17
	u1.test	1.0560	
2	u2.base	–	27,995.05
	u2.test	1.0206	
3	u3.base	–	24,450.00
	u3.test	1.0236	
4	u4.base	–	24,360.05
	u4.test	1.0236	
5	u5.base	–	23,687.43
	u5.test	1.0584	
Average score		**1.0364 ± 0.0190**	**24,578.54 ± 2078.23**

TABLE IX. THE SUMMARY OF PREDICTION RESULTS AND RUNNING TIMES ON THE MOVIELENS 1M. TAGS ARE THE MOVIES' ACTORS/ACTRESSES. – MEANS NO MEASUREMENT.

No.	Train-test sets	RMSE	Running times (ms)
1	u1.base	–	1,343,201.13
	u1.test	1.0354	
2	u2.base	–	1,594,259.98
	u2.test	1.0305	
3	u3.base	–	1,347,105.26
	u3.test	1.0324	
4	u4.base	–	1,379,111.58
	u4.test	1.0324	
5	u5.base	–	1,779,768.24
	u5.test	1.0368	
Average score		**1.0335 ± 0.0025**	**1,488,689.23 ± 193,062.32**

TABLE X. THE SUMMARY OF PREDICTION RESULTS AND RUNNING TIMES ON THE MOVIELENS 1M. TAGS ARE THE MOVIES' THE COMBINATION OF GENRES AND ACTORS/ACTRESSES. – MEANS NO MEASUREMENT.

No.	Train-test sets	RMSE	Running times (ms)
1	u1.base	–	726,754.79
	u1.test	1.0354	
2	u2.base	–	811,877.18
	u2.test	1.0304	
3	u3.base	–	735,897.88
	u3.test	1.0323	
4	u4.base	–	779,554.22
	u4.test	1.0323	
5	u5.base	–	693,466.31
	u5.test	1.0368	
Average score		**1.0334 ± 0.0025**	**749,510.07 ± 46,465.86**

IV. SYSTEM DEPLOYMENT

A. System Design

The movie recommender system implements the Model-View-Template model during creating an application with user interaction. This model includes HTML codes with Django Templage Language [43]. A controller is written to control the interaction between Model and View. When a user requests, the controller processes the user's request using Model, View, and Template. It acts as a Controller to check if it is available by URL mapping and if the URL is successful. The View will start interacting with the Model and return the Template to the user as Response. The website is written in Django's default SQLite database and it also integrates a lightweight server for application development and testing.

The system provides functionality for two groups of users, e.g. the administrator(s) and the user(s). An administrator performs functions of managing users, users' information, movies, movies' information, databases, and suggestions. Administrators have the highest rights in the system that can perform addition, editing, deletion, and search for movies and users. Users are allowed to register, log in, search for movies and actors/actresses. The overview of our proposed movie recommender system is illustrated in Fig. 1.

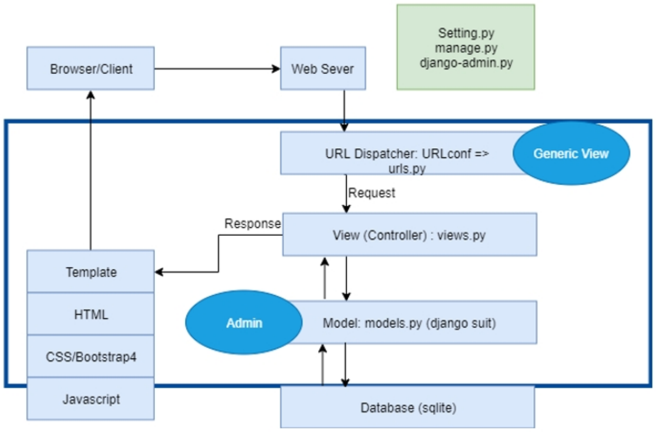


Fig. 1. The design of our movie recommender system.

B. System Implementation

As the implementation of our recommendation system, the authors have deployed a website for our movie recommender system[7]. The application has 15 preliminary features for both users and administrators. The functionality of our website is shown in Fig. 2. The database design is presented in Fig. 3. The website has been developed using Django framework[8] [44] and the relational database management system SQLite[9] [45]. A screenshot of our website can be seen in Fig. 4. The recommendation function is demonstrated in Fig. 5 where the watched movie is in the main position on the left and its list of similar movies is presented on the right. All the work of the model's training and prediction and the website's deployment are done on a normal laptop. The hardware configurations are the following: Intel Core i5, 12GB of RAM, 240GB high-speed SSD, and Windows 10.

Before using the system, users need to register an account without specifying their preferred movie genre. The interaction

[7] https://goiyphim.herokuapp.com/
[8] https://www.djangoproject.com/
[9] https://www.sqlite.org/index.html

with the recommender system can be done through the web interface. By watching any movies and rating them, user profiles are created. A list of recommended movies is generated every time a movie is watched. This system is deployed in real-time scenarios to generate an automatic recommendation.

V. Remarks and Discussion

The experimentation has been conducted on the MovieLens 100K and MovieLens 1M whose tags are missing originally. Remember to note that information of tags is only available in modern variations of the datasets, e.g. MovieLens 10M and MovieLens 20M. The authors interpolate the movie genres and actors/actresses as the tags. The experimental results lead us to believe that the proposed tag interpolation should work properly and yet improve the development of movie recommender systems whose tags are missing. We have achieved better RMSE scores as other approaches running on the tag-available MovieLens datasets [42]. From the experimental results conducted in [42] on a similar movie recommender system, we can agree that our proposed tag interpolation approach is more effective than probabilistic matrix factorization [46], collaborative topic regression [47], factorization machines [48], and regression latent factor model [49].

The RMSE scores are quite similar in all experimental scenarios. Regarding MovieLens 100K, the average score is achieved by 1.1457 ± 0.0126, 1.0556 ± 0.0250, and 1.0556 ± 0.0250 in case of genre tags, actors/actresses tags and the combination of genres and actors/actresses respectively. The running time is super fast in the case of movies' genres, e.g. less than 1 second. In case of MovieLens 1M, the RMSE scores are slightly better than those of 100K variation. the average score is achieved by 1.0364 ± 0.0190, 1.0335 ± 0.0025, and 1.0334 ± 0.0025 in case of genre tags, actors/actresses tags and the combination of genres and actors/actresses respectively. The running times increase through the extension of the number of interpolated tags. The effect is quite understandable that the more data processed, the more times required.

VI. Conclusion

The prevalence of movie recommendation systems has been an indispensable component in a wide range of websites and e-commerce applications. And tag usability is increasing in many recommendation systems, yet appropriate algorithms are available to exploit these tags. This work addresses a simple research question: what if the tags are missing or do not exist in the first place? Therefore, tags can be interpolated from any other characteristics of the movies themselves. Our proposed approach makes it highly convenient for users to get meaningful movie recommendations. Several experimental scenarios have validated the effectiveness of our proposed solution. The significant contribution of the paper is to the MovieLens-based research where previous work has never done on the 100K and 1M variations. As we illustrated in our experimental results, the effects of genres and actors/actresses as interpolated tags have proved the effectiveness and applicability. We have implemented a complete movie recommender system with 15 preliminary functions for both users and system administrators. The interpolated-tags datasets are also available on our Github repository. Future work will focus on the implementation of datasets that emerge the similar characteristics.

References

[1] C. C. Aggarwal *et al.*, *Recommender systems*. Springer, 2016.

[2] P. Melville and V. Sindhwani, "Recommender systems," *Encyclopedia of Machine Learning and Data Mining*, pp. 1056–1066, 2017.

[3] C. He, D. Parra, and K. Verbert, "Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities," *Expert Systems with Applications*, vol. 56, pp. 9–27, 2016.

[4] X. Amatriain and J. Basilico, "Past, present, and future of recommender systems: An industry perspective," in *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 2016, pp. 211–214.

[5] A. C. Rivera, M. Tapia-Leon, and S. Lujan-Mora, "Recommendation systems in education: A systematic mapping study," in *International Conference on Information Theoretic Security*. Springer, 2018, pp. 937–947.

[6] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: introduction and challenges," in *Recommender systems handbook*. Springer, 2015, pp. 1–34.

[7] B. Smith and G. Linden, "Two decades of recommender systems at amazon. com," *Ieee internet computing*, vol. 21, no. 3, pp. 12–18, 2017.

[8] M. Millecamp, N. N. Htun, Y. Jin, and K. Verbert, "Controlling spotify recommendations: effects of personal characteristics on music recommender user interfaces," in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, 2018, pp. 101–109.

[9] N. Duong-Trung, *Social Media Learning: Novel Text Analytics for Geolocation and Topic Modeling*. Cuvillier Verlag, 2017.

[10] A. Anandhan, L. Shuib, M. A. Ismail, and G. Mujtaba, "Social media recommender systems: review and open research issues," *IEEE Access*, vol. 6, pp. 15 608–15 628, 2018.

[11] N. Duong-Trung and L. Schmidt-Thieme, "On discovering the number of document topics via conceptual latent space," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 2051–2054.

[12] N. Duong-Trung, N. Schilling, and L. Schmidt-Thieme, "Near real-time geolocation prediction in twitter streams via matrix factorization based regression," in *Proceedings of the 25th ACM international on conference on information and knowledge management*, 2016, pp. 1973–1976.

[13] C. A. Gomez-Uribe and N. Hunt, "The netflix recommender system: Algorithms, business value, and innovation," *ACM Transactions on Management Information Systems (TMIS)*, vol. 6, no. 4, p. 13, 2016.

[14] H. Kim and H.-J. Kim, "A framework for tag-aware recommender systems," *Expert Systems with Applications*, vol. 41, no. 8, pp. 4000–4009, 2014.

[15] T. Bogers, "Tag-based recommendation," in *Social Information Access*. Springer, 2018, pp. 441–479.

[16] X. Shi, H. Huang, S. Zhao, P. Jian, and Y.-K. Tang, "Tag recommendation by word-level tag sequence modeling," in *International Conference on Database Systems for Advanced Applications*. Springer, 2019, pp. 420–424.

[17] S. Tang, Y. Yao, S. Zhang, F. Xu, T. Gu, H. Tong, X. Yan, and J. Lu, "An integral tag recommendation model for textual content," 2019.

[18] M. Rawashdeh, M. F. Alhamid, J. M. Alja'am, A. Alnusair, and A. El Saddik, "Tag-based personalized recommendation in social media services," *Multimedia Tools and Applications*, vol. 75, no. 21, pp. 13 299–13 315, 2016.

[19] A. Klašnja-Milićević, B. Vesin, M. Ivanović, Z. Budimac, and L. C. Jain, "Folksonomy and tag-based recommender systems in e-learning environments," in *E-Learning systems*. Springer, 2017, pp. 77–112.

[20] Y. Zheng, Y. Wang, L. Zhang, J. Wang, and Q. Qi, "A tag-based integrated diffusion model for personalized location recommendation," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 327–337.

[21] D. Lu, X. Liu, and X. Qian, "Tag-based image search by social re-ranking," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1628–1639, 2016.

[22] Y. Shen, P. Ai, Y. Xiao, W. Zheng, and W. Zhu, "A tag-based personalized news recommendation method," in *2018 14th International*

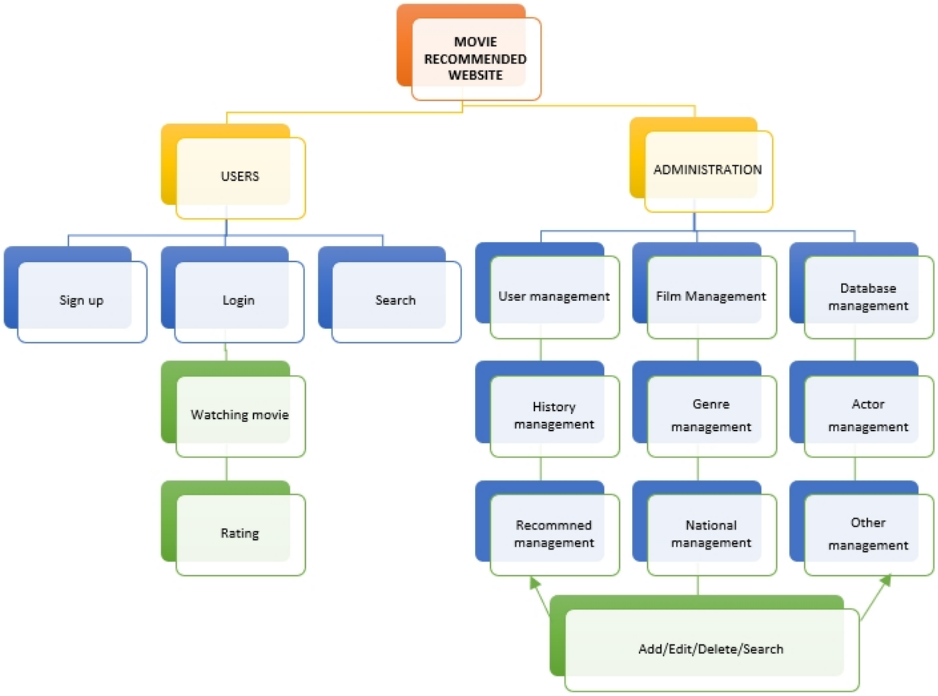


Fig. 2. The functionality of our deployed movie recommender system.

Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE, 2018, pp. 964–970.

[23] C. Musto, G. Semeraro, P. Lops, M. De Gemmis, and F. Narducci, "Leveraging social media sources to generate personalized music playlists," in *International Conference on Electronic Commerce and Web Technologies*. Springer, 2012, pp. 112–123.

[24] J.-H. Su, W.-Y. Chang, and V. S. Tseng, "Personalized music recommendation by mining social media tags," *Procedia Computer Science*, vol. 22, pp. 303–312, 2013.

[25] S. Wei, X. Zheng, D. Chen, and C. Chen, "A hybrid approach for movie recommendation via tags and ratings," *Electronic Commerce Research and Applications*, vol. 18, pp. 83–94, 2016.

[26] K. N. Jain, V. Kumar, P. Kumar, and T. Choudhury, "Movie recommendation system: hybrid information filtering system," in *Intelligent Computing and Information and Communication*. Springer, 2018, pp. 677–686.

[27] P. Lops, D. Jannach, C. Musto, T. Bogers, and M. Koolen, "Trends in content-based recommendation," *User Modeling and User-Adapted Interaction*, vol. 29, no. 2, pp. 239–249, 2019.

[28] M. F. Alhamid, M. Rawashdeh, M. A. Hossain, A. Alelaiwi, and A. El Saddik, "Towards context-aware media recommendation based on social tagging," *Journal of Intelligent Information Systems*, vol. 46, no. 3, pp. 499–516, 2016.

[29] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–71, 1992.

[30] Y. Koren, "Collaborative filtering with temporal dynamics," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 447–456.

[31] P. Mathew, B. Kuriakose, and V. Hegde, "Book recommendation system through content based and collaborative filtering method," in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*. IEEE, 2016, pp. 47–52.

[32] M. Gong, J. Hong, and J. Choi, "Music information and musical propensity analysis, and music recommendation system using collaborative filtering," in *Proceedings of the Korean Institute of Information and Commucation Sciences Conference*. The Korea Institute of Information and Commucation Engineering, 2015, pp. 533–536.

[33] R. Bharti and D. Gupta, "Recommending top n movies using content-based filtering and collaborative filtering with hadoop and hive framework," in *Recent Developments in Machine Learning and Data Analytics*. Springer, 2019, pp. 109–118.

[34] R. Raina, G. Rajaram, H. Ge, J. Pan, and J. Hegeman, "Selecting advertisements for users of a social networking system using collaborative filtering," Jun. 20 2013, uS Patent App. 13/330,502.

[35] H. S. Moon, Y. U. Ryu, and J. K. Kim, "Enhanced collaborative filtering: A product life cycle approach," *Journal of Electronic Commerce Research*, vol. 20, no. 3, pp. 155–168, 2019.

[36] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *Acm transactions on interactive intelligent systems (tiis)*, vol. 5, no. 4, p. 19, 2016.

[37] P. Rinearson and E. Speckman, "Relevancy rating of tags," May 31 2011, uS Patent 7,953,736.

[38] A. Azaria, A. Hassidim, S. Kraus, A. Eshkol, O. Weintraub, and I. Netanely, "Movie recommender system for profit maximization," in *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 2013, pp. 121–128.

[39] K. Wakil, R. Bakhtyar, K. Ali, and K. Alaadin, "Improving web movie recommender system based on emotions," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 2, pp. 218–226, 2015.

[40] Y. Deldjoo, M. G. Constantin, H. Eghbal-Zadeh, B. Ionescu, M. Schedl, and P. Cremonesi, "Audio-visual encoding of multimedia content for enhancing movie recommendations," in *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 2018, pp. 455–459.

[41] Y. Deldjoo, M. Schedl, and M. Elahi, "Movie genome recommender: A novel recommender system based on multimedia content," in *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2019, pp. 1–4.

[42] C. Zhang, K. Wang, E.-p. Lim, Q. Xu, J. Sun, and H. Yu, "Are features equally representative? a feature-centric recommendation," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[43] C. Burch, "Django, a web framework using python: tutorial presentation," *Journal of Computing Sciences in Colleges*, vol. 25, no. 5, pp. 154–155, 2010.

[44] N. Idris, C. F. M. Foozy, and P. Shamala, "A generic review of web

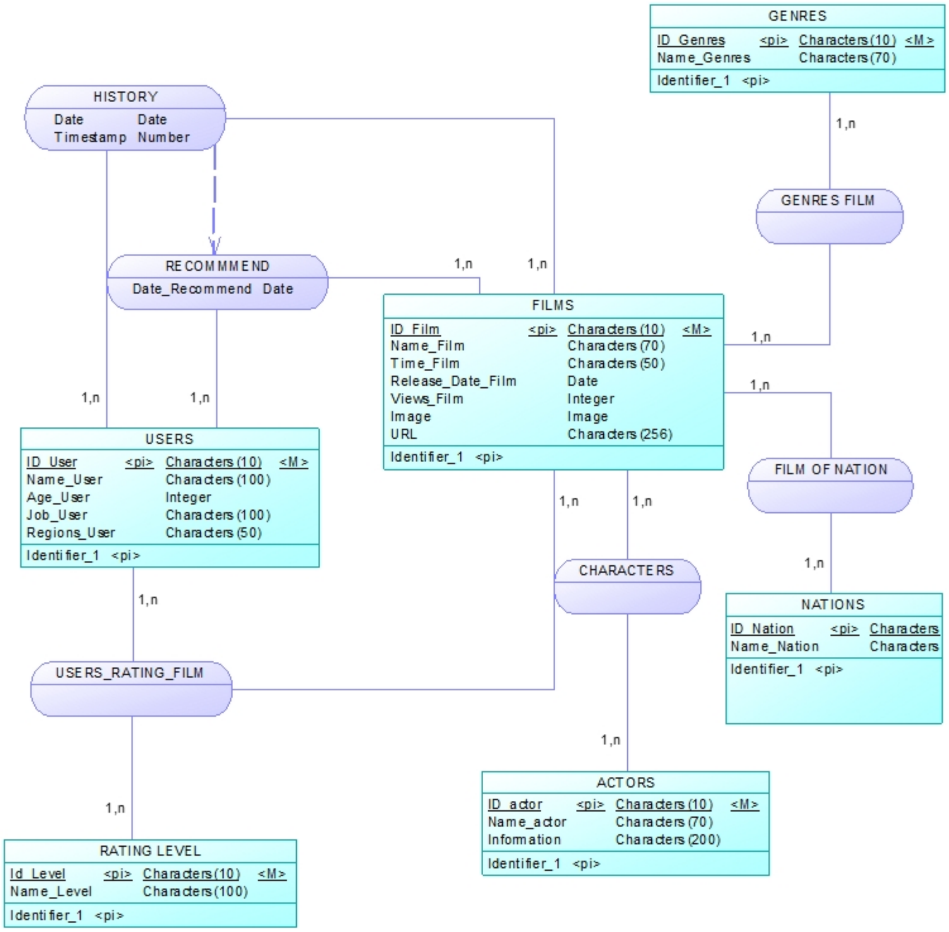

GENRES
ID_Genres <pi> Characters (10) <M>
Name_Genres Characters (70)
Identifier_1 <pi>
HISTORY
Date Date
Timestamp Number
GENRES FILM
RECOMMMEND
Date_Recommend Date
FILMS
ID_Film <pi> Characters (10) <M>
Name_Film Characters (70)
Time_Film Characters (50)
Release_Date_Film Date
Views_Film Integer
Image Image
URL Characters (256)
Identifier_1 <pi>
USERS
ID_User <pi> Characters (10) <M>
Name_User Characters (100)
Age_User Integer
Job_User Characters (100)
Regions_User Characters (50)
Identifier_1 <pi>
FILM OF NATION
CHARACTERS
NATIONS
ID_Nation <pi> Characters
Name_Nation Characters
Identifier_1 <pi>
USERS_RATING_FILM
ACTORS
ID_actor <pi> Characters (10) <M>
Name_actor Characters (70)
Information Characters (200)
Identifier_1 <pi>
RATING LEVEL
Id_Level <pi> Characters (10) <M>
Name_Level Characters (100)
Identifier_1 <pi>
1,n

Fig. 3. The database design of our deployed movie recommender system.

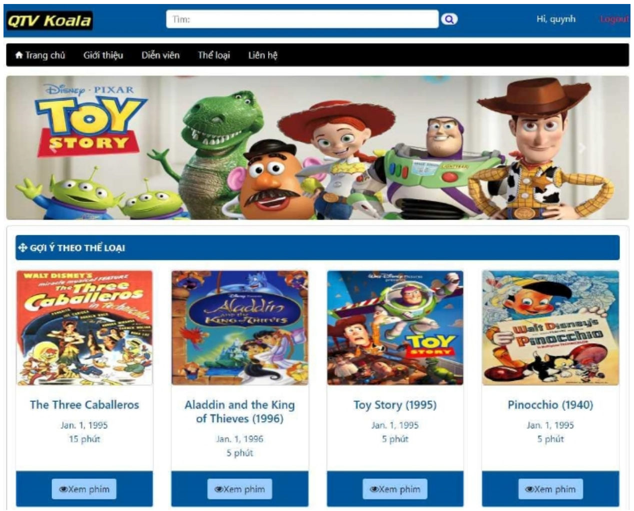
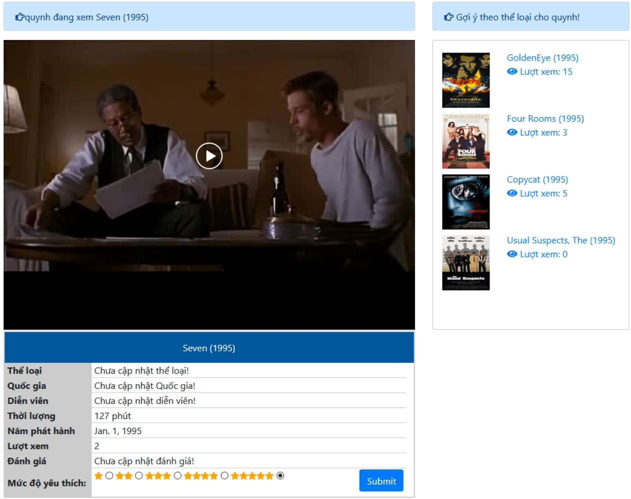


Fig. 4. A logged-in screenshot of our deployed movie recommender system.


quynh đang xem Seven (1995)

Gợi ý theo thể loại cho quynh!

GoldenEye (1995)
Lượt xem: 15

Four Rooms (1995)
Lượt xem: 3

Copycat (1995)
Lượt xem: 5

Usual Suspects, The (1995)
Lượt xem: 0

Seven (1995)

Thể loại	Chưa cập nhật thể loại!
Quốc gia	Chưa cập nhật Quốc gia!
Diễn viên	Chưa cập nhật diễn viên!
Thời lượng	127 phút
Năm phát hành	Jan. 1, 1995
Lượt xem	2
Đánh giá	Chưa cập nhật đánh giá!
Mức độ yêu thích:	★ ○ ★★ ○ ★★★ ○ ★★★★ ○ ★★★★★ ● Submit

Fig. 5. A watched movie and its list of recommendation on the right side.

technology: Django and flask," *International Journal of Engineering Information Computing and Application*, vol. 1, no. 1, 2019.

[45] S. Bhosale, T. Patil, and P. Patil, "Sqlite: Light database system," *International Journal of Computer Science and Mobile Computing*, pp. 882–885, 2015.

[46] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Advances in neural information processing systems*, 2008, pp. 1257–1264.

[47] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proceedings of the 17th ACM SIGKDD*

international conference on Knowledge discovery and data mining. ACM, 2011, pp. 448–456.

[48] S. Rendle, “Factorization machines,” in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 995–1000.

[49] D. Agarwal and B.-C. Chen, “Regression-based latent factor models,” in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 19–28.

Document Length Variation in the Vector Space Clustering of News in Arabic: A Comparison of Methods

Abdulfattah Omar[1*]
College of Science & Humanities
Prince Sattam Bin Abdulaziz University, Saudi Arabia
Department of English, Faculty of Arts, Port Said University

Wafya Ibrahim Hamouda[2]
Department of Foreign Languages
Faculty of Education
Tanta University, Egypt

***Abstract*—This article is concerned with addressing the effect of document length variation on measuring the semantic similarity in the text clustering of news in Arabic. Despite the development of different approaches for addressing the issue, there is no one strong conclusion recommending one approach. Furthermore, many of these have not been tested for the clustering of news in Arabic. The problem is that different length normalization methods can yield different analyses of the same data set, and that there is no obvious way of selecting the best one. The choice of an inappropriate method, however, has negative impacts on the accuracy and thus the reliability of clustering performance. Given the lack of agreement and disparity of opinions, we set out to comprehensively evaluate the existing normalization techniques to prove empirically which one is the best for the normalization of text length to improve the text clustering performance of news in Arabic. For this purpose, a corpus of 693 stories representing different categories and of different lengths is designed. Data is analyzed using different document length normalization methods along with vector space clustering (VSC), and then the analysis on which the clustering structure agrees most closely with the bibliographic information of the news stories is selected. The analysis of the data indicates that the clustering structure based on the byte length normalization method is the most accurate one. One main problem, however, with this method is that the lexical variables within the data set are not ranked which makes it difficult for retaining only the most distinctive lexical features for generating clustering structures based on semantic similarity. As thus, the study proposes the integration of TF-IDF for ranking the words within all the documents so that only those with the highest TF-IDF values are retained. It can be finally concluded that the proposed model proved effective in improving the function of the byte normalization method and thus on the performance and reliability of news clustering in Arabic. The findings of the study can also be extended to IR applications in Arabic. The proposed model can be usefully used in supporting the performance of the retrieval systems of Arabic in finding the most relevant documents for a given query based on semantic similarity, not document length.**

Keywords*—Arabic; document length; news clustering; semantic similarity; TF-IDF; VSC

I. Introduction

Variation in document length is widely considered an important factor in the validity of text clustering applications. It is essential in clustering applications that all documents within a collection corpus are equally represented [1-3]. Documents in any given corpus, however, can vary considerably in length. As a result, this characteristic can adversely affect the validity and thus reliability of clustering results. In document clustering applications, measuring the semantic similarity within texts can be greatly influenced by vectors that have the largest values. It is a tradition of all the proximity measurements to be dominated by longer documents. In vector space clustering (VSC), the distance between any two documents is determined by their length and the magnitude of the angle between the vectors. This means that if the length of the document increases, the number of times a particular term occurs in the document also increases. Consequently, length becomes an increasingly important determinant of vector clustering in the space. Vice versa, if the documents are short, the angles between the vectors become smaller and as a sequence, short documents will be clustered together [3].

The issue of document length variation has its implications to all text clustering applications including data organization, information retrieval (IR), document retrieval, information filtering, machine learning, text summarization, authorship detection and recognition, and even marketing purposes. In IR applications, for instance, documents that are longer have a higher number of words, hence the values or frequencies for those words are increased, and a document highly relevant for a given term that happens to be short will not necessarily have that relevance reflected in its term frequencies. So if length variation is not considered, longer documents come first irrespective of their relevance to the query. Longer documents have higher term frequency values and naturally, they have—for length reasons more distinct terms. The length factor results in raising the scores of longer documents, which is unnatural. So under the scoring scheme, longer documents are favored simply because they have more terms [4].

Numerous techniques have been devised to account for the variation of length within documents. However, very little has been done in relation to the language processing of Arabic in general and Arabic news in particular. This study addresses this gap in the literature by proposing an integrated model that considers the linguistic peculiarities of Arabic. By way of illustration, a corpus of 693 stories representing different

Paper Submission Date: January 30, 2020
Acceptance Notification Date: February 12, 2020
*Corresponding Author

categories and of different lengths is designed. These represent different topics including politics, sports, family, environment, health, education, technology, and business. Seven normalization methods are compared to choose the best normalization method. These are byte length normalization, cosine normalization, maximum tf normalization, mean normalization, pivoted-cosine normalization, probability normalization, and Z-score. The remainder of this article is organized as follows. Section 2 defines the research problem. Section 3 is a brief survey of VSC and document length normalization methods. Section 4 outlines the data selection and creation processes, methods, and procedures. Section 5 is an analysis of the data using different document length normalization methods. Section 6 is the conclusion.

II. Statement of the Problem

With the explosion in the amount of news and journalistic content being generated in Arabic, there is an increasing need for more reliable clustering tools that can effectively classify raw texts to make it easier for users to identify topics, obtain the information that is relevant to their queries using content-driven groupings of articles. This has been done over recent years using different VSC methods. One problem with these methods, however, is document length variation which is a normal issue. In spite of the development of different techniques for addressing the problem of variation in document length, they cannot be universally applied to all languages. In other words, standard normalization systems of document length have traditionally ignored the issue of language peculiarities which has negative impacts on the validity and thus reliability of such methods. In natural language processing (NLP) of Arabic, the specific linguistic properties play a significant role in the success of NLP applications [5-8]. It is essential for NLP systems thus to consider the peculiarities of Arabic for more reliable results. Furthermore, there is no agreement on the best method to be selected. In VSC applications, different length normalization methods can yield different analyses of the same data set, and that there is no obvious way of selecting the best one. The choice of an inappropriate method, however, will have negative impacts on the accuracy and thus the reliability of clustering performance. The proposed solution is to analyze the data using different document length normalization methods and then to select the analysis on which the clustering structure agrees most closely with the bibliographical information of the news stories.

III. Literature Review

The literature suggests that recent years have witnessed the development of numerous text clustering methods ad algorithms. These include Explicit Semantic Analysis (ESA), Latent Semantic Analysis (LSA), Self-Organizing Maps (SOMs), Sensitive Text Clustering, Vector Space Clustering (VSC), and Word Sense Clustering. (VSC), however, remains among the popular and reliable methods in text clustering applications for its accuracy and effectiveness in different clustering applications. VSC is still widely used in different natural language processing (NLP) applications including data mining, information retrieval (IR), document organization and browsing, corpus summarization, and document classification [9-12]. It is used in different tasks and for different purposes including marketing, grouping similar documents (news, tweets, academic articles, etc.) and the analysis of customer/employee feedback, and discovering meaningful implicit subjects across all documents.

VSC is simply a technique where documents are compared with each other than indexed or classified in terms of their similarity or distance based on the words they contain. It can be defined as the organization of a collection of documents usually represented by a vector space model into distinct clusters based on similarity. The theory was first developed by Salton [13] essentially for IR purposes four decades ago and since then it has become a standard tool in IR systems. The underlying formula of VSC is initially to extract all useful information within a document collection and record it in an index known as a vector space. Then a proximity measurement is used to compute the semantic similarity among the documents with the purpose of grouping similar documents together.

In spite of its popularity and extensive use, VSC has many challenges that have negative impacts on the clustering performance and accuracy. In this regard, many studies have doubted the effectiveness of VSM as it is wholly based on lexical semantics with no regard to the importance of context in identifying intended meanings [14-17]. Likewise, some studies have argued that VSC is less effective in clustering and ranking web pages since these have some special features such as hyperlinks and structural information, which inevitably have additional information and these are ignored in VSC applications.

The main problems with VCS are thus associated with the issue of selecting *appropriate features* of documents that should be used for clustering. Different studies have referred to the limitations of VSC methods in terms of extracting the most distinctive features within datasets [18-20]. For a better feature selection performance, however, some issues need to be addressed. These include document length variation. This issue represents a challenge to the accuracy of clustering performance. The problem is that in the representation of data, the same term usually occurs repeatedly in long documents and that the vocabulary of a long document is usually large. This has the effect that long documents are clustered together and in the same way, short documents are clustered together without any regard to thematic criteria [21]. In other words, clustering is generated based on document length, not semantic similarity. The literature suggests that different techniques have been developed in order to address the issue of document length variation in text classification. These are referred to as document length normalization (DLN) techniques. DLN is a way of penalizing the term weights for a document in accordance with its length. DLN has been one of the central topics of interest in IR and document clustering theory and applications for many years [2, 22, 23]. These include cosine normalization, relative frequency, maximum term frequency, mean term frequency, probability normalization, byte length normalization, and likelihood of relevance. The basic principle of all these techniques is that text length is adjusted so that long texts are not favored simply because they have more terms. Here is a short review of some of the most commonly used length normalization techniques.

A. Mean Document Length Normalization

Mean document length normalization is one of the simplest and most straightforward normalization methods. It involves the transformation of the row vectors of the data matrix in relation to the average length of documents in the corpus using the function.

$$M_i = M_i(\frac{\mu}{length\ (C_i)})$$

Where

M*i* is the matrix row representing the frequency profile of any document collection C,

Length (C*i*) is the total number of letter bigrams in C*i*, and

μ is the mean number of bigrams across all documents in C:

$$\mu = \sum_{i=1}^{m} \frac{length\ (C_i)}{m}$$

The values of each row vector M*i* are multiplied by the ratio of the mean number of bigrams per document across the collection C to the number of bigrams in document c*i*. The longer the document, the numerically smaller the ratio is, and vice versa. This has the effect of decreasing the values in the vectors that represent long documents, and increasing them in vectors that represent short ones, relative to average document length [3, 24-26].

B. Cosine Normalization

Cosine normalization is the most commonly used technique in the vector space model. Cosine normalization was developed some decades ago with early information retrieval (IR) efforts; nevertheless, it remains one of the best normalization methods. The underlying principle of cosine normalization is that all documents in a given collection are represented equally. In this process, all row vectors of the matrix are transformed so as to have unit length and are made to lie on a hypersphere of radius 1 around the origin so that all vectors are equal in length [27-30]. Accordingly, variation in the lengths of documents and, correspondingly, of the vectors that represent them cannot be a factor [31]. One main problem however with cosine normalization is that it tends to be more biased towards shorter documents. This observation is quite obvious in IR applications where it tends to retrieve shorter documents more than longer documents [32].

C. Probability Normalization

This is a widely used method whereby the frequency values in each vector row are divided by the sum of frequencies in that row. This has the effect of replacing absolute frequency values, whose magnitudes are dependent on document size, with probabilities, which are not. In practice, probability normalization gives satisfactory results when dealing with reasonably small numbers of variables [33, 34].

Examination of the literature shows that there is no one strong conclusion recommending one approach. Besides, many of these have not been tested for the clustering of news in Arabic to find the best approach. Given the lack of agreement and disparity of opinions, we set out to comprehensively evaluate the existing normalization techniques to prove empirically which approach is the best for the normalization of text length to improve the text clustering performance of news in Arabic.

IV. Methodology

To address the research problem, this study is based on experimenting with different normalization techniques to propose a reliable normalization method for the text clustering of news in Arabic. In so doing, a corpus of 693 stories representing different categories and of different lengths is designed. Stories were derived from four different newspapers. These are *Al-Ahram* (Egypt), *Ash-Sharq Al-Awsat* (Saudi Arabia, located in London), *Al-Bayan* (United Arab Emirates), and *Al-Ghad* (Jordan). The selected stories represent different topics including politics, sports, family, environment, health, education, technology, and business as shown in Table I.

The size of the documents ranges from 01 KB to 480 KBs. This is shown in Table II.

This study adopts the vector space model (VSM) for the mathematical representation of data. The reason is that it is conceptually simple as well as it is convenient for computing semantic similarity within documents. The model is usually referred to as a ‘bag of words’ where a text is represented as a string of words disregarding context and/or word order. Each document is represented by the number of occurrences of each word in the document in Euclidean vector space where each token in the vector corresponds to a unique/given word in the matrix [35, 36]. In VSM, a document is mathematically represented by a vector of index words extracted from the document, with associated weights representing the lexical frequency of these words in the document and within the whole corpus collection. A data Matrix M*ij* was built in which rows Mi represent the documents and columns M*j* the lexical type variables, and the value at the M*ij* is frequency of lexical type *j* in document *i*. The data matrix M*ij* was built out of the lexical variables representing the 693 texts.

TABLE. I. News Categories

Topic	**Number of Documents**
Business	113
Education	87
Environment	78
Family	81
Health	84
Politics	82
Sports	109
Technology	59
Total	693

TABLE. II. THE LENGTHS OF THE DOCUMENTS IN THE CORPUS

Size	Number of documents
From 01 KB to 10 KBs	97
From 11 KBs to 50 KBs	108
From 51 KBs to 100 KBs	102
From 101 KBs to 200 KBs	86
From 201 KBs to 300 KBs	84
From3 01 KBs to 440 KBs	111
From 401 KBs to 500 KBs	105
Total	693

V. ANALYSIS

In this section, the selected data is analyzed using different document length normalization methods using 𝒦-means clustering, one of the simplest and most popular VSC methods. In this process, every data point (the news stories in our case) is assigned to the closest center or nearest mean based on their Euclidean distance. Then, new centers are calculated and the data points are updated. This process continues until there are no further iterations and changes within the clusters as seen in Figure 1.

Initially, the selected texts were clustered without the use of any normalization method. The matrix M693 was assigned into two main clusters, which can be called A and B as shown in Figure 2.

Examination of the two clusters, however, shows that the texts do not cluster coherently in terms of thematic criteria, and the clustering, in fact, makes no obvious sense in terms of anything one knows about them and their subject matters. The reason is that there is a progression from the longest texts at the top of the tree to the shortest at the bottom; when correlated with cluster structure, it is easily seen that they have been clustered by length, so that A contains the longest texts and B the shortest. The idea is that in vector space, the distance between any two vectors in a space is determined by the size of the angle between the lines joining them to the origin of the space's coordinate system, and by the lengths of those lines [3, 23, 37]. Using external criteria methods, the clustering structure generated herein was evaluated in terms of the prior knowledge and information obtained about the news stories. Clustering accuracy was estimated to be only 17%. This supports the hypothesis that the lack to address variation in document length in VSC applications has negative effects on the accuracy and reliability of clustering performance. Clustering performance is thus improved when a normalization method compensates for length in all documents so all lexical entries are equally represented. This will have the effect that documents will be clustered based on semantic similarity, not document length. The next step then is to try different normalization methods to choose the most appropriate normalization method for the text clustering of news in Arabic where documents can be clustered based on semantic similarity, not document length. Seven normalization methods are used. These are alphabetically ordered as follows: byte length normalization, cosine normalization, maximum tf normalization, mean normalization, pivoted-cosine normalization, probability normalization, and Z-score.

Using byte length normalization method, the row vectors of the data matrix M693 were normalized to compensate for the variation in length among the texts so that their lexical frequency profiles could be meaningfully clustered. Texts were assigned into five clusters as shown in Figure 3.

This process is repeated with cosine normalization, maximum tf normalization, mean normalization, and probability normalization methods. Accuracy rates are represented in Table 3.

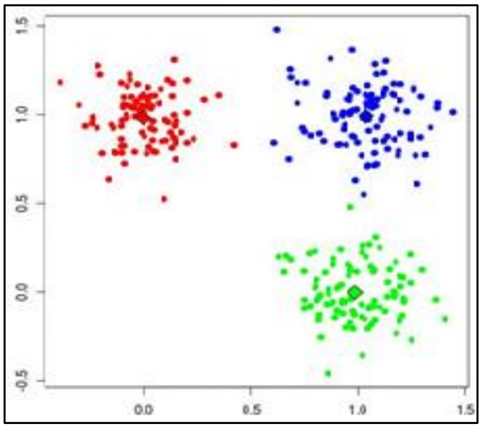


Fig. 1. Example of K-Means Clustering.

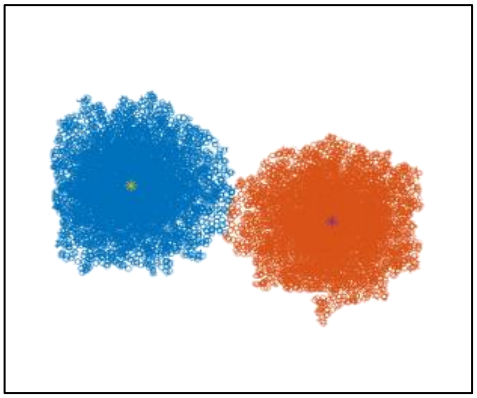


Fig. 2. K-Means Clustering of the Data Matrix M693 without the use of any Normalization Method.

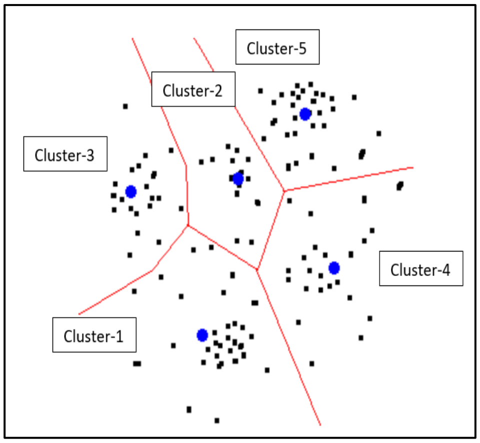


Fig. 3. K-Means Clustering of the Data Matrix M693 based on the Byte Length Normalization Method.

TABLE. III. ACCURACY RATES OF THE SELECTED DOCUMENT LENGTH NORMALIZATION METHODS

Normalization Method	Accuracy Rate
Byte length/size normalization	81%
Cosine normalization	73%
Maximum tf normalization	66%
Mean normalization	73%
Pivoted-cosine normalization	78%
Probability normalization	65%
Z-score	69%

The analysis indicates that byte normalization is the best method in terms of representing the terms within all the documents equally. One advantage of this method is addressing the issue of variation without distorting the byte size of documents. However, the analysis pointed to a major limitation with this method. It represents the documents equally without ranking of the lexical variables within the data set. For improving the document length normalization performance, thus, it is suggested that term frequency-inverse document frequency (TF-IDF) is used alongside the byte normalization method. The hypothesis is that TF-IDF will have the effect of ranking the words within all the documents so that only those with the highest TF-IDF values will be retained [20, 38-40].

Given that the highest TF-IDF variables are the most important, each column was calculated using the function:

$$tfid(t_j) = tf(t_j)log_2(\frac{M}{df_j})$$

Where *tf(t_j)* is the frequency of *term t_j* across all documents in the data matrix M693. Using the above formulation, the TF-IDF of some lexical type A that occurs once in a single document is 1 x $\log_2$ (1000 / 1) = 9.97, and the TF-IDF of a type B that occurs 400 times across 3 documents is 400 x $\log_2$ (1000 / 3) = 3352, that is, B is far more useful for document differentiation than A, which is more intuitively satisfying than the alternative. The variables are sorted in descending order as shown in Figure 4 and only the highest 1500 lexical variables within the data corpus were retained.

As a final step, a K-means clustering based on the byte normalization method and TF-IDF analysis was carried out. The documents were assigned to clearly define six groups (as seen in Figure 5) which correspond to a great extent to the information obtained about these documents with an accuracy rate of around 95.6%.

It can be thus claimed that the use of a single normalization method is not effective in terms of the document clustering of news in Arabic. The performance of normalization performance; however, can be improved with the use of TF-IDF alongside the byte normalization method.

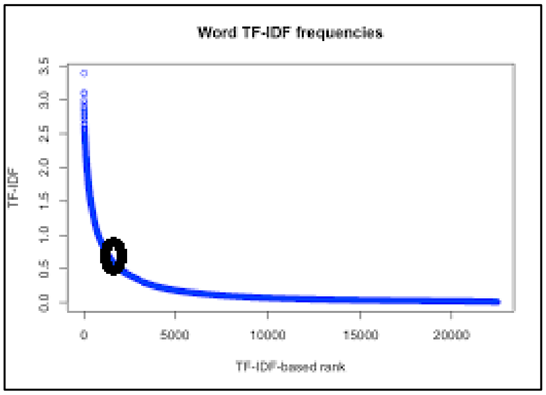


Fig. 4. Term Ranking using TF-IDF.

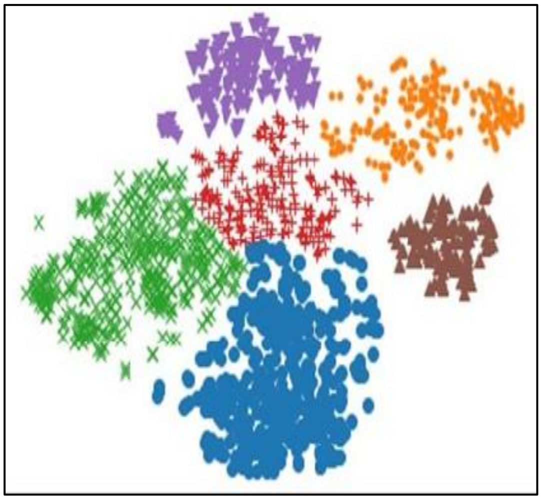


Fig. 5. K-Means Clustering of the Data Matrix M693 based on Byte Length Normalization Method and TF-IDF.

VI. CONCLUSION

This study addressed the issue of the effect of document length variation on the accuracy of the news clustering in Arabic. Different normalization methods were used and compared. It was found out that the byte length normalization method despite its limitations is the most appropriate for clustering applications of news in Arabic. In order to address these limitations, this study proposed the use of TF-IDF alongside this normalization method. The proposed model had the effect of improving the function of the byte normalization method and thus increasing the accuracy rate of the clustering performance. It can be finally concluded that the use of a single normalization method is not sufficient in addressing the issue of document length variation. The findings of the study can also be extended to IR applications in Arabic. The proposed model can be usefully used in supporting the performance of the retrieval systems of Arabic in terms of finding the most relevant documents for a given query based on semantic similarity, not document length.

ACKNOWLEDGMENTS

I take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Scientific Deanship, for all technical support it has unstintingly provided towards the fulfillment of the current research project.

REFERENCES

[1] Manning, P. Raghavan, and H. Schütze, An Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

[2] C. X. Zhai and S. Massung, Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. Association for Computing Machinery and Morgan & Claypool Publishers, 2016.

[3] H. Moisl, Cluster Analysis for Corpus Linguistics. De Gruyter, 2015.

[4] B. Mitra and N. Craswell, An Introduction to Neural Information Retrieval. Now Publishers, 2018.

[5] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," ACM Transactions on Asian Language Information Processing (TALIP), vol. 8, no. 4, 2009.

[6] I. Guellila, H. Saâdaneb, F. Azouaoua, B. Guenic, and D. Nouve, "Arabic natural language processing: An overview," Journal of King Saud University - Computer and Information Sciences, 2019.

[7] S. L. Maire-Sainte, N. Alalyani, S. Alotaibi, S. Ghouzali, and I. Abunadi, "Arabic Natural Language Processing and Machine Learning-Based Systems " IEEE Access, vol. 7, pp. 7011-7020, 2019.

[8] N. Y. Habash, Introduction to Arabic Natural Language Processing (Synthesis Lectures on Human Language Technologies). Morgan and Claypool Publishers, 2010.

[9] H. Lane, H. Hapke, and C. Howard, Natural Language Processing in Action: Understanding, analyzing, and generating text with Python. Manning Publications, 2019.

[10] H. Saggion and G. Hirst, Automatic Text Simplification. Morgan & Claypool Publishers, 2017.

[11] A. K. Luhach, D. Singh, P. A. Hsiung, K. B. G. Hawari, P. Lingras, and P. K. Singh, Advanced Informatics for Computing Research: Second International Conference, ICAICR 2018, Shimla, India, July 14–15, 2018, Revised Selected Papers (no. pt. 1). Springer Singapore, 2018.

[12] T.-U. Jang, W. Lim, Y.-M. Yang, and B. M. Kim, "Classification of the motor imagery EEG signal using vector quantization and K-nearest neighbors' algorithm," International Journal of Advanced and Applied Sciences, vol. 2, no. 12, pp. 72-77, 2015.

[13] G. Salton, The Smart retrieval system experiments in Automatic document processing. Englewood Cliffs: Prentice Hall Inc., 1971.

[14] S. Deerwester, T. D. Susan, W. F. George, K. L. Thomas, and H. Richard, "Indexing by latent semantic analysis," Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[15] T. K. Landauer, P. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," Discourse Processes, vol. 25, no. 2-3, pp. 259-84, 1998.

[16] E. Gabrilovich and S. Markovitch, "Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge," Proceedings of the Twenty-First National Conference on Artificial Intelligence, pp. 1301--1306, 2006.

[17] E. Gabrilovich and S. Markovitch, "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis," Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 1606--1611, 2007.

[18] R. Kaspar and B. Horst, Graph Classification And Clustering Based On Vector Space Embedding. World Scientific Publishing Company, 2010.

[19] A. E. Hassanien, C. Grosan, and M. F. Tolba, Applications of Intelligent Optimization in Biology and Medicine: Current Trends and Open Problems. Springer International Publishing, 2015.

[20] C. X. Zhai, Statistical Language Models for Information Retrieval. Morgan & Claypool, 2009.

[21] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," 19th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval pp. 21–29, 1996.

[22] L. M. Q. Abualigah, Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering. Springer International Publishing, 2018.

[23] T. Strzalkowski, Natural Language Information Retrieval. Springer Netherlands, 2013.

[24] W. Verhaegh, E. Aarts, and J. Korst, Algorithms in Ambient Intelligence. Springer Netherlands, 2013.

[25] A. W. Santoso et al., "A fuzzy approach for speckle noise reduction in SAR images," International Journal of Advanced and Applied Sciences, vol. 3, no. 5, pp. 33-38, 2016.

[26] R. Al-Jabar, "MFCC features with kernel PCA for speaker verification system " International Journal of Advanced and Applied Sciences, vol. 1, no. 6, pp. 37-44, 2014.

[27] A. Albalate and W. Minker, Semi-Supervised and Unsupervised Machine Learning: Novel Strategies. Wiley, 2013.

[28] U. S. Tiwary and T. Siddiqui, Natural Language Processing and Information Retrieval. OUP India, 2008.

[29] H. M. Blanken, A. P. de Vries, H. E. Blok, and L. Feng, Multimedia Retrieval. Springer Berlin Heidelberg, 2007.

[30] T. Sing, S. Siraj, R. Raguraman, P. Marimuthu, and K. Nithiyananthan, "Cosine similarity cluster analysis model based effective power systems fault identification," International Journal of Advanced and Applied Sciences, vol. 4, no. 1, pp. 123-130, 2017.

[31] H. Moisl, "Using Electronic Corpora in Historical Dialectology Research: The Problem of Document Length Variation," in Studies in English and European Historical Dialectology, vol. 98, M. Dossena and R. Lass, Eds., 2009, pp. 67-90.

[32] A. K. Singhal, Term Weighting Revisited. Cornell University, 1997.

[33] W. J. Stewart, Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling. Princeton University Press, 2009.

[34] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques. Elsevier Science, 2011.

[35] Y. Ozgur, "Empirical selection of nlp-driven document representations for text categorization," Syracuse University, 2006.

[36] T. Joachims, Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer Academic Publishers, 2002, p. 224.

[37] W. Wu, H. Xiong, and S. Shekhar, Clustering and Information Retrieval. Springer US, 2013.

[38] L. Azzopardi et al., Advances in Information Retrieval Theory: Second International Conference on the Theory of Information Retrieval, ICTIR 2009 Cambridge, UK, September 10-12, 2009 Proceedings. Springer, 2009.

[39] T. Roelleke, Information Retrieval Models: Foundations and Relationships. Morgan & Claypool Publishers, 2013.

[40] T. W. Miller, Web and Network Data Science: Modeling Techniques in Predictive Analytics. Pearson Education, 2014.

Impact of Information Technologies on HR Effectiveness: A Case of Azerbaijan

Aida Guliyeva[1], Ulviyya Rzayeva[2]
Department of Digital Economy and ICT
Azerbaijan State University of Economics
Baku, Azerbaijan

Aygun Abdulova[3]
International Magistrate and Doctorate Center
Azerbaijan State University of Economics
Baku, Azerbaijan

Abstract—**In the article, the impact of information technologies' (IT) implementation into the work of human resource departments for increased effectiveness is explored. Modern relations at the enterprise require the most important network-based enterprise's unit to be a strategic, flexible, cost-effective and service-oriented division of the organization. Although the influence of IT on Human Resources Management (HRM) has been a focus of scientists' attention, no empirical research has been conducted in this area in Azerbaijan. The authors use the experience and initiatives of enterprises and national banks to show the current state and results of the implementation of IT in HRM. Obtained data show that IT is not widely used in Azerbaijani organizations to perform HRM functions. The results also show that, although IT should have a certain impact on all sectors in terms of HRM, the types of IT used should vary significantly in recruitment, maintenance, and development tasks.**

Keywords—HR effectiveness; recruitment needs; maintenance and development tasks; management and planning tasks; performance of enterprises and banks; human capital; return on investment

I. Introduction

Modern radical transformations, intended to modify the form of ownership, the rights of organizations to economic independence, create objective economic conditions for the development of a personnel-training system at an enterprise. At the same time, in Azerbaijan, independent training of personnel by enterprises is not carried out for several reasons. The centralized distribution system for graduates that took place in Soviet times had ceased to exist, and with it, the interactions of enterprises and universities on staff training issues weakened.

At the same time, the rapid development of IT, as well as its implementation in all spheres of human life in recent years has led to a sharp expansion of information processing methods. It is impossible today to imagine an enterprise space without IT [1].

Enterprise-wide information systems are an organizational streamlined, interconnected set of techniques, and methods used for storage, processing, and dissemination of information for the achievement of desired objectives. Such an understanding of the information system requires computer literacy of employees in the process of working with special computer engineering tools in the field of information processing [2]. The creation of an information system assumes that the basic operations of accumulating, storing and processing information are assigned to computing equipment, while the professionals perform only a certain part of manual operations or accomplish procedures that require a creative approach in preparing management decisions [3]. At the same time, computing equipment works in close cooperation with specialists who control its actions, change the values of separate parameters in support of operational objectives, and enter input data for meeting challenges and management functions [4].

This study is devoted to the analysis of IT used in the personnel departments of organizations and several banks in Azerbaijan. The study aims to highlight several issues:

- What technologies in personnel management do enterprises and banks (national and international) in Azerbaijan use?
- Does the implementation of IT in HRM contribute to the overall performance of enterprises and banks, and how do they affect organizational indicators?

This paper focuses on the analysis of the impacts of IT in HRM using a case study of Azerbaijan. The study is prepared in four main parts. The paper first provides a review of the literature. In the second part, the methods used are described and explained. The third part is devoted to the analysis of empirical results. Finally, the fourth part provides a case study of Azerbaijan and scrutinizes the impacts of IT in HRM for increased efficiency in particular businesses. Section 5 provides discussions of the results. Section 6 concludes with some closing remarks and implications.

II. Literature Review

The automatic management of human resources (HR) in organizations gains objective strategic weight, and the importance of harmonized actions of HRM and elaborated business strategies is well known [5]. Many experts predict that appropriate software will become a central tool for all HR professionals [6]. It is expected that organizational efficiency grows with the increase in the use of specific methods of personnel management. Different companies may use different technological platforms, but usually, they all come together under the general title "Human Resource Information Systems (HRIS)". Such systems acquire, store, manipulate, analyze, extract, and distribute information in organizations on their human resources, staffing and organizational characteristics [7].

According to [8], HR technology has three main functional components: an input function (input of personnel information into the HR database), a processing function (adding new information and updating the database), and an output function (creating the resulting information, reports).

One of the potential positive effects of introducing IT in the enterprise is that it allows creating an IT-based workplace. This leads to the activation of basic managerial competencies, namely, responding to new challenges that require a restructuring of professional activities [9]. The chain of revolutionary changes in IT makes it possible to solve many of HRM problems, such as attracting, retaining and motivating employees, organizing the personnel functions under the company's strategy and managing the “human factor”. IT in the department of human resources can automate routine tasks such as payroll processing, compiling reports, conducting questionnaires and processing results, developing and comparing different motivational schemes, so that HR specialists can focus on more strategic issues related to identifying and implementing priorities of economic, scientific, technical and personnel policy [10]. Progressive companies already have a stake in HR tech and actively use applications and digital systems in their work. Author in [11] especially emphasizes the fact that technologies like shared databases, expert systems, telecommunication networks, decision-making tools, high-performance computers, etc. can radically change the methodological, informational and technological components of management processes and carry out them at a qualitatively new, more efficient level. However, in Azerbaijan, most HR departments continue to use the tools of the “old school” and are still not ready to integrate digital HR into their work [12]. Currently, there is an increasing power of New Information Technologies (NIT) on all management decisions. However, there are several objective factors that have a restraining effect on the pace of their implementation in Azerbaijan, which include, for example, the following: economic instability, gaps in the legislative framework, lack of formal education of managerial skills in IT, insufficient government funding for research and development related to NIT, clearly demonstrated lag relative to the West in the field of human resources development, computer equipment, and communications.

Not all HR professionals are ready to use modern technologies in their operations-and there are several explanations for this. First, people, by nature, tend to resist modification and do not want to change their usual way of working. The second reason is that HR technologies are developing so rapidly that specialists simply do not have time to keep track of all the software and mobile applications that appear on the market. The third reason is connected with the second factor - the level of user experience cannot cope with complex technological solutions [13].

However, observations show that in the existing literature the effect of the implementation of IT on personnel management in various organizations has not yet been systematically studied [14-16]. Although there is a large number of researches on information technologies and many articles are devoted to IT management, but they poorly cover the use of technologies in HRM [17, 18].

The implementation of IT can bring to many modifications and improvements in the work of enterprises, for example, reducing administrative costs, increasing productivity, reducing response time. According to [19], IT is actively involved in the process of preparing management decisions. The nature of the IT used varies depending on the specific information needs of an organization [20].

Nevertheless, the integration of developing Azerbaijan into the world information space cannot but contribute to the introduction of automation into all management processes, including HRM [21, 22]. Many problems, which no domestic enterprise is ready to fully resolve currently, accompanies the modern level of development of automation in the management sphere.

III. Materials and Methods

The objectives of the study are to determine the changes after the introduction of IT in the human resources department of an organization, the impact of ICT on the organization’s performance and the degree of participation of modern information technologies in the effective functioning of organizations in Azerbaijan. The authors also find out how various IT models support the activities of staff and increase their professional qualifications, study the impact of technology on the acquisition of professional knowledge in the workplace, the development of planning and organization skills using software products.

The work is based on a combination of methods of qualitative and quantitative analysis (multi-methodology). Research analytics is formed from data obtained as a result of observations, questionnaires, surveys, and the collection of necessary primary documentation. The questions of the questionnaire are focused on studying the current situation in enterprises in terms of the IT used in departments. A set of issues deals with identifying the intensity of software used. The questionnaire also includes questions to which answers can serve as confirmation or refutation of the hypotheses put forward. One-way analysis of variance (ANOVA) was used to process the results, implying that the averages of the total populations are equal. In other words, they all refer to the same population and the differences are random. For the purpose of testing the F-distribution is used, which accepts only positive or zero values.

Besides, the change in the productivity of personnel at enterprises and the efficiency of banks in the application of IT is described by calculating well-known indicators of the economic efficiency of the personnel service. Also indicators of the quality of personnel management, the cost of human capital and the effectiveness of banks are calculated. All the necessary data are taken from the official websites of the respective banks or during the interview with the responsible persons.

Conclusions are also presented by qualitative and quantitative results. One and a half hundred state, public and private organizations, including four banks, were taken as an object of study.

In meeting the challenges outlined in the work, the methods and techniques of economic and statistical analysis are applied.

On the one hand, the study involves conducting formalized interviews with representatives of human resources departments of organizations on a common questionnaire for all respondents. The empirical basis of the study is the combined answers to the questionnaire. On the other hand, financial and economic indicators, statistical and analytical materials of the studied banks, as well as data published in industry and periodicals and electronic media were also taken into account.

The scientific novelty of the results consists of solving the urgent task of developing the non-oil sector of Azerbaijan, in particular ICT, in terms of assessing the current situation and prospects for the implementation of information technologies in the activities of the organization's human resources department.

IV. Information Technologies in Reaching HR Effectiveness

Increasing the Efficiency in an Organization

The use of HR technologies may depend on various factors: the size and age of the enterprise, managerial diligence to personnel and innovations, the qualifications and experience of the personnel director, etc.

Depending on the complexity of the organizational structure of the enterprise, researchers allocate organizational-technological and other resources that are integrated by using computer networks.

To close the gap, the software used shall be analyzed, and in the paper for this purpose, two groups of hypotheses are presented:

1) The impact of information technologies on the activities of an organization's units;

2) The impact of types of information technologies on the activities of an organization's units

First, we provide the hypotheses of the first group.

Some companies in Azerbaijan use various information technologies for candidates' search, accepting applications, analyzing applications, using the reserve base. Nevertheless, of course, most organizations do not use IT at all, or the introduction of information technology is limited only to office applications. On the technological side of the above tasks, the authors highlight ethical methods (posting ads on the Internet, posting vacancy announcements on the company's website, attracting recruiting agencies, searching for candidates directly in educational institutions, notifying employment centers of open vacancies, participating in job fairs, etc.) and unethical ones (draining human resources away from a competitor company, illegal acquisition of a database, etc.). From the foregoing, the first hypothesis follows:

H1A. Depending on the specific requirements for hiring personnel in a particular enterprise, information technology should have different service models.

Modern research shows that the competence of personnel has a huge impact on the efficiency of the enterprise and the level of profit. Before selecting a candidate for a vacant position, it is necessary to describe his profile - a list of requirements for a nominee in this profession, specialty, and post. In compiling a complete set of personal data, methods of profession graph are used – based, for example, on fuzzy logic information technology of studying the requirements of the profession for personal qualities and psychophysiological characteristics, socio-psychological indicators, natural abilities, business qualities, professional knowledge and skills, and health. Therefore, the authors propose a second hypothesis:

H1B. Depending on the specifics of the maintenance and development tasks of a particular enterprise, information technologies should have different service models.

Organization staff is the most complex management object. Modern concepts of management and planning of the company's activities are based on the recognition of the growing importance of the employee's personality, the study of his motivations, the ability to correctly form them and adjust them in accordance with the strategic tasks facing the company. In modern companies focused on long-term success, information technologies for the selection of personnel based on business and personal qualities, as well as the official and professional promotion of employees based on the use of reasonable criteria for evaluating their activities, are developing. Thus, we put forward the third hypothesis:

H1C. Depending on the specific tasks of management and planning of a particular enterprise, information technologies should have different service models.

The hypotheses of the second group logically follow from these statements.

Modern information technologies concerning the tasks of personnel selection of different organizations can be used in:

- Practical activities of personnel services and personnel departments of organizations, recruitment agencies in the development of personnel selection technologies; when choosing special software; during certification and assessment of personnel of organizations; when searching, selecting and selecting candidates for the post.
- Practical activities of consulting agencies with preliminary, targeted and complete diagnostics of personnel; consulting on personnel's management problems; advising on staff development issues.
- Teaching in the preparation of textbooks and guidelines for university students and colleges; in the process of lecturing and conducting seminars and workshops.
- The work of employment services, training, and educational centers for the retraining of specialists and professional development as diagnostic and self-diagnosis methods of professional training; as a practical toolkit illustrating recruitment opportunities, etc.

Thus, we put forward our first hypothesis of the second group.

H2A. Depending on the employment requirements, at a particular enterprise-specific information technologies should be used.

Modern information technologies have a significant impact on the organizational structure and its main characteristics: configuration, complexity, level of formalization and centralization, coordination and control mechanisms, requiring managers to analyze the technologies used and planned for implementation to design an organizational structure for efficient service and enterprise development. Also the development of the organization creates standard problems, such as information asymmetry, information uncertainty, data multiplicity, data fuzziness, and data errors. Specific information technologies in these cases reduce or eliminate negative information factors.

The authors put forward the second hypothesis of the second group:

H2B. Depending on the tasks of service and development, at a particular enterprise-specific information technologies should be used.

Modern conditions of activity of most organizations require a goal-oriented, prompt staff selection when recruiting personnel. Not only specific job duties and requirements of the workplace should be taken into account, but also the tactical and strategic goals of the organization, ensuring the technologically competent implementation of the managerial and planned activities of the organization.

From the foregoing, the third hypothesis of the second group follows:

H2C. Depending on the managerial and planned tasks, at a particular enterprise specific information technologies should be used.

The above hypotheses are based on a thorough study of foreign literature in this field [23]. Further, the authors answer the question of how these hypotheses are applicable in a developing country, which is still lagging behind global trends in the use of ICT in various areas of economic life.

In the study, the data given below represents the result of a survey based on the questionnaire. A semi-formalized interview was conducted based on a questionnaire containing pre-prepared clear wording of questions and thought-out models of answers to them. Verbal responses were recorded in full, verbatim, with simultaneous primary coding on the attached scales. At the suggestion of the respondents, interviews were usually conducted on the territory of the enterprise itself. An obvious advantage was that the respondent, in this case, did not underestimate his role as the leader of the organization, was fully familiar with the problems raised during the interview.

The heads of 150 IT departments of state, public and private organizations (among them there are four large banks of Azerbaijan: international bank VTB, the state-owned International Bank of Azerbaijan (IBA) and the private banks Kapital Bank and Bank Respublika) acted as respondents. Since 2016, these companies have been participating in the Anniversary Azerbaijan International Telecommunications, Innovations and High Technologies Exhibition [24]. Filled out questionnaires were received from 89 companies, representing 59% of the total.

Through a thorough review of the available expert forms [25, 26], questionnaire points were formulated in Table I. This questionnaire contains seven questions that include seven variables to achieve the objectives of this study. The subsequent statistical adjustment procedure involves weighting the data, redefining the variables, and converting the scale. In the weighing procedure, a weighting coefficient reflecting the degree of its significance in comparison with other observations or respondents was assigned to each observation or respondent in the database. The procedure for redefining the variables was to transform the data to create new variables or modify existing ones. The purpose of this procedure was to create variables that best meet the main objectives of the study. The manipulation of the values of the scale (converting the scale) was carried out to be able to compare it with other scales, converting the data and making them suitable for analysis. The questionnaire is structured and does not contain open questions. Questions 3-6 list the answers from which the respondent must choose one or more alternatives. Except for the first question, all the rest are focused on HRM tasks.

TABLE I. QUESTIONNAIRE POINTS

Areas	Question	Range of values
Sector	What sector does your company belong to?	Government, public, private
IT use	Do IT operations depend on your HRM routine?	Yes/No
IT tools	What software tools are used to perform HR functions?	• Application software (DBMS, spreadsheets); • Data Mining tools (ERP, expert systems); • Information and communication technologies (LAN/WAN/ neural network, Internet/Intranet/Extranet, interactive services); • Others.
Recruitment	For which HR tasks does your company use IT?	• Monitoring of vacancies; • Search and selection of personnel using Internet resources; • Others.
Maintenance and development	For which maintenance and development tasks does your company use IT?	• Effective staff engagement; • High-quality and efficient selection; • Effective and targeted training, followed by measuring its effectiveness; • Retaining key specialists in the company; • Creating a comfortable microclimate; • Others.
Management and planning	For which management and planning tasks does your company use IT?	• Efficient use and development of the company's human resources and workforce planning; • Definition, creation, maximization, and support of benefits; • Identification and provision of a new generation of employees through mentoring and training; • Others
Restructure	Does your company require improved IT personnel management?	Yes/No

Source: Authors

In this research, indicators of recruitment, development, and maintenance, as well as management and planning are selected as independent variables, since they play an important role in shaping employee behavior [27, 28]. Besides, almost the entire system of personnel management software provides functions that correspond to these variables for the success of an organization. Existing studies do not represent a comparative analysis of software used for functions of various departments, including HR [29]. Nevertheless, these studies do not consider the relationships between types of software and the internal operations of an enterprise.

In this paper, the data were studied using ANOVA, a method in mathematical statistics aimed at finding dependencies in experimental data by examining the significance of differences in average values [30]. This method makes it possible to compare the mean values of three or more groups.

The results of the hypotheses of the first group (the influence of information technologies) are given below (Table II).

An analysis of the p-values in Table II showed that there was insufficient evidence to accept the hypotheses H1A or H1B. As in the case of H1A and H1B, if p>0.05, then the arguments presented are not enough to reject the null hypothesis. Without rejecting the null hypothesis, it can be stated that the results in this case are not significant. This means that IT does not significantly affect the hiring of employees, their support and development. The p-value for H1C is 0.02 and this hypothesis shall be accepted, that is, IT has a significant impact on management and planning tasks.

The results of the hypotheses of the second group (the influence of types of information technologies) are given below (Table III).

The last column of the table shows that, except for the H2C hypothesis, the results were significant at the level of 4-8 percent in this category. This means that the type of IT tool used for hiring, maintenance and development functions varies depending on the type of enterprise itself. The critical value of statistics in the first and third cases is 2.4 for a given level of significance. In both cases, regression is considered significant at a given level of significance. In the first and third cases, the p-value is 0.08, and in the second 0.98. Thus, in the second case, the confidence in the significance of the regression is significantly lower (the probability of error is significantly greater if the model is recognized as significant). On the other hand, the p-value for H2C indicates that there is not enough evidence to accept the hypothesis that the type of IT tool used affects management and planning functions.

TABLE II. RESULTS OF THE HYPOTHESES RELATED TO THE INFLUENCE OF INFORMATION TECHNOLOGIES

Test variables	Hypothesis	F-value	d.f.	p-value
Recruitment	H1A	1.9	2/9	0.2
Maintenance and development	H1B	0.78	2/9	0.48
Managing and planning	H1C	6.06	2/9	0.02

Source: Authors' calculations

TABLE III. RESULTS OF THE HYPOTHESES RELATED TO THE INFLUENCE OF TYPES OF INFORMATION TECHNOLOGIES

Test variables	Hypothesis	F-value	d.f.	p-value
Recruitment	H2B	2.4	3/32	0.08
Maintenance and development	H2C	0.06	3/32	0.98
Managing and planning	H2B	2.4	3/32	0.08

Source: Authors' calculations

A. Impact of IT in HRM on the Example of International and Local Banks

Further in the article, the influence of automation of information processes in the human resources department on increasing human capital in relation to banking is considered. Obviously, in modern conditions of fierce competition among banking organizations, modern information technologies are of great importance, which, albeit indirectly, is determined by the level of human capital of bank employees. By human capital, we will mean all non-monetary and intangible resources that are fully or partially controlled by the organization and are involved in creating value.

As authors have already noted, information technologies expand management capabilities when working with personnel; the banking sector is not an exception.

It is also expected that IT in HRM affect the efficiency of personnel management, which positively correlates with the organizational results of banks. To test this hypothesis, it is necessary to compare the results of the organizational activities of selected Azerbaijani banks. There are many different studies on the impact of IT on HRM on bank performance [31]. Our research is focused on HR analytics of selected banks and the comparison of results.

The main method of data collection for this section of the study is semi-structured interviews with HR directors of selected national and international banks. These interviews made it possible to collect information about personnel technologies in banks and the level of IT development in HRM. Return on investment in the staff of each bank is calculated to count the effectiveness of personnel [32].

Next, an overview of HR technologies in selected banks is presented below.

The authors interviewed the responsible persons of the international bank VTB, the state-owned International Bank of Azerbaijan (IBA) and the private banks Kapital Bank and Bank Respublika.

According to interviews with the heads of personnel departments, in all HR departments of the above-named banks, there is a payroll division, a division for development, promotion, and release of employees, a division for relations with trade unions and public organizations. Kapital Bank also has a personnel analysis division. In the Bank Respublika, there is an unspoken rule under which whatever the reason for dismissal (staff reduction due to production automation, desire for professional growth, or just wanting to change of scenery), an employee who left his place is not accepted back. According to the results of the study, the authors concluded that only the

Human Resources Management Department of Kapital Bank is involved in marketing and in building financial budget and strategy as a whole. This department supports line manager staff and employees and interacts with managers as business consultants. In the remaining banks, HR representatives are responsible for operational processes in the field of personnel management and do not act as strategic business partners. In all these banks, candidates can apply online to the bank's website. The Bank Respublika is an exception, where 10% of applications come at the email address of the HR department.

The functionality of modern personnel services is constantly changing, along with their names, reflecting the specificity and level of penetration into the internal affairs of an enterprise. The interviewees of all organizations confirm the impact of HRM on all processes of the enterprise. The ramified structure of VTB's HR department is responsible for communication with managers, regulates the human resources of the bank, for example, graduate recruitment, hiring experienced employees, relations with employees, their training and development, etc. HR managers communicate directly with managers of other departments, provide their personnel requirements and consult based on discussions with expert centers.

Many HR functions are automated by software solutions. As the interlocutors noted, in most cases each HR area is supported by a specific technological platform. In Kapital Bank, the Learning and Educational System administer the training and development; Graduate Recruitment System supports the recruitment, the remuneration of employees is also controlled by the automated system. In addition to the mentioned information technologies systems, the International Bank of Azerbaijan has a database that covers all records of employees from their recruitment upon retirement. These data cover personal information, qualifications, employee performance figures, vacation reports, salary information, etc.

Still, the main areas covered by the personnel departments of the banks under survey are personnel documentation and administration tasks under developing Azerbaijani legislation, and hiring and benefits management. It also shows poorly developed information systems in the field of personnel management in Azerbaijani banks.

Further, based on the formulas, the relationship between HR efficiency and performance of banks upon application of IT will be described; indicators of personnel management excellence, cost of human capital and efficiency of banks will be calculated. All necessary data were collected on the websites of the respective banks or in the process of interviews with responsible persons.

Now a few words about the general indicators of the economic efficiency of personnel service [33]. The HC ROI (Human Capital Return on Investment) indicator in various HR spheres demonstrates the effectiveness of investments, calculates the results of return on investments in personnel. This indicator is calculated by the formula:

HC ROI=(Revenue–(Expenses–Compensations))/Compensations

This method of evaluation is quite time-consuming. In the calculations, it is necessary to take into account not only the cost of a specific event, but also indirect costs associated with it, but the most difficult is a calculation of income from the event held in the field of personnel management. As indicated by [34], HC ROI is equal the value-added of investments in an organization's human assets. The numerator in this metric is the profit adjusted for the cost of people. The higher the HC ROI, the more effective is the personnel management in the bank [35].

The main indicator of employee productivity is human capital income (HCRF – Human Capital Revenue Factor) - the ratio of the total income by an indicator of employees working full-time (FTE – Full-time Equivalent):

HCRF = Income / FTE

It should be noted that this indicator is rapidly becoming obsolete.

The cost of human capital (HCCF – Human Capital Cost Factor) shows the proportion of staff costs in circulation and is calculated by the formula:

HCCF = Total Staff Costs / Turnover

The profitability of the "average" employee of the organization shows the indicator of HCVA (Human Capital Value Added):

HCVA = Income - (Costs - Salaries and Bonuses) / FTE

The results of the calculations, as well as a summary of interviews with the personnel directors of four banks, are presented in Table IV.

The data for calculations are collected on the official website of respective banks.

To assess the effectiveness of investment in human capital, many criteria and indicators can be used. We have focused on those that are used by organizations to evaluate their investments in personnel or are proposed by researchers for these purposes. The data presented in the table demonstrate that the international bank VTB has a higher return on human capital from investments. The HC ROI indicator shows how much each monetary unit invested in a specific activity brings. HC ROI, equal to 14.6, means that one monetary unit invested in the human capital of the bank returns 14.6 monetary units. Indicators of Azerbaijani banks demonstrate somewhat lower efficiency. Profit per employee of the bank is a mechanism for overall measuring the cost-effectiveness of all bank personnel. The ratio of income and expenses of employees shows that an international bank spends more money on its employees as compared to national banks. Since the level of development of HR technologies at the international bank is higher, and the efficiency of personnel management, as well as performance indicators, is better, a correlation between IT HR and personnel management efficiency can be observed.

TABLE IV. RESULTS OF THE CALCULATIONS RELATED TO EMPLOYEE PRODUCTIVITY*

	Loc.priv. Kapital Bank	**Local state IBA**	**Local private BR**	**Int Bank VTB**
Num.of Empl.	2700	1800	1050	300
HR Tech.	SAP	Spec. HR tech.app.developed by other bank headquarter	Spec. HR tech.app.developed by other bank headquarter	Spec. HR tech.app.developed by other bank headquarter
Num. of HR Prof.	20	20	11	9
HC ROI	5.824151	1.107982	3.1184105	14.68444
Empl. Exp. /Headcount	48641.071	96123.33	1076.1904	9894.033
Rev./ Headcount	188584.52	112402.2	22780	41493.66
HCRF	56452.380	447780	12739.047	24416.66
HCCF	0.153824	1.341210	0.0315648	0.005221
HCVA	168952.38	167033.8	22422.857	31816.3

Source: Authors' calculations

Thus, an increase in the bank's corporate human capital takes place when investing in information technologies, which put forward fundamentally new requirements for the employee, and the individual's ability to generate new knowledge and apply it to increase the efficiency of the bank.

V. Suggestions and Discussion

Continuous ICT education contributes to the formation of a professional basis for the human resources potential of the modern economy. Given the continuous changes in the technological environment, staffing for the development of the state economy should be based on continuing education in the field of information technology. The essence of the new technological revolution coming in the global business and banking system is to increase the yield of business due to its knowledge, qualification of personnel and technology. The banking system is being improved and every year becomes more complex and multifaceted. The results obtained show that the introduction and use of relevant information technologies can not only simplify the work of all branches of the bank but also make it more coherent.

The article analyzes the role of information technology in the formation of competitive advantage of an enterprise. This role is considered for optimizing information business processes when implementing IT in the organizational structure of the personnel department. Analysis of the impact of information technology on key organizational variables is based on questionnaires and interviews. Their role in diagnosing the state of the enterprise is also substantiated.

The article also analyzes the use of information technologies in banking management. The study shows the importance of information models and technologies in management reveals the features of information management in the banking sector.

The theoretical approaches to the definition of human capital differ in the interpretation of modern economists, but for banks, the very essence of the use and development of intellectual resources is a priority. New forms of organizing the activities of banks are based on modern approaches to determining the essence of capital, not only financial but also human since new business projects require a balanced approach to the abilities of personnel and modern developments in information technology.

However, an objective approach compels us to point out some contradictory trends, the emergence of stagnation and crisis processes, and the growth of problems in modern recruitment practices in Azerbaijan. Some existing approaches, techniques, procedures, and technologies for staff recruitment, including using information technology, adversely affect the effectiveness of organizations. The main problems of staff recruitment using traditional methods and techniques (for example, a document contest, and interview) are the disproportionality of the requirements and subjectivity in the assessments of applicants for vacancies and significant resource costs (especially for mass recruitment). The most acute and difficult to solve problems of staff recruitment using information technology are caused by other reasons.

Firstly, the provision of information and technology services by specialized firms does not always correspond to the modern requirements of personnel selection entities, and personnel management specialists are not typically focused on modern information and technological innovations.

Secondly, both personnel management and information technology are among the rapidly developing areas of professional activity. Therefore, specialists, loaded with their main professional activity, do not have enough motivation to track innovations in related fields. These results to a kind of stagnation manifested both in the use of outdated and inefficient information technologies, as well as outdated or unprofessionally developed personnel selection tools. These trends ultimately lead to a decrease in the quality of selection.

Thirdly, the capabilities of modern information resources are not fully used in organizations due to the lack of technologies for their use in the selection process.

Fourth, there is a rather complicated problem of creating the necessary and sufficient conditions for the effective use of information technology in the selection process.

VI. Conclusion

A breakthrough in the development of information technologies has greatly simplified the work of commercial institutions, making the internal structure and relationship system more convenient for employees; enterprises themselves are more accessible and more comfortable for customers. In the commercial sphere, information technologies are used in various forms, as in any organization. The most necessary of them are technologies for internal interaction of personnel and management. In modern conditions, the pace of technology development necessitates the constant updating of professional knowledge, skills, and competencies, as well as improving ICT skills of personnel.

The article substantiates the relevance of the use of information technology in personnel management systems as a key factor in ensuring sustainable growth and competitiveness of enterprises. A comparative analysis of approaches to understanding the economic essence and content of information technology in personnel management based on the hypotheses put forward is carried out. The circumstances that directly or indirectly determine the increasing demand of organizations on information technologies for personnel management are identified.

The article has examined the impact of information technologies personnel management as a set of software and IT, analyzed how the use of software products in personnel management can improve the efficiency of enterprises.

The introduction of modern information technologies, even with a wide supply in this market and regular price reductions, remains an expensive project. Nevertheless, the ultimate goal - to strengthen the market position of the enterprise - justifies the funds in the framework of a thoughtful and economically reasoned development strategy. Today, a versatile and multi-skilled manager will not engage in the project of implementing an information system without calculating the direct benefits of its operation, which is impossible without a thorough analysis and determination of its economic necessity, effectiveness, and expediency.

This article analyzes the prospects for implementing IT in a key area of organization management - the human resources department to achieve a competitive advantage. The authors analyzed the current situation in many Azerbaijani organizations and found that modern enterprises and industries are taking appropriate steps to introduce IT in the field of personnel management. However, there is a big gap in approaches to solving this problem between foreign and national organizations, which was demonstrated based on quantitative indicators of national and one international bank in Azerbaijan.

The authors have built a model for using IT tools to perform various functions of personnel management in enterprises and the banking sector. Based on the survey data, the results, firstly, showed that IT has a significant impact on all sectors in terms of management and planning tasks, and, secondly, the type of IT used varies considerably for recruitment tasks, as well as by functions of staff support and development. However, there is no standardization in integrating computer software into the core activities of HRM; there are no information systems in Azerbaijan that alone could cover the needs of a modern enterprise. Medium and large organizations usually operate at least a dozen multi-user systems. It can be explained by the gap between job requirements and the ability of employees to perform personnel management tasks. There are still problems with personnel in terms of elementary computer illiteracy. The survey shows that not all enterprises have special HR software. Most likely, it is expected that this situation will continue soon.

In future empirical research, the possibilities of introducing NIT into personnel management processes should be explored to improve HRM in the direction of optimizing personnel costs and to strengthen the efficiency of enterprise management as a whole through the rational use of its intellectual potential. Even though in Azerbaijan there is an acute need for the use of modern personnel management systems, insufficient attention is paid to the issues of staff case administrating by IT means on the part of supervisors.

References

[1] D. Soumyaja, T. Kamalanabhan, and S. Bhattacharyya, “Antecedents of employee readiness for change in the IT sector and the manufacturing sector: a comparative study”, International Journal of Human Resources Development and Management, Vol.18, 3/4, pp. 237-256, 2018. doi: 10.12816/0018079.

[2] P. Melo and C. Machado, Management and technological challenges in the digital age. Boca Raton, FL: CRC Press, 2018.

[3] E. Gimžauskienė and V. Varaniūtė, “Impact of information and communication technology for functionality of performance measurement system”, Economics and Management, vol.17 (1), pp. 15-21, 2012. doi: 10.5755/j01.em.17.1.2246.

[4] A. Tursunbayeva, “Human resource technology disruptions and their implications for human resources management in healthcare organizations”, BMC Health Services Research, vol.19(1), pp. 268-276, 2019. doi: 10.1186/s12913-019-4068-3.

[5] B. Phillips, “Information Technology Management Practice”, Journal of Organizational and End User Computing, vol.25(4), pp. 50-74, 2013. doi: 10.4018/joeuc.2013100103.

[6] D. Drum, A. Pernsteiner, and A. Revak, “Workarounds in an SAP environment: impacts on accounting information quality”, Journal of Accounting & Organizational Change, vol.13(1), pp. 44-64, 2017. doi: 10.1108/JAOC-05-2015-0040.

[7] S. Nawab, T. Nazir, M. Zahid, and S. Fawad, “Knowledge Management, Innovation and Organizational Performance”, International Journal of Knowledge Engineering-IACSIT, vol.1(1), pp. 43-48, 2015. doi: 10.7763/IJKE.2015.V1.7.

[8] M. Relich, “The impact of ICT on labor productivity in the EU, Information Technology for Development”, vol.23(4), pp. 706-722, 2017. doi: 10.1080/02681102.2017.1336071.

[9] P. Appiahene, N. Ussiph, and Y. Missah, “Information Technology Impact on Productivity”, International Journal of Information Communication Technologies and Human Development, vol.10(3), pp. 39-61, 2018. doi: 10.4018/IJICTHD.2018070104.

[10] R. Bahrini and A. Qaffas, “Impact of Information and Communication Technology on Economic Growth: Evidence from Developing Countries”, Economies, vol.7(1), pp. 21-34, 2019. doi: 10.3390/ economies7010021.

[11] L. Agu, “Information management in organizations: an overview. Information Impact”, Journal of Information and Knowledge Management, vol.8(4), pp. 123-136, 2018. doi:10.4314/iijikm.v8i4.10.

[12] R. Gasimov and N. Gurbanov, “Human Resource Management in Azerbaijan Companies: Evaluating on Functional Level, Economics and Management”, vol.18(1), pp. 165-176, 2013. doi:10.5755/j01.em.18. 1.4145.

[13] P. McEvoy, A.F. Ragab Mohamed, and A. Amr, “Review on the KM Applications in Public Organizations”, The Electronic Journal of Knowledge Management, vol.15(1), pp. 37-49, 2017. Available from: https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1016&context=buschgraart.

[14] H. Niavand, F. Nia, and R. Mahesh, “The Impact of Free Information Technology (IT) on Financial Markets”, Business Management and Strategy, vol.9(1), pp. 105-113, 2018. Available from: https://ideas.repec.org/a/mth/bmsmti/v9y2018i1p105-113.html.

[15] J. Dumay andd J. Guthrie, P. Puntillo, “IC and public sector: a structured literature review”, Journal of Intellectual Capital, vol.16(2), pp. 267–284, 2015. doi:10.1108/JIC-02-2015-0014.

[16] R. Tubey, K.J. Rotich, and A. Kurgat, “History, Evolution and Development of Human Resource Management: A Contemporary Perspective”, European Journal of Business and Management, vol.7(9), pp. 311-334, 2015. doi:10.4324/9781315299556-16.

[17] M. Shamekhi, H. Scheepers, and A. Ahmed, “The impact of Business Analytics on organisations: An Information Processing Theory perspective”, Australasian Conference on Information Systems, 2018. doi:10.5130/acis2018.ah.

[18] C. Seemiller, Complementary Learning Objectives: The Common Competencies of Leadership and Service-Learning, New Directions for Student Leadership, 150, 2016, pp. 23-35. doi:10.1002/yd.20168.

[19] H. Shahin and B. Topal, “Impact of information technology on business performance: Integrated structural equation modelling and articial neural network approach”, Scientia Iranica B, vol.25(3), pp. 1272-1280, 2018. doi: 10.24200/sci.2018.20526.

[20] R.Gould and M. Çetinkaya-Rundel, “Teaching Statistical Thinking in the Data Deluge”, Using Tools for Learning, Mathematics and Statistics, pp. 377-391, 2013. doi:10.1007/978-3-658-03104-6_27.

[21] R. Rehimli, The Management System of Republic of Azerbaijan. Ankara, Turkey; Kültür Ajans Tanıtım ve Organizasyon Ltd. Şti., 2011. ISBN: 978-605-4432-39-4.

[22] K. Kasemsap, “The Role of Information System within Enterprise Architecture and their Impact on Business Performance”, Advances in Business Information Systems and Analytics Technology, Innovation, and Enterprise Transformation, C, pp. 262–284, 2015. doi: 10.4018/978-1-4666-6473-9.ch012.

[23] M.I. Jambakr, A.F. Yuzaherdi, J. Jauhari, and R. Efendi, “Structural Equation Modeling Technique for Social Application of Information Technology User Requirements’ Identification”, J. Phys.: Conf. Ser. 1338 012049, 2019. doi:10.1088/1742-6596/1338/1/012049.

[24] https://bakutel.az/en-content/11-2016.html.

[25] K. Raworth, C. Sweetman, S. Narayan, J. Rowlands, and A. Hopkins, Conducting semi-structured Interviews, Oxford, UK: Oxfam; 2012.

[26] G. Kai and Z. Changzheng, “Effect of employees’ turnover rate on industrial firms’ production efficiency: A conceptual model”, The 2nd International Conference on Information Science and Engineering, 2010. doi:10.1109/icise.2010.5691528.

[27] M.F. Naim and U. Lenka, “Investigating the Impact of Social Media on Gen Y Employees Engagement”, International Journal of Human Capital and Information Technology Professionals, vol.8(3), pp. 29–48, 2017. doi: 10.4018/IJHCITP.2017070103.

[28] M. Tarafdar, Q. Tu, B. Ragu-Nathan, and T. Ragu-Nathan, The impact of Technostress on role stress and productivity, Journal of Management Information Systems, vol.24(1), pp. 301–328, 2007. doi:10.2753/ MIS0742-1222240109.

[29] W.D. Wet, E. Koekemoer, and J.A. Nel, “Exploring the impact of information and communication technology on employees’ work and personal lives”, SA Journal of Industrial Psychology, vol. 42(1), 2016. Available from: https://sajip.co.za/index.php/sajip/article/view/1330/ 1924.

[30] T. K. Kim, “Understanding one-way ANOVA using conceptual figures”, Korean J Anesthesiol, vol.70(1), pp. 22–26, 2017. doi: 10.4097/kjae. 2017.70.1.22.

[31] S. Gupta and A. Yadav, “The Impact of Electronic Banking and Information Technology on the Employees of Banking Sector”, Management and Labour Studies, vol.424, pp. 379–87, 2017. doi:10.1177/2393957517736457.

[32] M. Kesti and A. Syväjärvi, “Human Capital Production Function in Strategic Management”, Technology and Investment, vol.6(1), pp. 2015. doi:10.4236/ti.2015.61002.

[33] S-A. Barnes, “The differential impact of ICT on employees: narratives from a hi-tech organization”, New Technology, Work and Employment, vol.27(2), pp. 120-132, 2012. doi:10.1111/j.1468-005X.2012.00283.x.

[34] Botchkarev and P. Andru, “A Return on Investment as a Metric for Evaluating Information Systems: Taxonomy and Application”, Interdisciplinary Journal of Information, Knowledge, and Management, vol.6, pp. 245-269, 2011. Available from: http://www.ijikm.org/ Volume6/IJIKMv6p245-269Botchkarev566.pdf.

[35] R. Ryan and T. Raducha-Grace, Business of IT: How to improve service and lower costs, IBM Press, 2009. ISBN-10: 0-13-701890-8.

Discovering the Relationship between Heat-Stress Gene Expression and Gene SNPs Features using Rough Set Theory

Heba Zaki[1], Ahmed Farouk Al-Sadek[4]
Agricultural Research Center (ARC)
Giza, Egypt

Mohammad Nassef[2], Amr Ahmed Badr[3]
Department of Computer Science, Faculty of Computers and
Artificial Intelligence, Cairo University, Egypt

***Abstract*—Over the years of applying machine learning in bioinformatics, we have learned that scientists, working in many areas of life sciences, call for deeper knowledge of the modeled phenomenon than just the information used to classify the objects with a certain quality. As dynamic molecules of gene activities, transcriptome profiling by RNA sequencing (RNA-seq) is becoming increasingly popular, which not only measures gene expression but also structural variations such as mutations and fusion transcripts. Moreover, Single nucleotide polymorphisms (SNPs) are of great potential in genetics, breeding, ecological and evolutionary studies. Rough sets could be successfully employed to tackle various problems such as gene expression clustering and classification. This study provides general guidelines for accurate SNP discovery from RNA-seq data. Those SNPs annotations are used to find relation between their biological features and the differential expression of the genes to which those SNPs belong. Rough sets are utilized to define this kind of relationship into a finite set of rules. Set of (32) generated rules proved good results with strength, certainty and coverage evaluation terms. This strategy is applied to the analysis of SNPs in A. thaliana plant under heat-stress.**

Keywords—RNA sequencing (RNA-seq); variant calling; Single Nucleotide Polymorphisms (SNPs) analysis; rough sets; gene expression

I. Introduction

RNA sequencing (RNA-seq) technology has resulted in exceptionally fast and wide scale analysis of the genetic information exists in all organisms. This mainly includes the concurrent study of alternative splicing, Single nucleotide polymorphisms (SNPs) and differential expression. The approach of genome-guided transcriptome has been the standard RNA-seq analysis method for model organisms like A. thaliana. Some existing software packages are available to perform this task [1]. New tools are continuously developed to be used for RNA-seq analysis task starting from reads alignment ending with the pathway analysis mission. Unfortunately, some non-expert users for those tools cannot get the full power and capabilities of them on wide scale [2].

SNPs are single nucleotide base variations, caused by transitions or transversions, in the same position between individual genomic sequences. Genetics and breeding are the most two important studies using SNPs as significant molecular markers. In genetic studies, SNPs are ideal genomic resources used for functional gene identification for traits and characterization of genetic resources because of their extensive genome distribution, wide density and, high scalability [3]. Fortunately, SNPs discovery can be accomplished on both approaches of genome-guided and de novo on variety of organisms [4]. This is applied on many plants, including those with little or no available genetic information.

Among the various benefits of performing SNPs analysis using RNA-seq data, there are two important ones [5]. First the reasonable cost for simultaneous discovery of thousands SNPs together with expression levels of functional genes at the same time. Second is involving phenotypes which can be predicted according to genotypes and, the location of SNPs in coding regions related to the possibly identified plant biological and agronomical traits.

RNA-seq is considered the ideal method for gene expression profiling [6] and, it is commonly used for precision medicine due to its high capability of measuring dynamic gene activity in the genome for a specific tissue type. Moreover, when applying RNA-seq on some disease tissue samples [7], it detects most of mutations exist in expressed genes that are related to disease biology.

Machine learning techniques can support very interesting and critical analysis applications dedicated for the fields of molecular biology and bioinformatics. Particularly, rough set method is considered very commonly used for this task of data analysis due to its flexibility in handling qualitative data. Rough Set theory was proposed in 1982 by Z. Pawlak [8] and, has been used as a methodology of database mining or knowledge discovery. It can contribute in many processes like attribute selection, attribute extraction, data reduction, decision rule generation and pattern extraction.

Rough Set uses information system or information table to represent data. This table consists of objects (rows) and attributes (columns) [9]. There are two types of attributes named as the condition attribute and decision attribute. Each row of an information table defines a decision rule, which specifies the decision attribute values when conditions are indicated by condition attributes are satisfied. Additionally, a set of objects is classified using rough set theory by finding dependencies and relations between attributes [10]; reduction of unnecessary attributes; discovering the most important attributes; or by decision rule generation.

Rough set-based rule generation provides easier explanations and descriptions for complex biological systems [8]. The challenges of those complex systems can be summarized into determining the features or attributes that can demonstrate the biological phenomenon, and what combinations of features' values can define that phenomenon and make a significant added value to the system study [11]. The possible decision values can be the participation of some particular genes in a biological process. Furthermore, learning the set of minimum features that can determine the gene involvement in this process may be interesting issue for some biologists. High throughput problems have a great care about discovering which features, in which order and, in which combinations define decisions.

The main motivation for this research is to provide aid to find a reasonable answer about the relationship between expressed genes and SNPs under the effect of heat-stress phenomenon. This relationship is achieved through determining the set of features that best describe this biological process. Rough-set based rule induction method is applied on RNA-seq data for the A. thaliana plant.

The rest of the paper is organized as follows. Related work is reviewed in Section II. Section III describes the data used in the research and tools adopted to perform SNPs detection and analysis. Methodology and techniques utilized for SNPs Detection phases as well as SNPs analysis are discussed in Section IV. The results of the experiments in the form of evaluation terms, tables and charts are discussed in Section V. Section VI provides our derivations, outcome on the study and, recommendations for the future work.

II. Related Work

Various research efforts in the literature have been targeted to the two main focal topics of this research; SNPs identification and Rough set theory in bioinformatics. This section lists a summarization of these efforts as follows.

The authors in [7] investigated the most suitable method that can provide the greatest number of SNP calls with high specificity and sensitivity. Following the steps of alignment sequence reads to the genome, removing duplicates, and using SAMtools to call SNPs had achieved the required purpose. SAMtools proved higher consistency than GATK with 8–10% more variants identification.

Plant functions, related to climate adaptation, have leading genes involved in transcriptional mechanisms. In the study [12], they realized that neat and strong peaks of association were identified in expected functional variants in the extreme tail of genetic differentiation. Those results proved that climate adaptation can mainly cause the genomic variation when applied on A. thaliana at a small scale.

SNP-ML (SNP machine learning) suggested in [13], a novel tool, predicted true SNPs from sequence data using machine learning. It was designed for calling more trusted SNPs from polyploids. Moreover, it provided SNP machine learner (SNP-MLer), a functionality to train new models for customized use. Tetraploid peanut SNPs were identified using SNP-ML, and the validated true- and false-positive SNP mapping data improved the discovery process.

Another research [2] suggested Visualization Pipeline for RNA-seq analysis (VIPER) that combined stages of an RNA-seq analysis workflow. This workflow graded from raw RNA-seq data, then quality control and genome alignment, reaching to the differential expression and pathway analysis. VIPER listed the most popular tools used in the workflow like, RSEM for quantification, and SNPeff for annotating identified SNPs.

A reasonable amount of work has been performed on the usage of rough set methodology in solving bioinformatics issues and challenges [14]. These studies have focused on problems of classification and reduction of bioinformatics data. Some other literatures have dealt with topics related to selection of genes, classification of protein sequence and, prediction of protein structure.

A novel approach for tumor classification was proposed by [15]. This approach was based on Wavelet Packet Transforms (WPT) and Neighborhood Rough Sets (NRS) as tools for effective features extraction and selection. WPT performed features extraction, and then decision tables are formed. High classification with few attributes was reduced by NRS. The proposed method was applied on three gene expression datasets and experimental results showed feasibility and effectiveness.

A feature selection algorithm based on rough set theory had been suggested in [16]. It depended on selecting reduced set of genes from microarray data based on relevance and significance criteria of the selected genes. The importance of rough set theory here was computing both criteria to produce theoretical analysis justification. The proposed algorithm performance, along with a comparison with other related methods had obtained 100% predictive accuracy for three cancer and two arthritis data sets.

Suitable solutions were provided in [17] to solve two important issues exist in the data represented in information table. Those two solutions were applied based on concepts exist in Rough Set Methodology. The first issue was the indiscernible objects that were represented several times and solved by data reduction. This reduction included eliminating the unnecessary attributes and deletion of identical rows. The second issue was the existence of many redundant attributes and solved by dimensionality reduction. This solution used simplifying discernibility function to get reducts which used for generating if-then rules for classification.

A Promising framework was introduced in [18] to handle the complexities of protein structure prediction. Rough set improved harmony search quick reduct algorithm to be used for selecting the optimum number of features. More compact rules were generated via Rough set classification which showed a higher overall accuracy rates compared with classification algorithms in Weka.

III. Datasets and Materials

This section lists the datasets with their types and full description of the experiment conditions. Moreover, the tools, and computational power needed for accomplishing this study are presented. It involves Data Sources, Software Packages and Tools and, Computational Requirements.

A. *Data Sets*

The A. thaliana reference genome FASTA sequences have been downloaded from the Ensemble FTP (https://plants.ensembl.org/info/website/ftp/index.html). The RNA-seq FASTQ data files for A. thaliana under heat-stress were downloaded from the NCBI website. These data files represent an experiment that is performed on A. thaliana plants in Moscow, Russia. A Third leaf was collected from 15 plants of age 21 days after heat treatment at 42°C for 1, 3, 6, 12, and 24 hours. The experiment was accomplished with 2 replicates for each of the mentioned 5 different time points to resolve false-positive calls at the low end of signal detection [5]. This experiment was SINGLE stranded – Illumina Hi-Seq 2000 – RNA-seq libraries from TRANSCRIPTOMIC PolyA RNA. Ten files were downloaded and their attributes and description are listed in Table I.

B. *Software Packages and Tools*

The following list contains software tools and packages that were integrated with custom code to carry out the execution of the various processes along the presented workflow.

- STAR (Spliced Transcripts Alignment to a Reference): 2.5.3a [March 17, 2017] version available on BA-HPC.
- SAMtools (Sequence Alignment/Map tool): 1.9-intel-b [2018] version available on BA-HPC.
- BCFtools (Binary Counterpart Format tools): 1.9-foss-b [2018] version available on BA-HPC.
- SnpEff (variant annotation and effect prediction tool): 4.1d using (Java-1.7.0_80) [2015] version available on BA-HPC.
- Rosetta: version 1.4.41 [May 27 2001].

C. *Computational Requirements*

The experiments conducted in this research are based on the Unix-type operating systems (primarily Linux); it provides a command-line interface and is best run on a high-memory, multicore computer or in a high-performance computing environment. In general, having ~1 GB of RAM per 1 million paired-end reads is recommended. A typical configuration is a multicore server with 256 GB to 1 TB of RAM.

TABLE. I. FILES OF RNA-SEQ READS

Symbol	Replicate Name	Accession
H1_R1	1 hour Replicate 1	SRX1881868
H1_R2	1 hour Replicate 2	SRX1881876
H3_R1	3 hours Replicate 1	SRX1881880
H3_R2	3 hours Replicate 2	SRX1881883
H6_R1	6 hours Replicate 1	SRX1881886
H6_R2	6 hours Replicate 2	SRX1881888
H12_R1	12 hours Replicate 1	SRX1881889
H12_R2	12 hours Replicate 2	SRX1881897
H24_R1	24 hours Replicate 1	SRX1881908
H24_R2	24 hours Replicate 2	SRX1881912

For the research problem presented in this article, the lack of the required computing resources to accomplish the required work could be a challenge. In this study, the used computational resources were provided by The Bibliotheca Alexandrina (bibalex)[1]. The super computer BA-HPC capabilities are used to achieve this work.

IV. METHODOLOGY

This study proposes a promising framework to illustrate how SNPs can be discovered, annotated and, analyzed from RNA-seq data in order to be used to describe genes expression. Methodologies are divided mainly into two phases: (A) SNPs Detection, and (B) SNPs Analysis. Details of both phases and needed resources are being described below.

A. *SNPs Detection*

This phase shows an overview of the steps and methods that are employed to identify the most suitable performing of RNA-seq SNPs detection pipeline in Fig. 1. The employed steps start from creating genome indices, and go along till finding out annotated SNPs.

1) Creating genome indices: Using the A. thaliana reference genome (*.fna) file from (Data Sets) section, Indices are created using STAR tool.

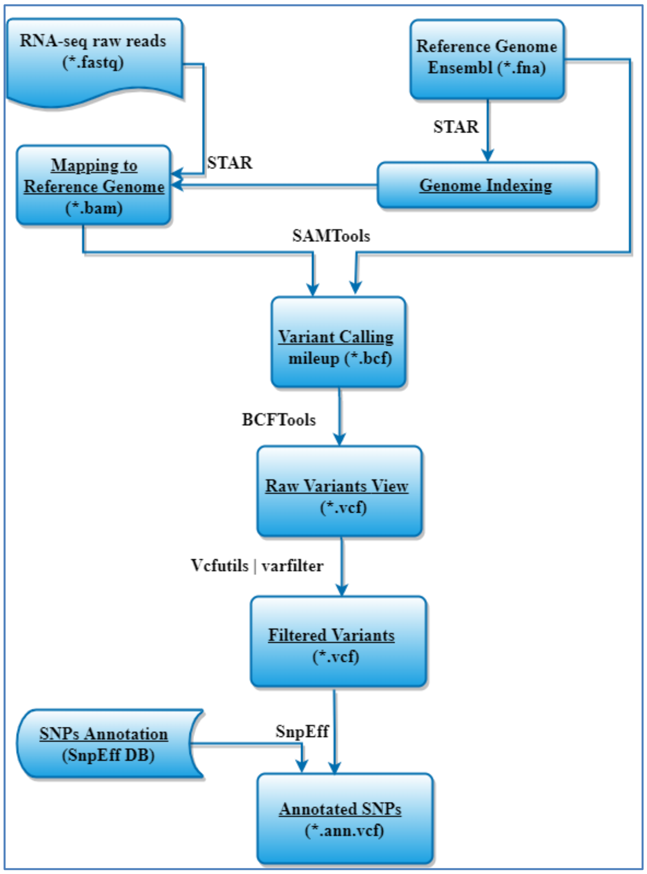


Fig. 1. A Framework for RNA-Seq SNP Detection.

[1] BA-HPC group, Alexandria Library, https://www.bibalex.org/, Egypt, (accessed Oct 2019).

2) Mapping raw reads to reference genome: This step aims to find matches between the reference genome and the sequences of the sampled RNA-seq short reads. During mapping, using existing gene models obtains the maximum advantage to some read mapper in order to map the coordinates accurately.

Using indices files generated by *STAR*, each *individual* read with the reference genome is mapped. Convert mapped reads from SAM to BAM, sort, and index. BAM files are generated sorted by coordinates, so they can be loaded much more quickly.

3) Variant calling: Find deviation from reference genome, the output of both previous steps (RefGenome indices and mapped reads) are used together to perform the variant calling task. This step is done using SAMtools which uses the mpileup command to compile information about the bases mapped to each reference position. It collects summary information in the input BAMs, computes the likelihood of data given each possible genotype and stores the likelihoods in the BCF format Output BCF file is a binary form of the text Variant Call Format (VCF).

4) Obtaining raw variants: BCF file came from the previous step is converted into VCF file using *BCFtools*. It is a collection of utilities to call SNPs and manipulate VCF files. Those utilities are calling SNPs and small indels, annotating and sub-selecting entries from VCF files, querying, filtering, merging VCF files, and converting BCF to human-readable VCF. VCF file has a nice header explaining what the columns mean. Below that header, there are rows of data describing potential genetic variants. Fig. 2 shows a sample for one of the produced (*.vcf) files header and content.

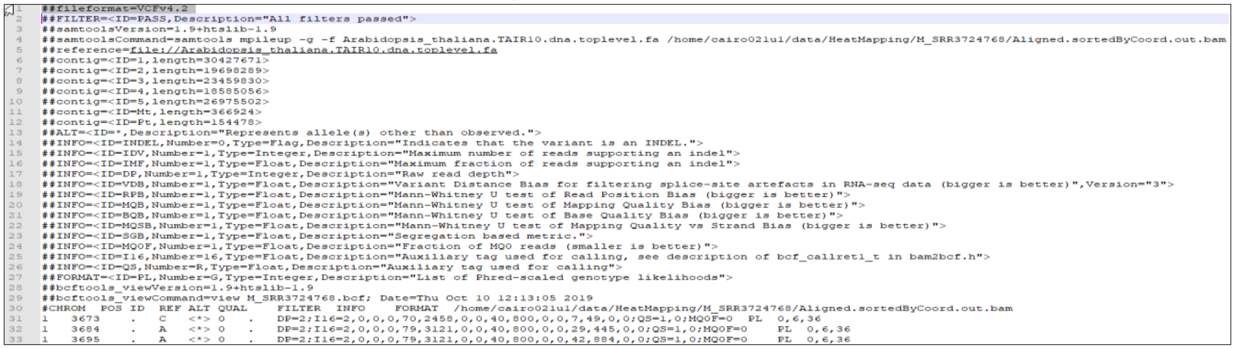
Header contains mandatory lines like the first line (1) and the line containing columns' headers (30). Lines (4) and (5) in the shown sample include data about the reference genome and bam files used to get that vcf variant calling. While lines from (6-12) have the contigs or chromosomes of the reference genome and length of each of them. Moreover, there are optional lines that describe some meta-data about the information in the VCF body shown in Table II.

SAMtools/BCFtools may write the following fields in the 'INFO' tag in VCF/BCF.

- DP: The number of reads covering or bridging POS.
- INDEL: Indicating the variant is an indel.
- I16: contains 16 integers like; sum of reference base qualities, sum of ref mapping qualities, sum of tail distance for ref bases, etc.

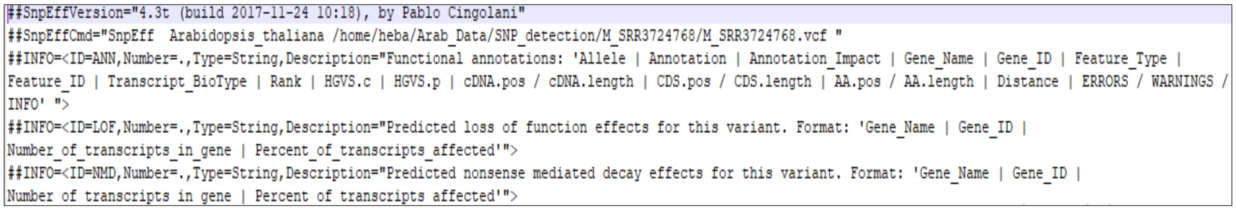
5) Variant filtering: This step applies the prior and does the actual calling. It performs filtering short variants using vcfutils.pl varFilter. This filtration contains; delete duplication, remove low-quality reads (defined by sequencing device), filter unmapped reads and, filter low quality reads/mappings.

6) Finding out annotated SNPs: The last step in the phase of SNPs Detection is to discover the categorization of the variants effects in genome sequences. SnpEff (an abbreviation of "SNP effect") tool is able to analyze and annotate thousands of variants per second and predict their possible genetic effects. Since many databases containing genomic annotations are available with SnpEff distribution, the SNPs annotation is called through SnpEff DB. The output information provided using SnpEff (*.ann.vcf) includes some additional lines at the end of the header which are concerned with the annotation part, as presented in Fig. 3. In this research, 'ANN' field is the main target which includes the information needed to determine the value of the variant as shown in Fig. 4.



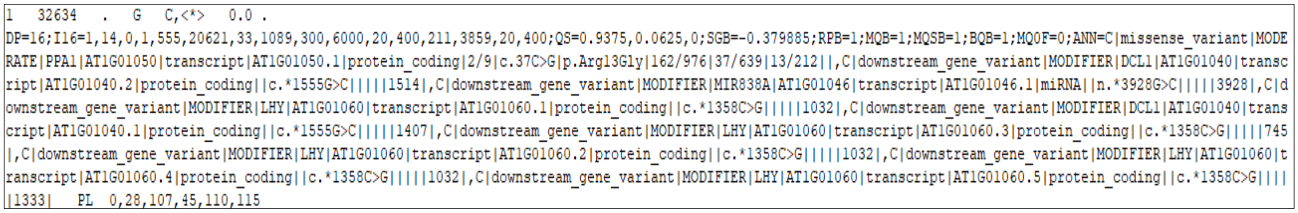
```
1   ##fileformat=VCFv4.2
2   ##FILTER=<ID=PASS,Description="All filters passed">
3   ##samtoolsVersion=1.9+htslib-1.9
4   ##samtoolsCommand=samtools mpileup -g -f Arabidopsis_thaliana.TAIR10.dna.toplevel.fa /home/cairo021ul/data/HeatMapping/M_SRR3724768/Aligned.sortedByCoord.out.bam
5   ##reference=file://Arabidopsis_thaliana.TAIR10.dna.toplevel.fa
6   ##contig=<ID=1,length=30427671>
7   ##contig=<ID=2,length=19698289>
8   ##contig=<ID=3,length=23459830>
9   ##contig=<ID=4,length=18585056>
10  ##contig=<ID=5,length=26975502>
11  ##contig=<ID=Mt,length=366924>
12  ##contig=<ID=Pt,length=154478>
13  ##ALT=<ID=*,Description="Represents allele(s) other than observed.">
14  ##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
15  ##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">
16  ##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of reads supporting an indel">
17  ##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
18  ##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version="3">
19  ##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">
20  ##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">
21  ##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">
22  ##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
23  ##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
24  ##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
25  ##INFO=<ID=I16,Number=16,Type=Float,Description="Auxiliary tag used for calling, see description of bcf_callret1_t in bam2bcf.h">
26  ##INFO=<ID=QS,Number=R,Type=Float,Description="Auxiliary tag used for calling">
27  ##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
28  ##bcftools_viewVersion=1.9+htslib-1.9
29  ##bcftools_viewCommand=view M_SRR3724768.bcf; Date=Thu Oct 10 12:13:05 2019
30  #CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  /home/cairo021ul/data/HeatMapping/M_SRR3724768/Aligned.sortedByCoord.out.bam
31  1   3673    .   C   <*> 0   .   DP=2;I16=2,0,0,0,70,2458,0,0,40,800,0,0,7,49,0,0;QS=1,0;MQ0F=0  PL  0,6,36
32  1   3684    .   A   <*> 0   .   DP=2;I16=2,0,0,0,79,3121,0,0,40,800,0,0,29,445,0,0;QS=1,0;MQ0F=0    PL  0,6,36
33  1   3695    .   A   <*> 0   .   DP=2;I16=2,0,0,0,79,3121,0,0,40,800,0,0,42,884,0,0;QS=1,0;MQ0F=0    PL  0,6,36
```

Fig. 2. VCF File Header and Content.



```
##SnpEffVersion="4.3t (build 2017-11-24 10:18), by Pablo Cingolani"
##SnpEffCmd="SnpEff  Arabidopsis_thaliana /home/heba/Arab_Data/SNP_detection/M_SRR3724768/M_SRR3724768.vcf "
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation | Annotation_Impact | Gene_Name | Gene_ID | Feature_Type |
Feature_ID | Transcript_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS /
INFO' ">
##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss of function effects for this variant. Format: 'Gene_Name | Gene_ID |
Number_of_transcripts_in_gene | Percent_of_transcripts_affected'">
##INFO=<ID=NMD,Number=.,Type=String,Description="Predicted nonsense mediated decay effects for this variant. Format: 'Gene_Name | Gene_ID |
Number_of_transcripts_in_gene | Percent_of_transcripts_affected'">
```

Fig. 3. Header of (*.ann.vcf) File.



```
1   32634   .   G   C,<*>   0.0 .
DP=16;I16=1,14,0,1,555,20621,33,1089,300,6000,20,400,211,3859,20,400;QS=0.9375,0.0625,0;SGB=-0.379885;RPB=1;MQB=1;MQSB=1;BQB=1;MQ0F=0;ANN=C|missense_variant|MODE
RATE|PPA1|AT1G01050|transcript|AT1G01050.1|protein_coding|2/9|c.37C>G|p.Arg13Gly|162/976|37/639|13/212||,C|downstream_gene_variant|MODIFIER|DCL1|AT1G01040|transc
ript|AT1G01040.2|protein_coding||c.*1555G>C|||||1514|,C|downstream_gene_variant|MODIFIER|MIR838A|AT1G01046|transcript|AT1G01046.1|miRNA||n.*3928G>C|||||3928|,C|d
ownstream_gene_variant|MODIFIER|LHY|AT1G01060|transcript|AT1G01060.1|protein_coding||c.*1358C>G|||||1032|,C|downstream_gene_variant|MODIFIER|DCL1|AT1G01040|trans
cript|AT1G01040.1|protein_coding||c.*1555G>C|||||1407|,C|downstream_gene_variant|MODIFIER|LHY|AT1G01060|transcript|AT1G01060.3|protein_coding||c.*1358C>G|||||745
|,C|downstream_gene_variant|MODIFIER|LHY|AT1G01060|transcript|AT1G01060.2|protein_coding||c.*1358C>G|||||1032|,C|downstream_gene_variant|MODIFIER|LHY|AT1G01060|t
ranscript|AT1G01060.4|protein_coding||c.*1358C>G|||||1032|,C|downstream_gene_variant|MODIFIER|LHY|AT1G01060|transcript|AT1G01060.5|protein_coding||c.*1358C>G||||
|1333|   PL  0,28,107,45,110,115
```

Fig. 4. Content Sample of 'ANN' Field.

TABLE. II. VCF META-DATA

Tag	Description
CHROM	No. of chromose that variant belongs to
POS	Position of that variant on that chromosome
REF	Reference sequence at POS involved in the variant
ALT	Comma delimited list of alternative sequence(s)
QUAL	Phred-scaled probability of all samples being homozygous reference
INFO	Semicolon delimited list of variant information

B. SNPs Analysis

The input of this phase is the 10 (*.ann.vcf) files created in the first phase, which include analysis ready variants. This phase includes set of well-ordered processes which are applied to determine the relationship between SNPs biological features and gene expression.

1) Adjustment: This process handles the 10 (*.ann.vcf) files and put them into another flexible form enabling the separation of some specific information to be analyzed (INFO, ANN) tags.

2) Separation of variants from indels: The main goal is analyzing SNPs and their effect on the genome sequence. So, indels are removed to focus on SNPs only.

3) SNPs selection: Choose only SNPs located in the common heat-stress genes of A. thaliana, published in reference databases; DRASTIC[2] and TAIR10[3].

4) Detection of SNPs biological features: Some biological features of SNPs mainly describe the biological value of the detected SNP. They are isolated and been prepared for analysis. Table III lists some of those features, description and, their possible values.

5) Discovery of relationship between SNPs' features and genes differential expression using Rough Set: The most suitable technique to represent this kind of relation is Rough set. Rough set theory has been a methodology of database mining or knowledge discovery in relational databases [9]. The target is to find the set of rules that translate the relationship between the values of the biological features of detected SNPs for some gene and the differential expression of the same gene.

Rough Set Analysis approach has many important advantages like; Discovery of hidden patterns in data, Data reduction (finds minimal sets of data), Evaluating the importance of data, Representing data as sets of decision rules and, Providing the interpretation of obtained result [19].

Rosetta is a general-purpose tool that is not geared towards any particular application domain. The name ROSETTA can be construed as an acronym, for a Rough Set Toolkit for Analysis of Data. It has been put to use by a large number of researchers world-wide, and has resulted in scientific publications in a wide variety of areas. Moreover, it implements features relevant to build and evaluate rough set models in different domains, and offers a highly user friendly environment in which to conduct experiments. In this study, Rosetta is used to generate rough set rules for the predicted SNPs. This will be discussed obviously in the (Generation of Rough set rules) section [20].

6) Measuring the generated rules: To quantify the generated rules, several numerical measures for the rules are illustrated in Definition 1, 2, and 3 and described in Table IV [21] [22].

S is called a decision table, which is denoted by $S = (U, C, D)$. They are called C, D condition and decision attributes, respectively.

Definition 1: Let $S = (U, C, D)$ be a decision table, $\Phi \in For(C)$ and $\Psi \in For(D)$ .The expression *if* Φ *then* Ψ is called a *decision rule* and is denoted by $\Phi \rightarrow \Psi$.

Definition 2: Let $S = (U, C, D)$ be a decision table and $\Phi \rightarrow \Psi$ a decision rule in S. The *certainty factor* of this rule is defined as:

$$Cer_S(\Phi\,,\Psi) = \frac{card(\| \Phi \wedge \Psi \|_S)}{card(\| \Phi \|_S)}$$

It is obvious that $0 \leq Cer_S(\Phi,\Psi) \leq 1$ for every $\Phi \rightarrow \Psi$. This coefficient is widely used in data mining and is called confidence coefficient too.

Definition 3: Let $S = (U, C, D)$ be a decision table and $\Phi \rightarrow \Psi$ a decision rule in S. The coverage factor of this rule is defined as:

$$Cov_S(\Phi\,,\Psi) = \frac{card(\| \Phi \wedge \Psi \|_S)}{card(\| \Psi \|_S)}$$

It is obvious that $0 \leq Cov_S\,(\Phi,\Psi) \leq 1$ for every $\Phi \rightarrow \Psi$.

[2] Gary Lyon, The DRASTIC gene expression database, http://www. drastic.org.uk, (accessed Nov 2018).

[3] The Arabidopsis Information Resource (TAIR), http://www .Arabidopsis.org, (accessed Oct 2019).

TABLE. III. SNPs BIOLOGICAL FEATURES

Feature	Description, and Possible value(s)
Gene Name	Common gene name (PPA1)
Gene ID	Gene ID (AT1G01050)
Annotation (effect or consequence)	Annotated using Sequence Ontology (SO) terms (e.g. chromosome_number_variation, exon_loss_variant, stop_gained, stop_lost, start_lost, etc.)
Annotation_impact	A simple estimation of putative impact / deleteriousness : {HIGH, MODERATE, LOW, MODIFIER}
Feature type	Which type of feature is in the next field (e.g. transcript, motif, miRNA, etc.). It is preferred to use SO terms
Transcript biotype	The bare minimum is at least a description on whether the transcript is {Coding, Noncoding}
cDNA_position / (cDNA_len)	Position in cDNA and trancript's cDNA length
Protein_position / (Protein_len)	Position and number of AA

TABLE. IV. EVALUATION MEASURES FOR ROUGH SET RULES

Measure	Description
Rule Support	The number of samples that represent this rule
Rule Strength	The Rule Support divided by the total number of samples. (The more cases support a rule, the stronger it is)
Rule Certainty (accuracy)	the frequency of objects having Ψ in the set of objects with the property Φ
Decision Coverage	the frequency of objects with the property Φ in the set of objects with the property Ψ

V. RESULTS AND EVALUATION

A. Identification of SNPs in Heat-Stress Genes

Continuous decision values may cause a challenge. In most practical approaches, there are about two to five decision classes. So, if the problem has continuous decision values, they can be split into 2 or 3 intervals [11]. In another example of exon expression values, the decision experimentally was split into three classes by taking 20%: 60%: 20% corresponding to highly expressed, medium expressed and low expressed exons [23]. Similarly, in this study the decision is divided only into two classes, by taking the highest expressed (Yes): the lowest expressed (No) genes.

DRASTIC and TAIR10 reference databases are used as trusted sources for the highest expressed heat-stress genes for A. thaliana. The union of heat-stress genes in those two databases are (225) unique genes. Next, all SNPs that are located into those set of genes exist in the resulting 10 (*.ann.vcf) files are being selected. The number of heat-stress genes detected in each replicate and also numbers of their identified SNPs are listed in details in Table V.

For balance, an equalized set of the lowest expressed heat-stress genes are selected from work presented in [24], to analyze their SNPs features too. About (200) genes are selected and their SNPs are got from the 10 (*.ann.vcf) files. The number of the lowest expressed heat-stress genes detected in each replicate and, number of their identified SNPs are listed in details in Table VI.

B. Capturing Biological Features of Identified SNPs

After that, the biological features of SNPs for both groups of genes, for the highest and lowest expressed heat-stress genes, are picked up from the 10 (*.ann.vcf) files. The most effective biological features exist in these annotation files due to the rough set are (Gene Name, Gene ID, Annotation, Annotation Impact, Feature Type and, Transcript BioType). The top-ranked attributes (biological features) are used to build a rule-based classifier using the Rosetta system.

C. Generation of Rough Set Rules

An information system or information table can be viewed as a table, consisting of objects (rows) and attributes (columns). The captured set of SNPs are used to discover finite set of rules that can describe whether the genes, those SNPs belong to, are heat-stress or not. Rules were generated in *Rosetta* [20] with the manual reducer which determines decision rules (Heat Expressed: Yes; No) based on characterization of a set of objects in terms of attribute values (SNPs biological features). Table VII explores the given replicates and their total number of objects, number of (Yes) decision, number of (No) decision, and the number of resulting rules. Total number of Rules (251) represents the sum of all rules generated over the 10 replicates. However, the set of non-repeatable rules shared between the 10 replicates is (32) rules.

Table VIII presents samples of the resulting rules after applying the rough set characterization. It shows the values of the chosen condition attributes based on the Rough set reduction, and the decision attribute for each rule.

D. Rules Evaluation

To quantify the generated rules, three main numerical parameters for the rules are defined: Rule Strength, Rule Certainty (accuracy) and, Decision Coverage.

Those parameters are calculated for the generated set of rules by applying Definition 1, 2, 3 and, Table IV mentioned in section (Measuring the generated rules). The pie chart shown in Fig. 5 displays the Rule Strength of all rules, showing rules that have the highest strength percentages. Moreover, Fig. 6 shows the different Rule Certainty in a bar chart. Rules that have the same condition values but different decision value (Yes, No) are represented in adjacent bars. Finally, Decision Coverage of rules is represented in Fig. 7 for (Yes) decision rules and Fig. 8 for (No) decision rules.

TABLE. V. THE HIGHEST EXPRESSED HEAT-STRESS GENES AND THEIR SNPS IN ALL REPLICATES

Rep. Name	H1_R1	H1_R2	H3_R1	H3_R2	H6_R1	H6_R2	H12_R1	H12_R2	H24_R1	H24_R2
No.Genes	140	64	172	171	184	186	186	190	183	173
No. SNPs	2353	440	5158	2971	4860	2799	3930	3020	3446	3041

TABLE. VI. THE LOWEST EXPRESSED HEAT-STRESS GENES AND THEIR SNPS IN ALL REPLICATES

Rep. Name	H1_R1	H1_R2	H3_R1	H3_R2	H6_R1	H6_R2	H12_R1	H12_R2	H24_R1	H24_R2
No.Genes	95	22	134	144	157	159	165	170	158	144
No. SNPs	424	58	889	898	1181	1013	1241	1195	1177	853

TABLE. VII. REPLICATES OBJECTS AND THEIR RULES

Rep. Name	Objects	(Yes)	(No)	Rules
H01_R1	2775	2351	424	24
H01_R2	498	440	58	16
H03_R1	6047	5158	889	23
H03_R2	3863	2965	898	27
H06_R1	6037	4856	1181	24
H06_R2	3806	2793	1013	26
H12_R1	5164	3923	1241	27
H12_R2	4210	3015	1195	28
H24_R1	4622	3445	1177	29
H24_R2	3891	3038	853	27
Total	**40913**	**31984**	**8929**	**251**

TABLE. VIII. SAMPLE OF GENERATED RULES

Annotation	Annotation_Impact	Feature_Type	Transcript_BioType	Heat_Expressed
missense_variant	MODERATE	Transcript	protein_coding	Yes
stop_lost	HIGH	Transcript	protein_coding	Yes
downstream_gene_variant	MODIFIER	Motif	protein_coding	No
intergenic_region	MODIFIER	intergenic_region	Noncoding	Yes
start_lost	HIGH	Transcript	protein_coding	No

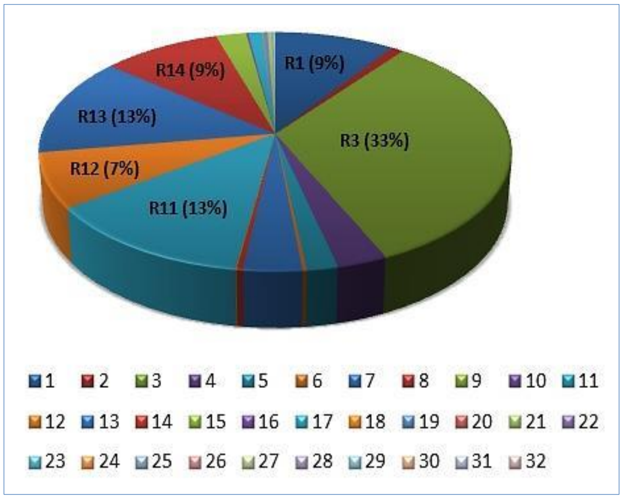


Fig. 5. Rules Strength.

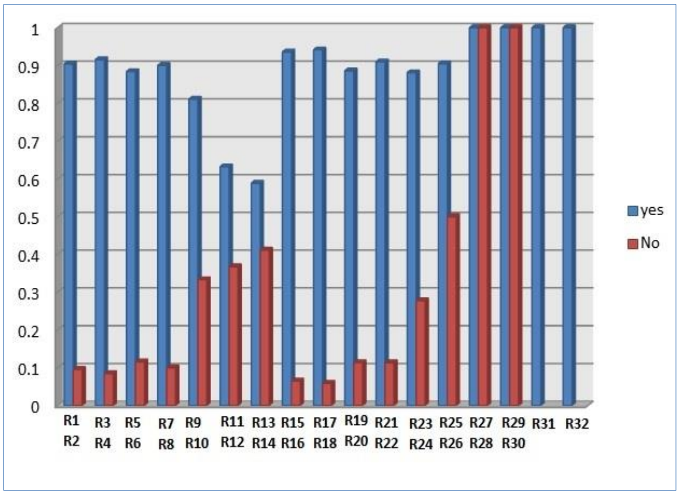


Fig. 6. Rules Certainty.

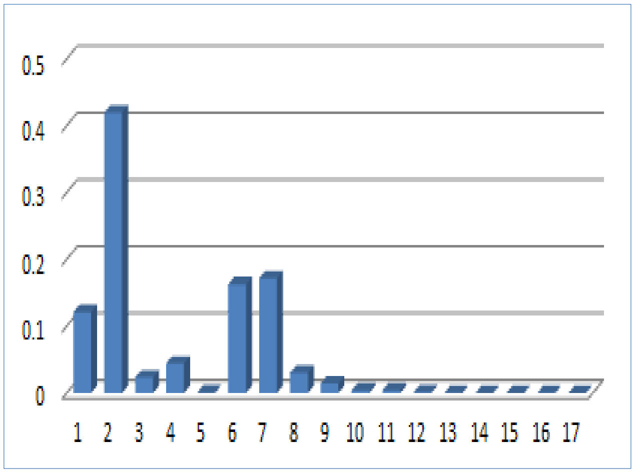


Fig. 7. Decision Coverage (Yes).

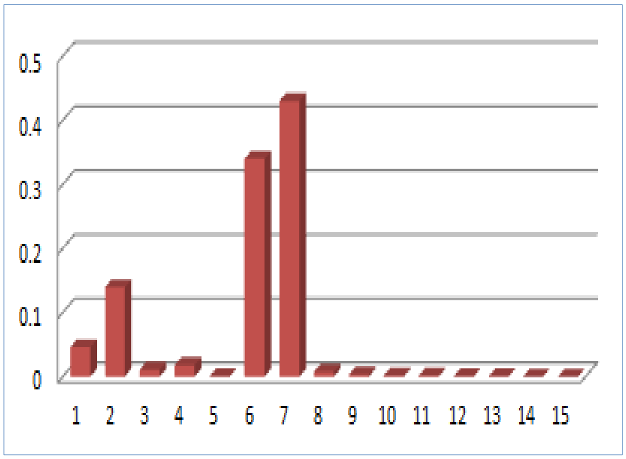


Fig. 8. Decision Coverage (No).

VI. Conclusion

The ultimate goal of this research is to find the relationship between the set of heat-stress expressed genes and their detected SNPs biological features in A. thaliana RNA-seq raw reads. Utilizing rough set-based rule induction resulted in set of descriptive rules which can draw the correlation between those two significant concepts; genes and SNPs. A promising analysis framework was presented to detect SNPs in RNA-seq raw reads then using annotations of those SNPs to figure out their biological features. Additionally, about (225) unique genes got from DRASTIC and TAIR10 databases were used to represent the highly expressed heat-stress genes for A. thaliana. However, (200) genes were selected to represent the lowly expressed heat-stress genes for the same plant. The top-ranked biological features of SNPs for both groups of genes with decision rules (Heat Expressed: Yes; No) were utilized to build a rule-based classifier using the Rosetta system.

The system stated set of (32) non-repeatable rules. Results showed acceptable outcomes and, evaluation had been applied to check the suitability of the generated rules using Rule Strength, Rule Certainty and Decision Coverage. In conclusion, relation between SNP calls and expressed genes in RNA-seq data can be a very useful by-product and increases the amount of knowledge for SNPs discovery and analysis in functional genomics research. With this important result in mind, this method can be verified using in vivo tests to improve the work results. Moreover, generating rules for more species of the same plant may improve complete and well-defined base for machine learning approach to researchers of all expertise levels.

References

[1] Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J. & MacManes, M. D. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols, 8(8), 1494.

[2] Cornwell, M., Vangala, M., Taing, L., Herbert, Z., Köster, J., Li, B., & Pun, M. (2018). VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. BMC bioinformatics, 19(1), 135.

[3] Zhao, Y., Wang, K., Wang, W. L., Yin, T. T., Dong, W. Q., & Xu, C. J. (2019). A high-throughput SNP discovery strategy for RNA-seq data. BMC genomics, 20(1), 160.

[4] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., & McKenna, A. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics, 43(5), 491.

[5] Yu, Y., Wei, J., Zhang, X., Liu, J., Liu, C., Li, F., & Xiang, J. (2014). SNP discovery in the transcriptome of white Pacific shrimp Litopenaeus vannamei by next generation sequencing. PloS one, 9 (1), e87218.

[6] Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P., & Kocher, J. P. A. (2016). Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. Briefings in bioinformatics, 18(6), 973-983.

[7] Quinn, E. M., Cormican, P., Kenny, E. M., Hill, M., Anney, R., Gill, M., Corvin AP & Morris, D. W. (2013). Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. PloS one, 8(3), e58815.

[8] Hvidsten, T. R., & Komorowski, J. (2007). Rough sets in bioinformatics. In Transactions on rough sets VII (pp. 225-243). Springer, Berlin, Heidelberg.

[9] Rissino, S., & Lambert-Torres, G. (2009). Rough set theory—fundamental concepts, principals, data extraction, and applications. In Data mining and knowledge discovery in real life applications. IntechOpen, (pp. 35-60).

[10] Dong, J., Zhong, N., & Ohsuga, S. (1999). Probabilistic rough induction: the GDT-RS methodology and algorithms. In International Symposium on Methodologies for Intelligent Systems (pp. 621-629). Springer, Berlin, Heidelberg.

[11] Komorowski, J. (2014). Learning rule-based models-the rough set approach. Amsterdam: Comprehensive Biomedical Physics.

[12] Frachon, L., Bartoli, C., Carrère, S., Bouchez, O., Chaubet, A., Gautier, M.,Rody, D. & Roux, F. (2018). A genomic map of climate adaptation in Arabidopsis thaliana at a micro-geographic scale. Frontiers in plant science, 9, 967.

[13] Korani, W., Clevenger, J. P., Chu, Y., & Ozias-Akins, P. (2019). Machine Learning as an Effective Method for Identifying True Single Nucleotide Polymorphisms in Polyploid Plants. The Plant Genome, 12(1).

[14] Hassanien, A. E., Al-Shammari, E. T., & Ghali, N. I. (2013). Computational intelligence techniques in bioinformatics. Computational biology and chemistry, 47, (pp. 37-47).

[15] Zhang, S. W., Huang, D. S., & Wang, S. L. (2010). A method of tumor classification based on wavelet packet transforms and neighborhood rough set. Computers in biology and medicine, 40(4), (pp. 430-437).

[16] Maji, P., & Paul, S. (2011). Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. International Journal of Approximate Reasoning, 52(3), (pp. 408-426).

[17] Jain, K., & Kulkarni, S. (2020), Multi-reduct Rough Set Classifier for Computer-Aided Diagnosis in Medical Data. In Advancement of

Machine Intelligence in Interactive Medical Image Analysis. (pp. 167-183). Springer, Singapore.

[18] Bagyamathi, M., & Inbarani, D. H. H. (2017). Prediction of Protein Structural Classes using Rough Set based Feature Selection and Classification Framework. Journal of Recent Research In Engineering and Technology (JRRET), 4.

[19] Pawlak, Z. (1997). Vagueness - a Rough Set View, In: Lecture Notes in Computer Science- 1261, Mycielski, J; Rozenberg, G. & Salomaa, A. (editors), (pp. 106-117), Springer, ISBN 3-540- 63246-8, Secaucus-USA.

[20] Komorowski J., Øhrn A., Skowron A., (2002) The ROSETTA Rough Set Software System. In: W. Klo¨sgen, Zytkow J, eds. Handbook of Data Mining and Knowledge Discovery, Oxford University Press.

[21] Pawlak, Z. (2002). Rough sets and intelligent data analysis. Information sciences, 147(1-4), 1-12.

[22] Vashist, R., & Garg, M. L. (2011). Rule generation based on reduct and core: A rough set approach. Int. J. Comput. Appl, 29(9), 0975-8887.

[23] Enroth, S., Bornelöv, S., Wadelius, C., & Komorowski, J. (2012). Combinations of histone modifications mark exon inclusion levels. PloS one, 7(1), e29911.

[24] Zaki. H., Nassef M., Farouk A., & Badr A. (2019). A proposed RNA-seq analysis workflow to study heat-stress genes in Arabidopsis thaliana. Bioscience Research, 16(3), (pp. 2641-2654),

Feature Selection in Text Clustering Applications of Literary Texts: A Hybrid of Term Weighting Methods

Abdulfattah Omar
College of Science and Humanities
Prince Sattam Bin Abdulaziz University, Saudi Arabia
Department of English, Faculty of Arts, Port Said University

Abstract—**The recent years have witnessed an increasing use of automated text clustering approaches and more particularly Vector Space Clustering (VSC) methods in the computational analysis of literary data including genre classification, theme analysis, stylometry, and authorship attribution. In spite of the effectiveness of VSC methods in resolving different problems in these disciplines and providing evidence-based research findings, the problem of feature selection remains a challenging one. For reliable text clustering applications, a clustering structure should be based on only and all the most distinctive features within a corpus. Although different term weighting approaches have been developed, the problem of identifying the most distinctive variables within a corpus remains challenging especially in the document clustering applications of literary texts. For this purpose, this study proposes a hybrid of statistical measures including variance analysis, term frequency-inverse document frequency, TF-IDF, and Principal Component Analysis (PCA) for selecting only and all the most distinctive features that can be usefully used for generating more reliable document clustering that can be usefully used in authorship attribution tasks. The study is based on a corpus of 74 novels written by 18 novelists representing different literary traditions. Results indicate that the proposed model proved effective in the successful extraction of the most distinctive features within the datasets and thus generating reliable clustering structures that can be usefully used in different computational applications of literary texts.**

Keywords—Feature selection; frequency; PCA; term weight; text clustering; TF-IDF; variance; VSC

I. Introduction

With the increasing access to e-texts and the availability and power of computational tools, there has been an increasing amount of humanities computing literature on text analysis and interpretation. Studies of this kind are generally classified under the broad heading computer-assisted text analysis (CATA). CATA includes numerous applications including authorship attribution, stylometric analysis, theme analysis, the use of imagery, genre classification, characterization, and textual analysis [1-4]. In spite of the effectiveness of VSC methods in resolving different problems in these disciplines and providing evidence-based research findings, the problem of feature selection remains a challenging one. For reliable text clustering applications, a clustering structure should be based on only and all the most distinctive features within a corpus. For this purpose, this study proposes a hybrid of statistical measures including variance analysis, term frequency-inverse document frequency, TF-IDF, and Principal Component Analysis (PCA) successively for selecting only and all the most distinctive features that can be usefully used for generating more reliable document clustering that can be usefully used in authorship attribution tasks. The study is based on a corpus of 74 novels written by 18 novelists representing different literary traditions.

II. Literature Review

The literature suggests that text clustering (simply putting similar texts together) is central in almost all CATA applications [5, 6]. It is used as a starting point for many of the CATA applications including thematic analysis, genre classification, stylometry, and authorship attribution [5, 7-14]. It is known that studies in these disciplines have always been done using non-computational methods. With the development of computational approaches; however, critics and researchers have come to think about how effective computational approaches are in identifying meanings within texts. Now, it is often assumed that computational approaches prove effective in better understanding texts in question [15]. This is best described as a process of decoding meanings within texts [16]. Despite the relative success of studies of this kind, they are met with a strong wave of objections from a number of critics and scholars. They still think that their success in the interpretation of texts is still far from detecting what a text is exactly about [17, 18]. This can be attributed to the unfamiliarity of the world of computational theory and methodology to literary scholars. Ramsay [19] suggests that "the inability of computing humanists to break into the mainstream of literary critical scholarship may be attributed to the prevalence of scientific methodologies and metaphors in humanities computing research" [19, P. 167]. One might even suggest that the unfamiliarity with computational and mathematical approaches has generated in literary scholars the belief that all computational and statistical approaches are somehow antithetical to literary critical approaches. This would explain the gap we see between literary critical theory on the one hand and computer-based text analysis and quantitative approaches on the other: the majority of critical theory researchers have never argued the need for using computational mathematical approaches to supplement widely

Paper Submission Date: January 19, 2020
Acceptance Notification Date: February 12, 2020

used critical approaches [20-22]. Critics of the involvement of computational methods in literary criticisms always argue that human reasoning is crucial and can never be replaced in understanding and interpreting texts. They argue that so far there is no computer-assisted system that is capable of accounting only for all the linguistic and meta-linguistic features of texts.

Defenders of computational text analysis, on the other hand, argue that the use of a computational framework in literary studies is objective, quantifiable, and methodologically consistent [23-27]. Hockey asserts that computational tools are useful adjuncts to literary criticism. She contends that without computational tools, critics have only human reading, intuition, and serendipity to use in literary criticism. Many of the defenders even go beyond that, arguing "without the computer, the interpreter is nothing more than some Romantic Aeolian harpist drowning in the phenomenological abyss of their own impressions" [19, P. 168]. This can be reflected in the significant increase in the application of computational methods in literary studies over the recent years. In numerous thematic reviews of different literary texts, text clustering is central in thematic analysis applications. This is the arrangement of texts by topic with the purpose of investigating thematic interrelationships within texts [7, 9, 14, 28, 29]. The main assumption is that text clustering methods are effective in identifying what a text is about. Consequently, thematic hypotheses can be based on clustering results. It is even argued that computational techniques are effective in generating new insights and interpretative ideas about thematic reviews of different literary texts [14, 28]. Likewise, Ramsay [13] indicates that genre classification which remained distant from computational and mathematical applications for a long time, is now making use of computation technologies and more specifically text clustering approaches to adjudicate some genre classification problems and objectively assign literary texts to appropriate genres. With the high development of text clustering algorithms and methods, genre classification studies draw more heavily on computational methods for more accurate results and better performance [13, 30-35]. Interestingly, the works of Shakespeare have been the subject of many computer-based genre classifications [13, 34, 36]. Using cluster analysis methods, Jockers classified 37 Shakespearean plays into three main clusters, comedy, history, and tragedy as shown in Fig. 1.

The literature also suggests that text clustering methods are now used in stylometry- the investigation of the quantitative properties of an author's style, and authorship attribution [33, 37-45]. The claim is that results based on computer-based methods are accepted by many as more accurate than those based on conventional non-computational methods. In spite of the potentials of computational approaches and text clustering methods especially the capacities for analyzing large quantities of data and generating results that are objective and replicable, there are still many problems and challenges with these approaches that may affect the reliability and acceptability of such methods [46-49]. One main problem is the effectiveness of text classifiers to identify and extract only and all the most distinctive features or variables within a corpus for generating clustering structures that can be usefully used in different applications. Although the issue has been extensively investigated in different disciplines including data mining and information retrieval, very little has been done in relation to the problem of feature selection in text clustering applications on literary texts. This study addresses this gap in the literature by proposing a model that combines together three statistical methods, namely variance, TF-IDF, and PCA.

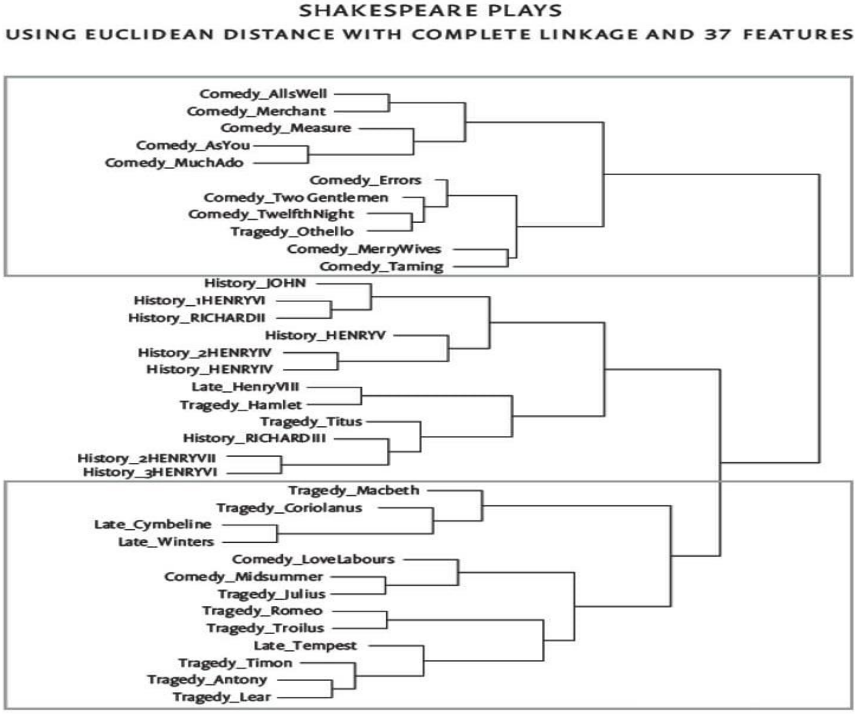


Fig. 1. Jockers' Genre Classification of 37 Shakespearean Plays.

III. Methodology

A. Methods

For the purposes of the study, an experimental study is used where different term-weighting methods are tried to develop a model that best identifies and extracts only and all the most distinctive variables within datasets. Term weighting is a pre-processing step in text clustering applications where each term is assigned its appropriate weight in all documents within a corpus with the purpose of enhancing the text clustering performance [50-52]. Term frequency is still one of the most widely used term weighting approaches in text clustering applications [53-57]. However, term frequency approaches alone are unsuitable for the text clustering of literary texts. This study experiments a combination of different term weighting methods including variance, TF-IDF, and PCA.

1). Variance: Document clustering depends on there being variation in the characteristics of interest to the research question; if there is no variation, the documents are identical and cannot be classified relative to one another [57-60]. The assumption is that variables describing the characteristics of interest are thus only useful for clustering if there is significant variation in the values they take. The intuition for variance is that if a word is used in all or most of the documents in a document collection then that word is more likely to be more important than words that do not vary considerably [53]. Accordingly, documents can be clustered according to the basis of variance. The implication is that variables of significant variation can be retained and variables with little or no variation can be removed. Although variance is an important factor in the assessment of variable importance, retaining the variables that have significant significance is not a guarantee that the data matrix is built up of the most distinctive vectors. Consequently, it should be used along with different term-weighting methods.

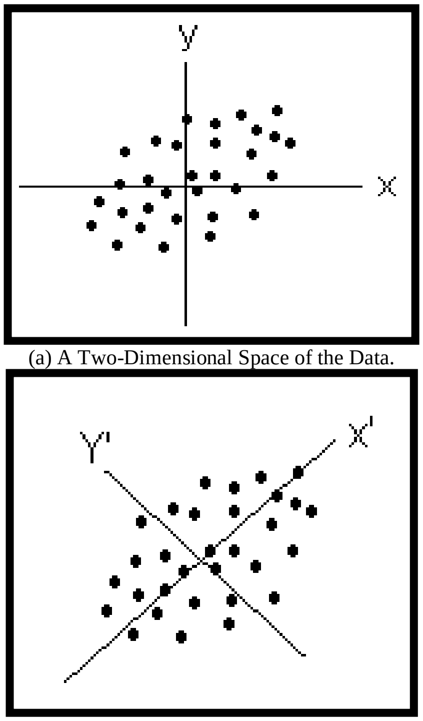
2). TF-IDF: TF-IDF is currently the most common method of calculating term frequency. It is widely used in information retrieval and text mining for identifying the most important variables within datasets. Numerous studies have concluded that TF-IDF works well but they do not explain why this happens [51, 59, 61-64]. The development of IDF came at the hands of Karen Spärck Jones in 1972 with the publication of her article “A statistical interpretation of term specificity and its application in retrieval”. Spärck Jones [65] was the first to propose the measure of term specificity and the term came to be known as Inverse Document Frequency IDF later. The underlying principle of specificity is the selection of particular terms, or rather the adoption of a certain set of effective vocabulary that collectively characterizes the set of documents. In statistical terms, specificity is a statistical property of index terms. Statistical specificity is explained in relation to term frequency. This is based on counting the number of documents in the collection being searched which contain the query [61, 65]. Given that the term frequency of a document is the number of terms it contains, specificity of a term is the number of documents to which it pertains [65]. Logically, if descriptions are longer, terms will be used more often. This may lead to the assumption that if a query is frequently repeated in a document, this document is related to the query. This assumption can be, however, falsified. Spärck Jones [65] argues that a query term that occurs in many documents is not necessarily a good discriminator, and should be given less weight than one which occurs in a few documents. Spärck Jones’ specificity or inverse document frequency IDF was later coupled with term frequency where it has been extensively used in many term weighting schemes [61, 66, 67]. In TF-IDF, the most discriminant terms are the highest TF-IDF variables. This is computed by summing the TF-IDF for each query term and a high weight in TF-IDF is reached by a high term frequency in the given document and a low document frequency of the term in the whole collection of documents [51, 59, 66, 67]. The implication to document clustering is that if the highest TF-IDF variables, which are taken to be the most discriminant terms, are identified, then unimportant variables can be deleted and data dimensionality is reduced.

3). PCA: PCA is one of the basic geometric tools that are used to produce a lower-dimensional description of the rows and columns of a multivariate data matrix [50, 68-70]. The main function of PCA is to find the most informative vectors within a data matrix. Jolliffe [71] explains “The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data sets [71]. It can be thus described as a technique for data quality [69]. To put it simply, PCA performs two complementary tasks: (1) organizing sets of data and (2) reducing the number of variables without much loss of information. In many text clustering applications, PCA is used along with cluster analysis so that clustering is based on the most distinctive vectors within data sets. The literature suggests that PCA is used a great deal in text clustering applications prior to performing cluster analysis. The link between both cluster analysis and PCA is that both are concerned with finding patterns in data. It is sometimes advised that cluster analysis is based on PCA results so that the clustering structure is built on uncorrelated vectors. In spite of the computational mathematical nature of PCA, this section is only concerned with the idea of data reduction.

The main assumption behind PCA is that a matrix with huge data sets can be reduced so that the most distinctive vectors are identified with the purpose of best expressing the data and revealing hidden structures. Although some of the discarded or deleted variables can be important for clustering, PCA works to perform a ‘good’ dimensionality reduction with no great loss of information. The underlying principle of PCA is that it removes correlated variables within datasets so that it describes the covariance relationships among these variables. Fielding [72] explains that PCA “transforms an original set of variables (strictly continuous variables) into derived variables that are orthogonal (uncorrelated) and account for decreasing amounts of the total variance in the original set" [72, P. 16]. The process is done by means of computing the principal components scores by measuring all the variables in the data set. In so doing the variables that have the highest loading or weight are identified as principal components and other variables are discarded. The resulting principal components can then be used in subsequent analyses. Given a two-dimensional vector space with dimensions x and y shown in Fig. 2A, it is possible to transform the distribution of the data as an orthogonal linear representation as shown in Fig. 2B.

The data vector coordinates are then recalculated relative to the new basis. This has the effect of generating a highly correlated 2-dimensional vector space, as shown in Fig. 3.

Finally, the data vector coordinates are then computed on a given principal component. The variables are weighted in such a way that the resulting components account for a maximal amount of variance in the dataset. This is shown in Fig. 4.



(a) A Two-Dimensional Space of the Data.

(b) An Alternative Orthogonal Basis for Data.

Fig. 2. A Representation of a 2-Dimensional Space in Two different Ways.

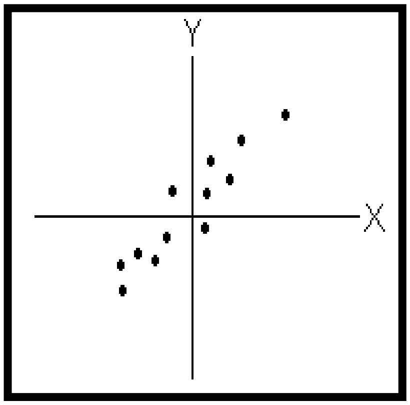


Fig. 3. A Highly Correlated 2-Dimensional Vector Space.

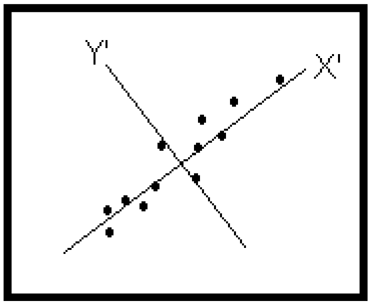


Fig. 4. Testing Variance in Data using TF-IDF.

As seen in the above figure, X' captures almost all the variation in the data, and Y' only a small amount. If Y' is simply disregarded, then the data can be restated in just one rather than the original two dimensions with minimal loss of information, and the data dimensionality has been reduced. The idea is extended to any data dimensionality. So given a data matrix of 100 rows and 1000 columns, the data matrix can be re-described in a lower number of dimensions given that there is redundancy among the variables; that is, they overlap with one another in terms of the information they present. One of the main issues in PCA, however, is determining the number of meaningful principal components (PCs).

B. Data

This is based on a corpus of 74 novels written by 18 novelists representing different literary traditions. These were alphabetically ordered and coded as shown in Table I.

C. Procedures

For text clustering purposes, a data matrix M was built. The matrix included all the 74 novels. Three pre-processing steps were carried out. First, all non-alphabetical ad punctuation marks were removed. The texts were converted into what is called bag of words (BOW). Second, stemming was carried out where only lexical types were retained. Third, texts were normalized in terms of length so that variation in text length has no negative impacts on the reliability of text clustering results. A matrix M was thus generated consisting of 74 rows (the number of texts) and 37435 vectors (all the lexical types in the texts). One major problem with this matrix is data dimensionality. That is, the matrix is composed of so many variables which makes it impossible for any text clustering system to generate reliable clustering structures. In the face of this problem, a model of three term weighting methods was proposed.

First, a variance analysis test using ANOVA was carried out for the $M_{74,\ 37435}$. It was found out that the only 1000 variables are the highest density ones. So it was decided that variables 1-1000 to be retained and variables 1001-37435 to be removed. This can be shown in Fig. 5.

Second, a TF-IDF analysis was carried out. Based on the TF-IDF test shown in Fig. 6, only the variables with the highest TF-IDF values are retained. It was decided that the highest 200 TF-IDF frequencies to be retained. So far, the Matrix is composed of only 200 variables (M74, 200).

TABLE. I. A LIST OF THE SELECTED NOVELS AND SHORT STORIES

Code	Title of the novel/short story	Author
M01	A Daughter of Isis	Nawal El-Saadawi
M02	A Portrait of the Artist as a Young Man	James Joyce
M03	A Shabby Genteel Story	Thackeray
M04	Adventures of Huckleberry Finn	Mark Twain
M05	Aisha	Ahdaf Soueif
M06	Arabian Jazz	Diana Abu Jaber
M07	Basil	Wilkie Collins
M08	Beloved	Toni Morrison
M09	Bird Summons	Leila Aboulela
M10	Birds of Paradise	Diana Abu Jaber
M11	Catherine	Thackeray
M12	Colored Lights	Leila Aboulela
M13	Daisy Miller	Henry James
M14	David Copperfield	Charles Dickens
M15	Dubliners	James Joyce
M16	Elsewhere, Home	Leila Aboulela
M17	Emma	Jane Austen
M18	Far From the Madding Crowd	Thomas Hardy
M19	God Help the Child	Toni Morrison
M20	Hard Times	Charles Dickens
M21	Home	Toni Morrison
M22	I Think of You	Ahdaf Soueif
M23	In Love and Trouble: Stories of Black Women	Alice Walker
M24	In the Eye of the Sun	Ahdaf Soueif
M25	Jude the Obscure	Thomas Hardy
M26	Lady Chatterley's Lover	D. H. Lawrence
M27	Memoirs of a Woman Doctor	Nawal El-Saadawi
M28	Meridian	Alice Walker
M29	Minaret	Leila Aboulela
M30	Mrs. Dalloway	Virginia Woolf
M31	My Name is Salma	Fadia Faqir
M32	Nisanit	Fadia Faqir
M33	Northern Abbey	Jane Austen
M34	Oliver Twist	Charles Dickens
M35	Origin	Diana Abu Jaber
M36	Orlando: A Biography	Virginia Woolf
M37	Paradise	Toni Morrison
M38	Persuasion	Jane Austen
M39	Pillars of Salt	Fadia Faqir
M40	Pride and Prejudice	Jane Austen
M41	Sandpiper	Ahdaf Soueif
M42	Sense and Sensibility	Jane Austen
M43	Song of Solomon	Toni Morrison
M44	Sons and Lovers	D. H. Lawrence
M45	Sula	Toni Morrison
M46	Tar Baby	Toni Morrison
M47	Tess of the D'Urberville	Thomas Hardy
M48	The Bluest Eye	Toni Morrison
M49	The Captain's Doll	D. H. Lawrence
M50	The Cask of Amortillado	Edgar Allan Poe
M51	The Celebrated Jumping Frog of Calaveras County	Mark Twain
M52	The Color Purple	Alice Walker
M53	The Fox	D. H. Lawrence
M54	The Glided Age	Mark Twain
M55	The Luck of Barry Lyndon	Thackeray
M56	The Map of Love	Ahdaf Soueif
M57	The Mayor of Casterbridge	Thomas Hardy
M58	The Moon Stone	Wilkie Collins
M59	The Portrait of a Lady	Henry James
M60	The Rainbow	D. H. Lawrence
M61	The Raven	Edgar Allan Poe
M62	The Tell Tale Heart	Edgar Allan Poe
M63	The Translator	Leila Aboulela
M64	The Voyage Out	Virginia Woolf
M65	The Waves	Virginia Woolf
M66	The Woman in White	Wilkie Collins
M67	To the Lighthouse	Virginia Woolf
M68	Ulysses	James Joyce
M69	Under the Greenwood Tree	Thomas Hardy
M70	Vanity Fair	Thackeray
M71	Washington Square	Henry James
M72	Willow Trees Don't Weep	Fadia Faqir
M73	Women in Love	D. H. Lawrence
M74	Zeina	Nawal El-Saadawi

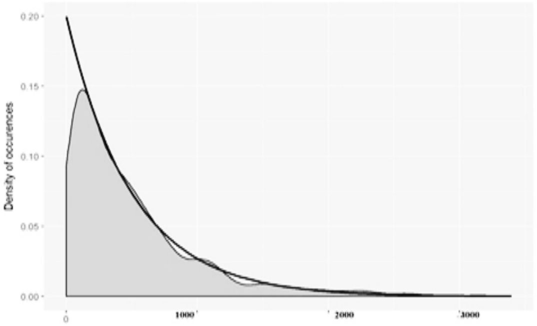


Fig. 5. Variance Analysis Test of the Matrix $M_{74, 37435}$ using ANOVA.

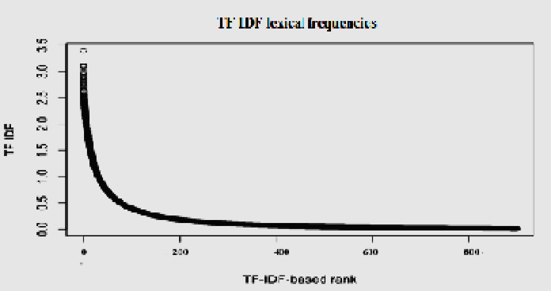


Fig. 6. TF-IDF Test of the Data Matrix $M_{74, 1000}$.

As a final step, PCA was carried out in order to extract only the most distinctive variables within the matrix $M_{74, 200}$. Based on the PCA test shown in Fig. 7, only the first 50 variables were retained. The matrix thus is reduced to only 50 variables which are thought to be the most distinctive features within the corpus.

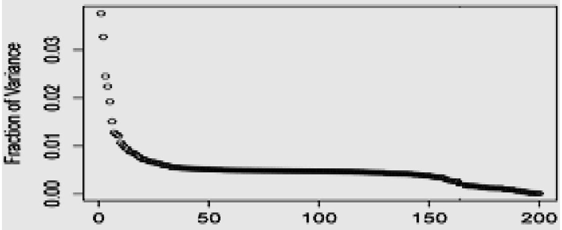


Fig. 7. A PCA of the Data Matrix $M_{74, 200}$.

IV. Analysis

In order to test the effectiveness of the proposed model, cluster analysis is used. This is a technique whereby similar texts are grouped together. The assumption is that there is a strong association between members of the same group or cluster as sharing the same characteristics. The closer texts to each other, the more similar they are and vice versa. These should be texts that can be classified under a given genre and/or written by the same author. $\mathcal{K}$-means clustering, one of the simplest and most popular cluster analysis methods, is used for the task [73-75]. In this process, every data point (the novels in our case) is assigned to the closest center or nearest mean based on their Euclidean distance. Then, new centers are calculated and the data points are updated. This process continues until there is no further iterations and changes within the clusters as seen in Fig. 8.

Using K-means clustering, the texts or data points of the matrix $M_{72, 50}$ were assigned to three groups as seen in Fig. 9. This is based on the number of centroids within the clustering structure. It should be noted, however, that the identification of the number of classes can be different from one researcher to another.

In order to validate the results of the clustering performance, hierarchical cluster analysis is used. Hierarchical clustering is as simple as $\mathcal{K}$-means clustering and it results in a clustering structure consisting of nested partitions. The results can be seen in Fig. 10.

In testing the clustering performance based on our proposed model, results of the K-means clustering are compared to those of hierarchical cluster analysis. Results indicate that there is complete agreement between the members of each cluster/group in the two clustering structures. In the two clustering structures, there are three main distinct classes. These are shown as follows.

Group 1 includes 20 texts. These are 20 novels and short stories. The most distinctive lexical features of this group are words like Islam, veil, marriage, obedience, exile, young, woman, and virginity. Texts included in this cluster are Ahdaf Soueif's Aisha, I Think of You, In the Eye of the Sun, Sandpiper, and The Map of Love; Diana Abu Jaber's Arabian Jazz, Birds of Paradise, and Origin; Fadia Faqir's My Name is Salma, Nisanit, Pillars of Salt, and Willow Trees Don't Weep; Leila Aboulela's Bird Summons, Colored Lights, Elsewhere, Home, Minartet, and The Translator; and Nawal EL-Saadawi's A Daughter of Isis, Memoirs of a Woman Doctor, and Zeina. These texts can be suggested to be belonging to a class of literature known as Anglophone Arabic literature [76-78].

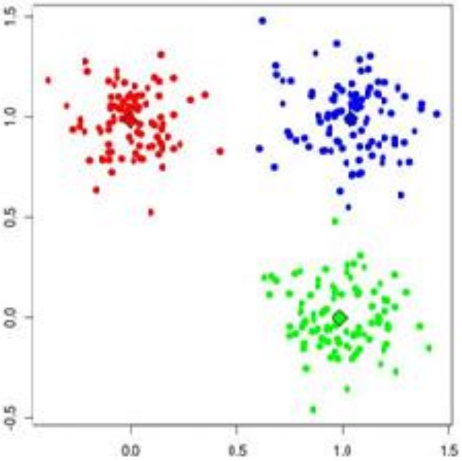


Fig. 8. K-Means Clustering.

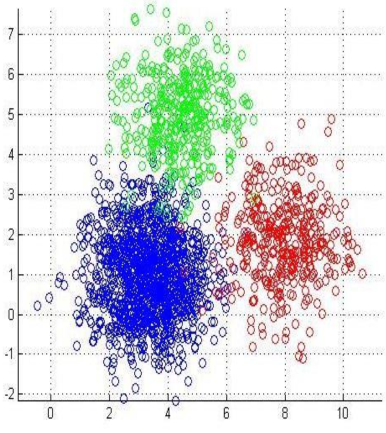


Fig. 9. K-Means Clustering of the Data Matrix $M_{74, 50}$.

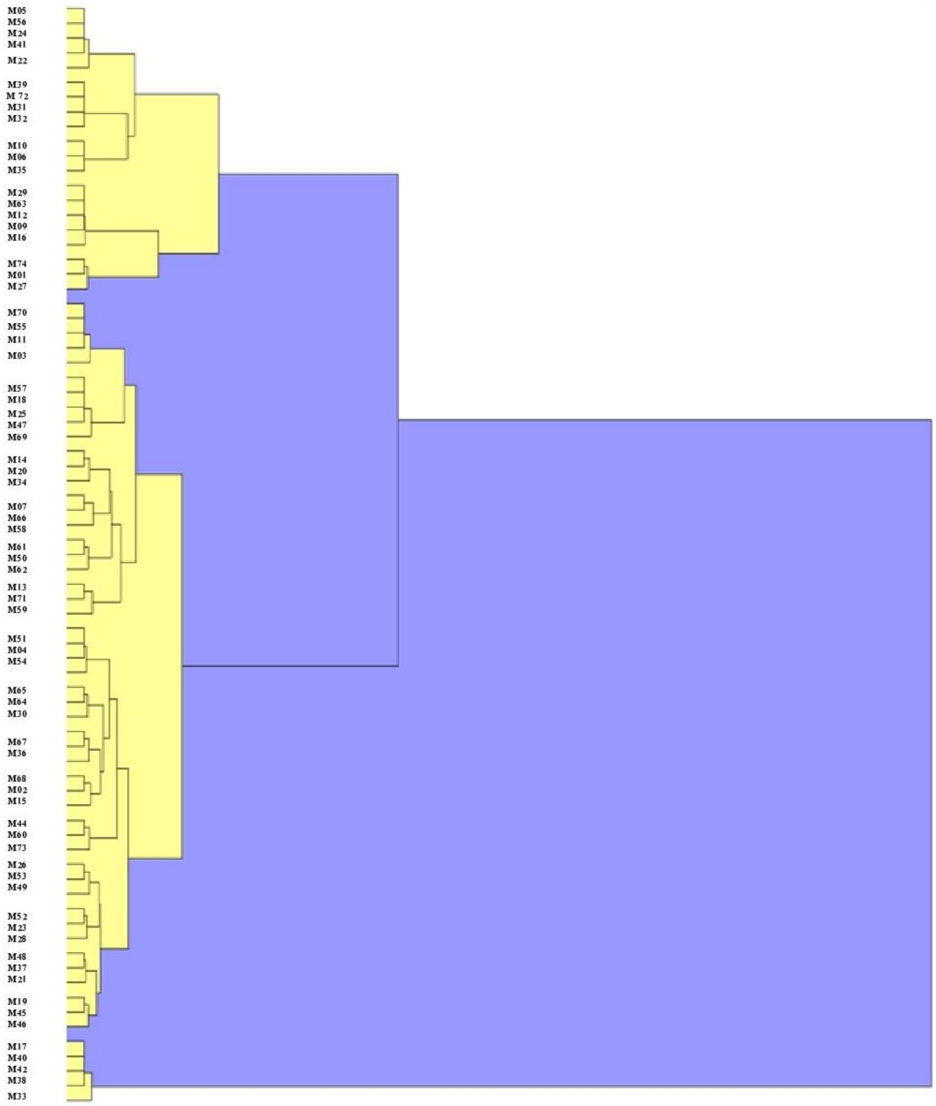


Fig. 10. A Cluster Analysis of the Data Matrix M $_{74, 50}$.

Group 2 is the biggest one as it includes 49 novels and short stories. These include Charles Dickens' Bleak House, David Copperfield, Great Expectations, Hard Times, and Oliver Twist; Thomas Hardy's Jude the Obscure, Far From the Madding Crowd, Tess of the D'Urbervilles, The Mayor of Casterbridge, and Under the Greenwood Tree; Henry James' Washington Square, D. H. Lawrence's Sons and Lovers, and Virginia Woolf's Mrs. Dalloway and The Wave. It can be seen that these texts share some features such as the portrayal of the world as we know it and the discussion of realistic problems. This cluster includes the novels that can be described as realistic novels.

Within Cluster 2, however, we can identify 4 sub-clusters or subclasses. The first subclass includes the texts written by Charles Dickens, Thomas Hardy, William Thackeray, and Wilkie Collins. These are described as social realistic novels [79, 80]. The second subclass includes the texts written by American Victorian writers Henry James, Mark Twain, and Edgar Allan Poe. Poe's texts are, however, distant from those of James and Twain as Poe is adopting a different style, the Gothic tradition, in addressing some realistic problems. The third subclass includes the novels and short stories that best described as modernist novels. These are the books written by James Joyce, D. H. Lawrence, and Virginia Woolf. These represent the modernist novels. The fourth subclass includes 11 novels. These are Toni Morrison's novels Beloved, God Help the Child, Home, Paradise, Song of Solomon, Sula, Tar Baby, and The Bluest Eye; and Alice Walker's In Love and Trouble: Stories of Black Women, Meridian and The Color Purple. These texts are similar to other members of the same group (Cluster 2) in the sense that they all address realistic problems. However, they form a distinct class by themselves as focusing more on the problems of the Black communities.

Group 3 includes only 5 novels. These are Emma, Northanger Abbey, Persuasion, Pride and Prejudice, and Sense and Sensibility. These are all written by Jane Austen and belong to the same literary tradition of what is referred to as the Romanticism [81-83]. It is also clear that the four texts

Emma, Persuasion, Pride and Prejudice, and Sense and Sensibility are very close to each other forming a subclass while Northanger Abbey represents a separate subclass. This hints that the first four texts are thematically similar to each other while Northanger Abbey has a different theme.

It is obvious that the intra-cluster similarity is high. That is, members of each group are similar to each other as the data inside each cluster is similar to one another. It is also clear that each cluster holds information that isn't similar to the other clusters. It can be claimed then that the clustering performance based on our proposed model generated a distinct structure even though different interpretations can be suggested.

V. Conclusion

This study addressed the problem of feature selection in the text clustering applications of literary texts. It proposed an integrated model for extracting the most distinctive features within datasets. The proposed model combines together three different term weighting methods: variance, TF-IDF, and PCA. In order to test the proposed model, a corpus of 74 novels and short stories was designed. Using VSC methods, the selected texts were classified into three distinct classes. It can be concluded that the proposed model is successful in extracting the most distinctive features within datasets. The findings of this study support the claim that traditional or conventional term weighting methods based solely on frequency methods are not sufficient or effective in extracting the most distinctive features within datasets. The proposed model is suggested to be usefully used in CATA applications for its high accuracy in grouping similar texts together.

Acknowledgment

I take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Scientific Deanship, for all technical support it has unstintingly provided towards the fulfillment of the current research project.

References

[1] R. Popping, Computer-assisted Text Analysis, London: SAGE, 2000.

[2] G. Wiedemann, Text Mining for Qualitative Data Analysis in the Social Sciences: A Study on Democratic Discourse in Germany. Springer Fachmedien Wiesbaden, 2016.

[3] D. N. Bengston and U. S. F. S. N. C. R. Station, Applications of Computer-aided Text Analysis in Natural Resources. U.S. Department of Agriculture, Forest Service, North Central Research Station, 2000.

[4] B. D. Hirsch, Digital Humanities Pedagogy: Practices, Principles and Politics. Open Book Publishers, 2012.

[5] B. Yu, "An Evaluation of Text Classification Methods for Literary Study," Lit Linguist Computing, vol. 23, no. 3, pp. 327-343, September 1, 2008 2008.

[6] C. Crompton, R. J. Lane, and R. Siemens, Doing Digital Humanities: Practice, Training, Research. Taylor & Francis, 2016.

[7] B. Yu and J. Unsworth, "Toward Discovering Potential Data Mining Applications in Literary Criticism," presented at the Digital Humanities, 5-9 July 2006, Paris-Sorbo, 2006.

[8] J. Unsworth, "Scholarly Primitives: What Methods do Humanities Researchers Have in Common, and How Might Our Tools Reflect This? ," Symposium on Humanities Computing: Formal Methods, Experimental Practice, King's College, London, 13 May 2000 2000.

[9] S. Argamon and M. Olsen, "Toward Meaningful Computing," Communications of ACM, vol. 49, no. 4, pp. 33-35, 2006.

[10] G. Tambouratzis and M. Vassiliou, "Employing Thematic Variables for Enhancing Classification Accuracy Within Author Discrimination Experiments," Lit Linguist Computing, vol. 22, no. 2, pp. 207-224, June 1, 2007 2007.

[11] C. Labbe and D. Labbe, "A Tool for Literary Studies: Intertextual Distance and Tree Classification," Lit Linguist Computing, vol. 21, no. 3, pp. 311-326, September 1, 2006 2006.

[12] J. Nakamura and J. Sinclair, "The World of Woman in the Bank of English: Internal Criteria for the Classification of Corpora," Lit Linguist Computing, vol. 10, no. 2, pp. 99-110, January 1, 1995 1995.

[13] S. Ramsay, "In Praise of Pattern," TEXT Technology: the Journal of Computer Text Processing, vol. 14, no. 2, pp. 177-190, 2005.

[14] T. Horton, C. Taylor, B. Yu, and X. Xiang, "'Quite Right, Dear and Interesting': Seeking the Sentimental in Nineteenth Century American Fiction," presented at the Digital Humanities, Paris-Sorbonne, France, 5-9 July 2006, 2006.

[15] G. Rockwell, "What is Text Analysis, Really?," Lit Linguist Computing, vol. 18, no. 2, pp. 209-219, June 1, 2003 2003.

[16] P. Boot, "Decoding Emblem Semantics," Lit Linguist Computing, vol. 21, no. suppl_1, pp. 15-27, January 1, 2006 2006.

[17] T. Rommel, "Literary Studies," in ACompanion to Digital Humanities, S. Schreibman, R. Siemens, and J. Unsworth, Eds. Oxford: Blackwell, 2004, pp. 88-97.

[18] T. N. Corns, "Computers in the Humanities: Methods and Applications in the Study of English Literature," Lit Linguist Computing, vol. 2, no. 2, pp. 127-130, January 1, 1987 1987.

[19] S. Ramsay, "Special Section: Reconceiving Text Analysis: Toward an Algorithmic Criticism," Lit Linguist Computing, vol. 18, no. 2, pp. 167-174, June 1, 2003 2003.

[20] T. W. Machan, "Late Middle English Texts and the Higher and Lower Criticisms," in Medieval Literature: Texts and Interpretation. Medieval and Renaissance Texts and Studies, T. W. Machan, Ed. New York: Binghamton, 1991, pp. 3–16.

[21] R. Siemens, "A New Computer-assisted Literary Criticism?," Computers and the Humanities, vol. 36, no. 3, pp. 259-267, 2002.

[22] R. Cohen, The Future of literary theory. New York: Routledge, 1989, pp. xx, 445 p.

[23] S. M. Hockey, Electronic Texts in the Humanities: Principles and Practice. Oxford: Oxford University Press, 2000, pp. xii, 216 p.

[24] M. Terras, J. Nyhan, and E. Vanhoutte, Defining Digital Humanities: A Reader. Taylor & Francis, 2016.

[25] T. H. Howard-Hill, Literary Concordances: A Complete Handbook for the Preparation of Manual and Computer Concordances. Elsevier Science, 2014.

[26] M. L. Jockers, Macroanalysis: Digital Methods and Literary History. University of Illinois Press, 2013.

[27] N. Dershowitz and E. Nissan, Language, Culture, Computation: Computing - Theory and Technology: Essays Dedicated to Yaacov Choueka on the Occasion of His 75 Birthday (no. pt. 1). Springer Berlin Heidelberg, 2014.

[28] C. Plaisant, J. Rose, and B. Yu, "Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces," presented at the Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06), Chapel Hill, North Carolina, 11-15 June 2006, 2006.

[29] R. Horton, M. Olsen, G. Roe, and R. Voyer, "Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the Encyclope´die," presented at the Digital Humanities,. Urbana-Champaign, Illinois, 2-8 June 2007, 2007.

[30] Z. Xiao and A. McEnery, "Two Approaches to Genre Analysis: Three Genres in Modern American English," Journal of English Linguistics, vol. 33, no. 1, pp. 62-82, March 1, 2005 2005.

[31] M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically Categorizing Written Texts by Author Gender," Lit Linguist Computing, vol. 17, no. 4, pp. 401-412, November 1, 2002 2002.

[32] M. Wolters and M. Kirsten, "Exploring the Use of Linguistic Features in Domain and Genre Classification," presented at the Proceedings of the

ninth conference on European chapter of the Association for Computational Linguistics, Bergen, Norway, 1999.

[33] D. I. Holmes, "The Evolution of Stylometry in Humanities Scholarship," Lit Linguist Computing, vol. 13, no. 3, pp. 111-117, September 1, 1998 1998.

[34] M. L. Jockers. (2009, 16 March 2010). Machine-Classifying Novels and Plays by Genre. Available: https://www.stanford.edu/~mjockers/cgi-bin/drupal/node/27.

[35] B. Kessler, G. Numberg, and H. Schtze, "Automatic Detection of Text Genre," presented at the Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, Madrid, Spain, 1997.

[36] S. Ramsay, "Algorithmic Criticism," in A companion to digital literary studies, vol. A companion to digital literary studies, R. G. Siemens and S. Schreibman, Eds. no. Blackwell companions to literature and culture) Malden, MA: Blackwell Publishers, 2007, pp. xx, 620 p.

[37] J. F. Burrows, "Modal Verbs and Moral Principles: An Aspect of Jane Austen's Style," Lit Linguist Computing, vol. 1, no. 1, pp. 9-23, January 1, 1986 1986.

[38] J. F. Burrows, "Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style," Literary and Linguistic Computing, vol. 2, pp. 60-71, 1987.

[39] J. F. Burrows, Computation into criticism : a study of Jane Austen's novels and an experiment in method. Oxford: Clarendon, 1987, pp. xii,255p.

[40] J. F. Burrows, "'An ocean where each kind. . .': Statistical analysis and some major determinants of literary style," Computers and the Humanities, vol. 23 (4), no. 4, pp. 309-321, 1989.

[41] R. A. J. Matthews and T. V. N. Merriam, "Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher," Lit Linguist Computing, vol. 8, no. 4, pp. 203-209, January 1, 1993 1993.

[42] M. Q. Patton, Qualitative Research & Evaluation Methods, 3rd ed ed. London: Sage, 2002, pp. xxiv, 598, [65].

[43] R. S. Forsyth and D. I. Holmes, "Feature-finding for test classification," Lit Linguist Computing, vol. 11 (4), no. 4, pp. 163-174, December 1, 1996 1996.

[44] D. I. Holmes, "Authorship Attribution," Computers and the Humanities, vol. 28, pp. 87-106, 1994.

[45] D. I. Holmes and R. S. Forsyth, "The Federalist Revisited: New Directions in Authorship Attribution," Lit Linguist Computing, vol. 10, no. 2, pp. 111-127, January 1, 1995 1995.

[46] M. W. A. Smith, "Shakespeare, Stylometry and "Sir Thomas More"," Studies in Philology, vol. 89, no. 4, pp. 434-444, 1992.

[47] M. W. A. Smith, "An investigation of Morton's method to distinguish Elizabethan playwrights," Comput. Hum., vol. 19, no. 1, pp. 3-21, 1985.

[48] C. Delcourt, "About the statistical analysis of co-occurrence," Computers and the Humanities, vol. 26, no. 1, pp. 21-29, 1992.

[49] T. Sing, S. Siraj, R. Raguraman, P. Marimuthu, and K. Nithiyananthan, "Cosine similarity cluster analysis model based effective power systems fault identification," International Journal of Advanced and Applied Sciences, vol. 4, no. 1, pp. 123-130, 2017.

[50] M. W. Berry, Survey of Text Mining: Clustering, Classification, and Retrieval. Springer New York, 2013.

[51] I. Zelinka, P. Vasant, V. H. Duy, and T. T. Dao, Innovative Computing, Optimization and Its Applications: Modelling and Simulations. Springer International Publishing, 2017.

[52] S. Sirmakessis, Text Mining and its Applications: Results of the NEMIS Launch Conference. Springer Berlin Heidelberg, 2012.

[53] T. Jo, Text Mining: Concepts, Implementation, and Big Data Challenge. Springer International Publishing, 2018.

[54] C. C. Aggarwal and C. X. Zhai, Mining Text Data. Springer New York, 2012.

[55] K. L. Du and M. N. S. Swamy, Neural Networks and Statistical Learning. Springer London, 2019.

[56] C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms and Applications. CRC Press, 2018.

[57] R. Nisbet, G. Miner, and K. Yale, Handbook of Statistical Analysis and Data Mining Applications. Elsevier Science, 2017.

[58] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer New York, 2010.

[59] S. M. Weiss, N. Indurkhya, and T. Zhang, Fundamentals of Predictive Text Mining. Springer London, 2015.

[60] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery, Model-Based Clustering and Classification for Data Science: With Applications in R. Cambridge University Press, 2019.

[61] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," Journal of Documentation, vol. 60, no. 5, pp. 503-520, 2004.

[62] D. H. Kraft, E. Colvin, and G. Marchionini, Fuzzy Information Retrieval. Morgan & Claypool Publishers, 2017.

[63] B. Mitra and N. Craswell, An Introduction to Neural Information Retrieval. Now Publishers, 2018.

[64] M. Gopal, Applied Machine Learning. McGraw-Hill Education, 2019.

[65] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval " Journal of Documentation, vol. 28, pp. 11-21, 1972.

[66] G. Salton and C. Buckley, "Term-weighing approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513-523, 1988.

[67] G. Salton and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval," Cornell University1987.

[68] W. Härdle and L. Simar, Applied multivariate statistical analysis. Berlin ; New York: Springer, 2003, p. 486 p.

[69] J. E. Jackson, A user's guide to principal components (Wiley series in probability and mathematical statistics. Applied probability and statistics). New York: Wiley, 1991, pp. xvii, 569.

[70] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R. Springer New York, 2013.

[71] I. T. Jolliffe, Principal component analysis, 2nd ed. ed. (Springer series in statistics). Berlin ; London: Springer, 2002, p. 500 p.

[72] A. Fielding, Cluster and Classification Techniques for the Biosciences. Cambridge, UK ; New York: Cambridge University Press, 2007, pp. xii, 246 p.

[73] A. Khan, S. Baseer, and S. Javed, "Perception of students on usage of mobile data by K-mean clustering algorithm," International Journal of Advanced and Applied Sciences, vol. 4, no. 2, pp. 17-21, 2017.

[74] P. Kaur, S. Singla, and S. Singh, "Detection and classification of leaf diseases using integrated approach of support vector machine and particle swarm optimization," International Journal of Advanced and Applied Sciences, vol. 4, no. 8, pp. 79-83, 2017.

[75] Z. Ullah, S. Lee, and M. Fayaz, "Enhanced feature extraction technique for brain MRI classification based on Haar wavelet and statistical moments," International Journal of Advanced and Applied Sciences, vol. 6, no. 7, pp. 89-98, 2019.

[76] L. Maleh and L. A. Maleh, Arab Voices in Diaspora: Critical Perspectives on Anglophone Arab Literature. Rodopi, 2009.

[77] G. Nash, The Anglo-Arab Encounter: Fiction and Autobiography by Arab Writers in English. Peter Lang, 2007.

[78] Z. Halabi, Unmaking of the Arab Intellectual: Prophecy, Exile and the Nation. Edinburgh University Press, 2017.

[79] E. Freedgood, Worlds Enough: The Invention of Realism in the Victorian Novel. Princeton University Press, 2019.

[80] D. David, D. Deirdre, P. E. E. D. David, and C. U. Press, The Cambridge Companion to the Victorian Novel. Cambridge University Press, 2001.

[81] C. Lamont and M. Rossington, Romanticism's Debatable Lands. Palgrave Macmillan UK, 2007.

[82] S. Ailwood, Jane Austen's Men: Rewriting Masculinity in the Romantic Era. Taylor & Francis, 2019.

[83] M. Ferber, Romanticism: A Very Short Introduction. OUP Oxford, 2010.

Fast FPGA Prototyping based Real-Time Image and Video Processing with High-Level Synthesis

Refka Ghodhbani[1], Layla Horrigue[2], Taoufik Saidani[3], Mohamed Atri[4]
Laboratory of Electronics and Microelectronics, Faculty of Sciences of Monastir, University of Monastir, Monastir, Tunisia[1, 2, 3]
Department of Computer Science, Faculty of Computing and Information Technology[1, 3]
Northern Border University Rafha, Saudi Arabia[1, 3]
College of Computer Science, King Khalid University, Abha, Saudi Arabia[4]

Abstract—**Programming in high abstraction level is known by its benefits. It can facilitate the development of digital image and video processing systems. Recently, high-level synthesis (HLS) has played a significant role in developing this field of study. Real time image and video Processing solution needing high throughput rate are often performed in a dedicated hardware such as FPGA. Previous studies relied on traditional design processes called VHDL and Verilog and to synthesize and validate the hardware. These processes are technically complex and time consuming. This paper introduces an alternative novel approach. It uses a Model-Based Design workflow based on HDL Coder (MBD), Vision HDL Toolbox, Simulink and MATLAB for the purpose of accelerating the design of image and video solution. The main purpose of the present paper is to study the complexity of the design development and minimize development time (Time to market: TM) of conventional FPGA design. In this paper, the complexity of the development™ can be reduced by 60% effectively by automatically generating the IP cores and downloading the modeled design through the Xilinx tools and give more advantages of FPGA related to the other devices like ASIC and GPU.**

Keywords—High-level synthesis; FPGA; fast prototyping; real-time image processing; video surveillance; computer-aided design; model-based design; HDL coder; FPGA

I. Introduction

Image processing is taking place in increasingly numerous and complex fields to perform essentially control, inspection and data acquisition tasks [21]. We can cite industrial vision, video surveillance and spatial imagery, medical analysis, robotics ... The last in the list is the field of multimedia with its many recent applications. Image processing follows a well-defined process: to establish, from a raw image, a list of characteristics of the scenes viewed (or objects present in this image) to interpret the content of the image to guide or take a decision [1,2].

Advances in the integration capability of electronic circuits have opened up new perspectives for real-time image and video processing on embedded systems. On the one hand, specific processors can commonly perform billions of operations per second, and on the other hand, reprogrammable components will have billions of logical gates in the near future. These circuits make it possible to realize applications with performances in terms of speed of processing which are constantly increasing.

The past years have seen the explosion of the embedded systems market in many industrial and consumer domains such as telecommunications, satellites, and medical imaging. These increasingly important needs generate an industrial competition where factors such as cost, performance and especially the "Time To Market" become preponderant for the success of a product [25].

In this context, the Field Programmable Gate Array (FPGA) with its large integration and reconfiguration capabilities make it a key component for rapidly developing prototypes. In order to encourage the widespread diffusion of such circuits, it is necessary to improve the development environments to make them more accessible to non-experts in electronics [12,16].

Some applications of advanced computer vision algorithms include video histogram, color conversion system that can be found in modern cameras and many video surveillance [3,4]. Although it might not be necessary to have live video processing capability for many applications, some applications such as color conversion and histogram equalization used for autonomous driving system would require an input stream from cameras to be processed at real time in order to send signals back to the powertrain and steering control unit to respond properly [5,6,7]. FPGAs are a good choice platform for real-time video processing because energy efficiency and the potential to extract highly-parallelized calculations [7,8]. However, hardware development consumes typically more time and human resources than a similar software development would consume [20,22]. For a traditional development based on FPGAs, a good knowledge of digital logic circuit is necessary for Hardware Description Languages (HDLs) such as Verilog and VHDL to construct and config Register-Transfer Level (RTL) circuits in an FPGA [7,17].

Each software offers users with its model block. These tools can help users build the Simulink model with the provided block to generate HDL codes. As compared to the above three software, Simulink HDL Coder by which the generated HDL codes is characterized by its flexibility [18,19].

The goal for this paper is to conceive an automatically very high-level synthesis (VHLS) framework with the following features:

- A short time automatically creating for RTL desired rather than hours or even days.

- To examine the algorithm behavior described in very high-level languages.
- To achieve performances of the designs with the hardware constraints, including area of the target device or frequency.
- To be able to use the currently tools for hight level synthesis available.
- To boost code reuse from 0 to 60%.

The remainder of this paper is organized as follows: in the Section 2, related work on VHLS for image and video processing are presented. Section 3 present high-level synthesis proposed method for image and video prototyping, it discusses the challenge that we met when prototyping this conception, as well as the solutions. Proposed method prototyping and experiment results are given in Section 4. Finally, this paper is finished by a conclusion in Section 5.

A. *Selecting a Template*

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. If you are using A4-sized paper, please close this file and download the file “MSW_A4_format”.

B. *Maintaining the Integrity of the Specifications*

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

II. Related Works

To accelerate embedded real time image and video processing, in the last years, there are some given hardware-accelerated with different high-level synthesis method have been introduced in the automotive area, dedicated DSP-based systems and ASIC solutions compete against FPGA platforms and GPU applications [24].

Abdelgawad, Safar, and Wahba have developed a Canny Edge Detection Algorithm on Zynq Platform [1]. They utilized the Targeted Reference Design (TRD) for Zynq from Xilinx as the platform for the experiment of performance comparison between the CPU processor and the hardware accelerator. They declared memory accesses as the major bottleneck for a real-time video processing system. With proper buffering and directive-based optimizations, they were able to achieve a speedup of 100x on Zynq’s hardware accelerator. They also provide the utilization estimation of the Canny Edge Detector hardware accelerator, but power analysis is missing. By using the TRD, they could inspect the performance improvement more directly thanks to the QT GUI interface. However, the TRD design gave rise to less control of the hardware design as well as software development.

Moreover, by the same research team, Monson, Wirthlin, and Hutchings attempted to optimize another popular image processing algorithm, Sobel filter, using Vivado HLS targeting a Zynq based FPGA [2]. Their first goal was to restructure an existing Sobel filter written in C to a C synthesizable version in Vivado HLS because the original code contains some non-synthesizable portions. Besides the restructuring, the authors discovered and applied three incremental optimizations that can be synthesized in Vivado HLS. The incremental optimization helped their design to achieve a performance of 388 FPS at a resolution of 640x480.

According to [5], the proposed approach operates on the building block level. All these devices seem to depend on hardware simulation and synthesis technology to derive performance scenarios. These figs are only available at a very late stage of the design process after a final FPGA integration.

Cai et al. utilized the capability of Vivado HLS to transform a software face recognition program to a corresponding hardware design based on Zynq platform [9,11]. Their intention was to improve the face detection performance, and the result indicates the performance was improved by up to 80% after migrating the computation onto the hardware. Their face location algorithm relies on color segmentation to detect human faces. The algorithm involves transforming from RGB color space to YCbCr color space, converting the query image to grayscale, and locating the skin color region after erosion and dilation. This algorithm results in straightforward and fast computations. Using color segmentation can be computationally efficient and it is possible to achieve real-time image processing performance. However, a relatively clean background is required for face detection using color segmentation. Also, misrecognition could occur if hands and arms are exposed in the query image.

Now, As opposed to the low-level design approach, Model Based design for FPGA are one of the methods that are based of high-level modeling for image and video processing applications on a very higher level of abstraction. Many various industrial and academic design approaches are available such as Simulink/ Xilinx System generator models, which can convert automatically into a hardware (VHDL) description [10,11].

Author in [13] provides a survey of HLS FPGA design flows for image and video processing applications. Although the given solutions focus on the composition, implementation or HDL generation of an optimal FPGA design, FPGA resources at execution time are neglecting considering reuse of.

III. Proposed Very High-Level Synthesis for Image Processing

According to [14,15], from September 2013 when MathWorks presented its hardware/software workflow for Zynq-7000 focusing Model-Based Design (MBD). Based on this new proposed workflow presented in Fig. 1, models are designed in Simulink using HDL toolbox that can show a completely dynamic system. These include a Simulink model for algorithms targeted for the Xilinx Zynq SoC platform, and Quickly create software- hardware implementations for Zynq platform directly from the algorithm and system design.

A. Rapid Prototyping Flow with HDL coder

This paper presents a video processing system rapid prototyping flow that allows engineers with little to no HDL experience to develop an FPGA based high-performance video processing system.

Fig. 2 demonstrates the rapid prototyping flow from a high-level point of view, and this paper focuses on three of the most essential steps in the flow:

- An FPGA-based SoC video processing system architecture needs to be designed. The system should allow integration of generated IPs from high-level synthesis tools to realize real-time video processing capability.
- The design enables the adoption of FPGA acceleration kernels developed by high-level synthesis tools so that engineers can quickly reconfigure the functionality of the system.
- System-level communications allow users to use software for initializing and configuring modules that are developed in the hardware system.

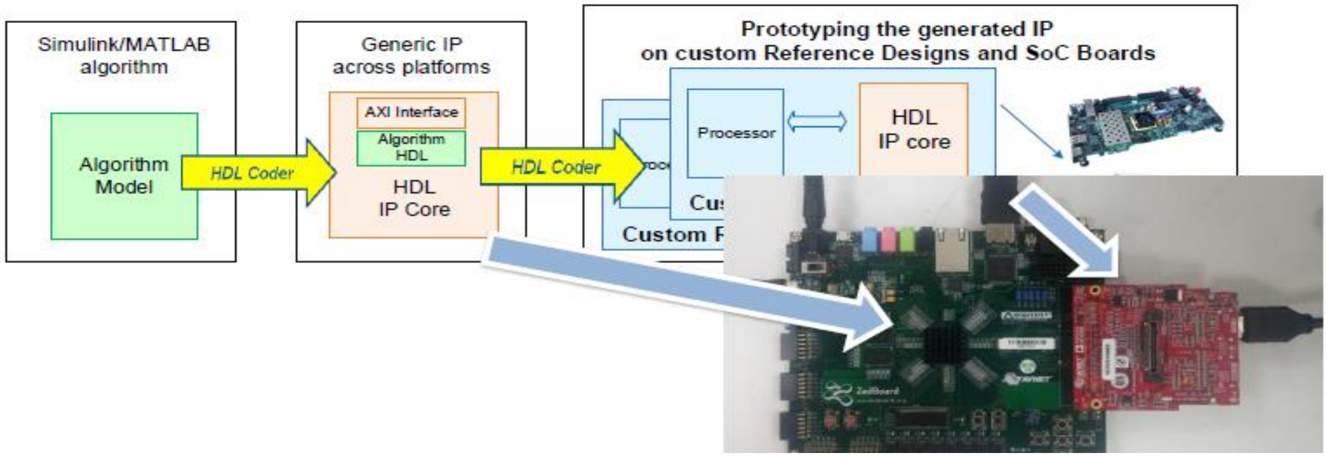

Simulink/MATLAB algorithm
Algorithm Model
HDL Coder
Generic IP across platforms
AXI Interface
Algorithm HDL
HDL IP Core
HDL Coder
Prototyping the generated IP on custom Reference Designs and SoC Boards
Processor
HDL IP core
Cus
Custom F

Fig. 1. HDL Coder Workflow.

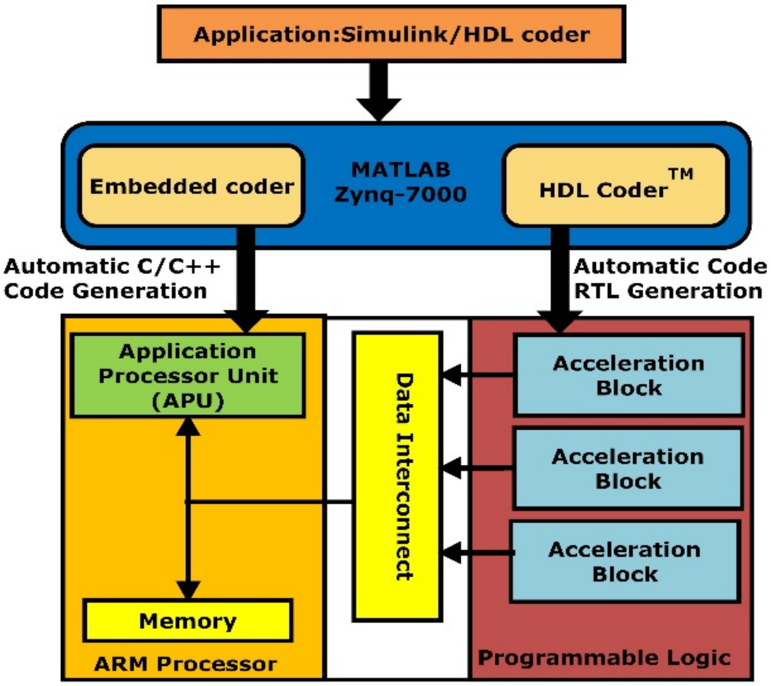


Fig. 2. Model based Design Prototyping with MATLAB/HDL Coder.

The development of the proposed real-time video processing system is divided into two parts: 1) Video processing system architecture design, and 2) Video processing algorithms design. The first part discusses the major components contributed to the video processing system on the Zynq platform including the AXI4 Interfaces used for high throughput data transmissions, while the second part discusses the approaches and optimizations, we have taken for building video processing algorithms using Vivado HLS [25]. The proposed approach is based on the following key:

- Simulation in Simulink used by system designers and algorithm developers is utilized for two reasons. The designer involves creating models for a complete system – communications, image and video processing components. The second reason is to facilitate partition model between hardware and software component and make a good compromise for high-level synthesis.

- High-speed I/O cores for the Xilinx Zynq 7000 platform and IP cores creating can be easily generated by using HDL code generation from HDL coder TM.

- The Zynq Cortex-A9 cores programming, by using of embedded coder from Simulink support rapid embedded software iteration [23].

- Relating to the ARM processing system and programmable logic with support for Xilinx Zynq 7000, automatic AXI4 interfaces cores can be generated.

- Integration with downstream tasks, including software compilation, the executable for the ARM and bit stream generation using Xilinx implementation tools like Vivado and downloading directly to Zynq 7000 platform boards permits a rapid prototyping workflow.

IV. Experiments Applications

The experimental of the proposed approach are investigated by utilizing real-time applications for image and video applications (Fig. 3).

As clarified in, the design is structured and verified in MATLAB and Simulink. Then, it targeted to the Zynq-7000 on the Xilinx Zed board development kit [23]. The real-life application algorithm is achieved on the FPGA fabric through HDL Coder for system acceleration, and it is executed on the ARM Cortex-A9 processor, as shown in Fig. 4.

A. *Color Histogram Equalization*

A histogram can be defined as a diagram that describes how many pixels of an image or a video frame have a particular intensity. It includes different applications in image and video processing [1]. This is due to the simplicity of extracting histogram features. Its characteristics are invariant to image rotation. Moreover, it has low storage demands as compared to the size of the image.

Fig. 5 presents histogram equalization module flowchart operations. The flowchart is composed by two states of operation. When executing, the ready input signal is approved, using a lookup table the input value is transformed, and the histogram is generated. The module enters the second mode of operations once the complete image has been streamed via the module, if the input is not ready so that the new lookup table can be calculated. For the new transformation lookup table generation, the size of the input image approves the accumulating and normalizing histogram module. Lookup table is updated by normalized values, and running mode of operation is done once all 255 values have been updated.

1) Simulink HDL coder Model

a) Video Partition: The video partition component in this design divides a big input frame to 4 small images. For each small frame histogram is generated. The big input image is divided into 160 by 120 small images. There is a connection between each small partition, Frame, pixel and block. This video partition module generates pixel stream and corresponding control signals.

b) HDL Histogram: this module is a part of hardware acceleration. It is designed with HDL coder toolbox and Simulink library. Using the vision HDL toolbox Histogram, the pixel stream of histogram is calculated. The grayscale input pixels are classified into 256 bins.

The model presented by Fig. 6 reads the calculated histogram bins sequentially once the block asserts the read Rdy signal. The bin values are sent for cumulative histogram calculation. After all 256 bin values are read, the model asserts binReset to reset all bins to zero. The collected histogram of each small image is then added together to compute the accumulated histogram of the big image (Fig. 7).

Equalization module: The calculated and accumulated histogram for the current frame generated by a histogram module is processed by equalization module to store the input video. This last input video is delayed by one frame. The uniform equalization is performed to the original video. Finally, a comparison between the original video and the equalized video is done.

2) Synthesis and FPGA implementation: Once the Histogram process is completed and Simulink code of design is successfully converted into hardware design, generated VHDL code of histogram equalization is verified through co-simulation using ModelSim 10.3d software. A further design is processed in Vivado 17.4 Design Suite for synthesis and implementation on Xilinx Zynq xc7z020clg484-2 FPGA device. The logic resources utilized by design with timing performance are presented in Table I. Table I represents the total number of slices and look-up tables used in this design, which indicates entire area occupied in the target device. From the Table II, it is found that the proposed design is working with an estimated speed of 170 MHz by utilizing only 3350 slices. Proposed model is using 2770 lookup tables.

Hardware consumption in any design determines its cost. Therefore, the cost of proposed design is decreased due to lesser hardware utilization. Hence, the suggested design methodology improves efficiency in area and provides good choice in terms of low-cost hardware. The resource usage and maximum frequency for this module are shown in Table II.

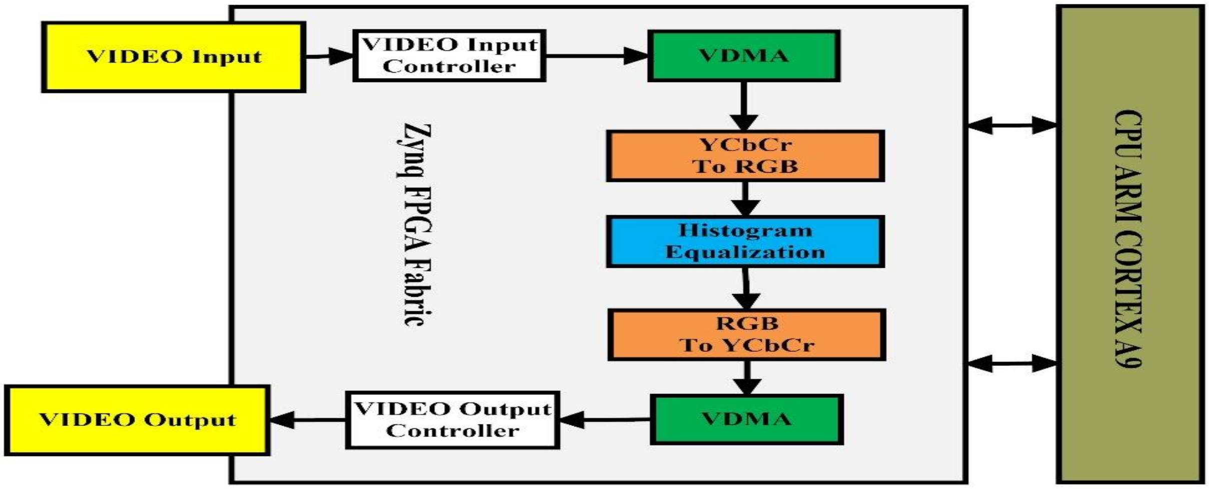


Fig. 3. Low-Light Video Processing Architecture Implemented on the FPGA.

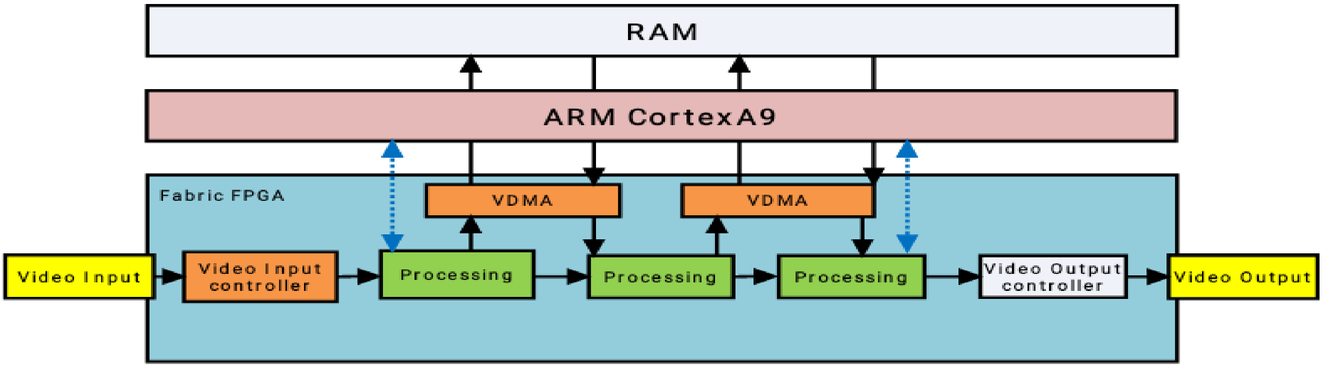


Fig. 4. Real-Time Video Processing Architecture based on Zynq 7000 FPGA and ARM Processor.

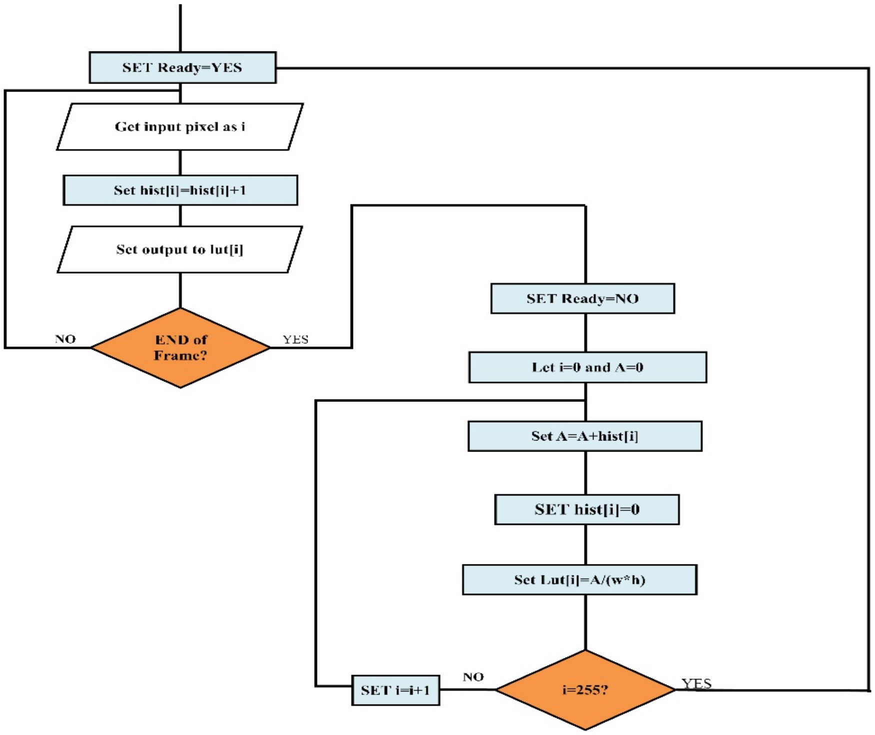


Fig. 5. Histogram Equalization Module Flowchart.

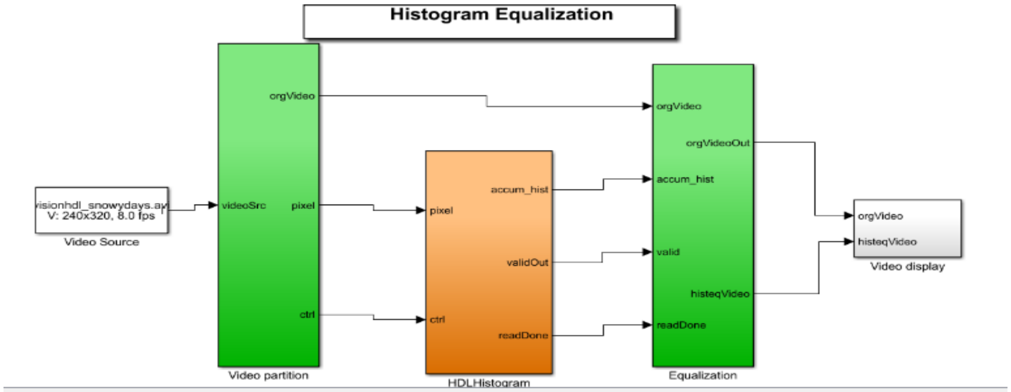


Fig. 6. Model based Design of Histogram Equalization using HDL Coder.

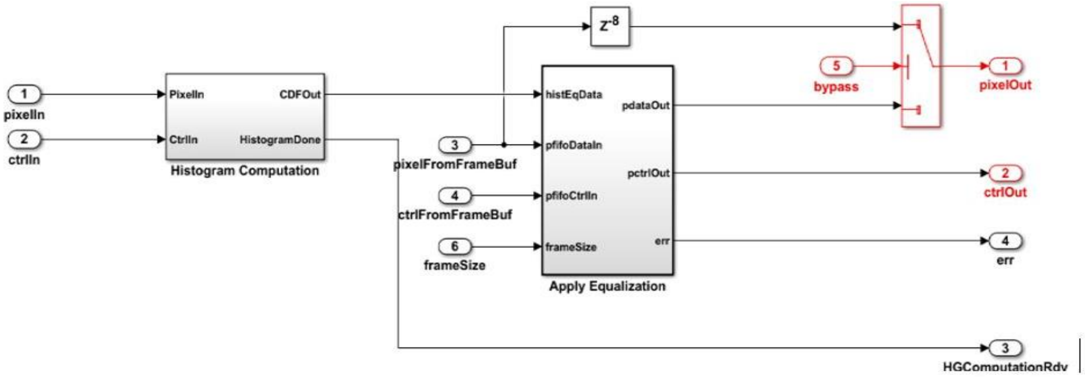


Fig. 7. HDL Histogram Equalization Subsystem.

TABLE. I. UTILIZATION OF THE AVAILABLE RESOURCES IN THE ZYNQ XC7Z020CLG484-2 PART

Bit Depth	8	
Channels	1	
LUT-FF Pairs	3740	7%
LUTs as Logic	2770	5%
LUTs as Memory	289	1.66%
Slice Registers	3350	3%
RAM 36/18	0.5	0.36%
DSP48	0	0.00%

TABLE. II. UTILIZATION AND MAXIMUM FREQUENCY FOR THE HISTOGRAM EQUALIZATION MODULE

Max frequency: 170 MHz		
Resolution	***Pixel Per Frame***	***Maximum Frame Rate***
1920x1080	2073600	82.1FPS
1440x900	1296000	129.6 FPS
1024x1024	1048576	159.7 FPS
1280x720	921600	181.0 FPS
1024x768	786432	211.5 FPS
640x480	307200	536.9 FPS
512x512	262144	628.6 FPS

B. *Color Conversion System*

Color conversion converts the raw image having colors belonging to the color space of the sensor into values in a standard color space independent of the sensor. The RGB color space is the standard widely adopted by the image and video processing system. Hence the interest of making the conversion directly to this color space. The conversion is performed using a standard method which is the use of a 3x3 conversion matrix.

For the forward conversion module, the conversion module uses the following matrix conversion:

$$Y = \frac{54.4256}{256} \cdot R + \frac{183.0912}{256} \cdot G + \frac{18.4832}{256} \cdot BCb$$

$$= -\frac{29.3305}{256} \cdot R - \frac{98.6695}{256} \cdot G + \frac{128}{256} \cdot B + 128Cr$$

$$= \frac{128}{256} \cdot R - \frac{116.2631}{256} \cdot G - \frac{11.7369}{256} \cdot B + 128$$

The following equations present the backward conversion from YCbCr space to RGB space:

$$R = Y + \frac{403.1488}{256} \cdot Cr - 201.5744G$$

$$= Y - \frac{47.9550}{256} \cdot Cb - \frac{119.8398}{256} \cdot Cr + 83.8974B$$

$$= Y + \frac{475.0336}{256} \cdot Cb - 237.5168$$

1) Simulink HDL coder model: Color image processing is a logical extension to the processing of grayscale images. The essential difference involves the fact that each pixel is composed of a vector of components rather than a scalar. Usually, a pixel from an image has three parts: red, green and blue. These are defined by the human visual system. A three-dimensional vector and user mainly present color can determine how many bits each component have.

The pre-defined video reference design, which contains other IPs to handle the HDMI input and output interfaces. The HDL DUT IP processes a video stream coming from the HDMI input IP, generates an output video stream and sends it to the HDMI output IP. All of these video streams are transferred to AXI4-Stream Video interface. The HDL DUT IP can also include an AXI4-Lite interface for parameter tuning. Compared to the AXI4-Lite interface, the AXI4-Stream Video interface transfers data much faster, making it more suitable for the data path of the video algorithm.

The color conversion system IP core was established in Matlab Simulink environment with HDL coder. This tool allows the user to drive the Zynq programmable logic (PL) part at a high level without having to deal with low-level hardware details. The proposed architecture is shown in Fig. 8.

Fig. I gives an overview of the entire color system conversion Simulink system. The other blocks are identical with those in color correction model except the RGB2YCbCr kernel block. The input bus signal is represented in Xilinx video data format, which is 32'hFFRRBBGG. So, we first separate the bus signal to three color components RGB. Red component is bit 23 to 16; Green is bit 7 to 0; Blue is bit 15 to 8. The gain blocks implement multiplication. The sum blocks calculate add and subtract result, which is defined in block parameter. Matrix multiplication is performed using these gain and sum blocks. Finally, again to a bus signal. The delay blocks inserted in between will be transferred to registers in hardware, which break down the critical path to achieving higher clock frequency.

2) Synthesis and FPGA implementation: A block design for color space conversion that implements an entire image processing system can be created.

The vivado project for the video processing system is generated by a proposed worksflow based on model-based design (MBD). This project can be opened with Xilinx Vivado version 2017.4.

Fig. 10 presents a full diagram for color conversion system. This diagram is composed by many blocks such as RGBtoYCbCr conversion IP, YCbCrtoRGB conversion IP. Also, the video processing HDMI input output, the Zynq's CPU cores, the Video DMA engines, and all the supporting blocks.

We validate the entire color conversion system design on a Zynq FPGA platform using generated HDL code. The logic resources utilized by design with timing performance are presented in Table III. The reported maximum frequency is 302 MHz for RGB to YcbCr and 260 MHz for YCbCr to RGB. The resources utilization of the YCbCr to RGB system on Zynq xc7z020clg484-2 FPGA is as follows: 1850 slice registers and 2280 slice LUTs.

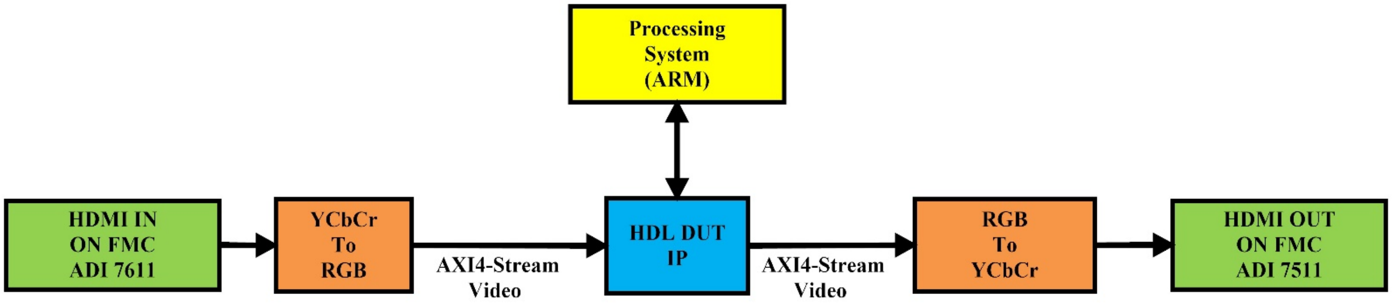


Fig. 8. AXI4-Stream Video Interface in Zynq 7000.

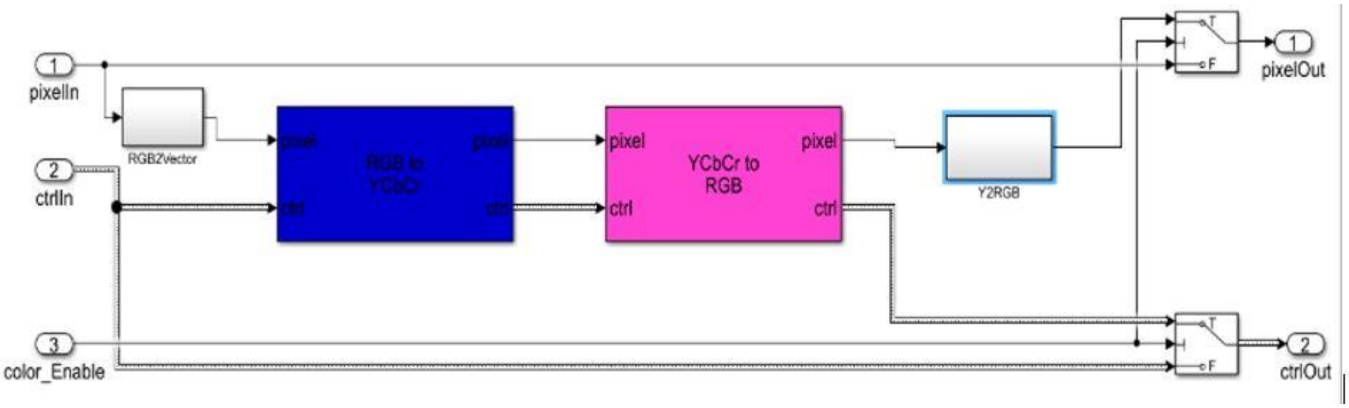


Fig. 9. Hardware Prototype for Color System Conversion.

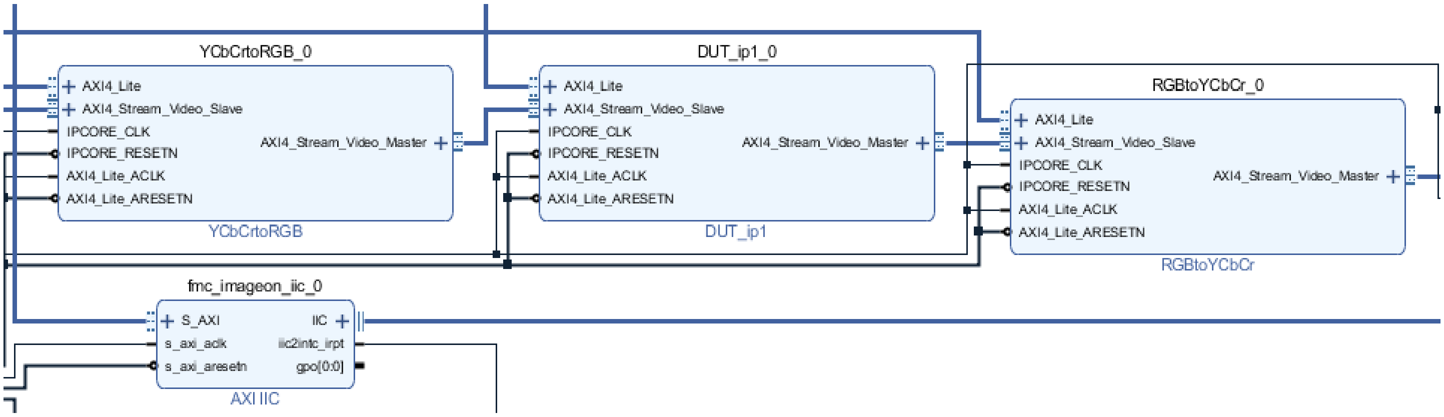


Fig. 10. Vivado IP Integrator with Several IP-Blocks for Color System Conversion.

TABLE. III. UTILIZATION OF AVAILABLE RESOURCES IN THE ZYNQ XC7Z020CLG484-2 PART

	YCbCr to RGB		**RGB to YCbCr**	
Bit Depth	8		8	
Channels	3		3	
LUT-FF Pairs	2280	4.20 %	4400	8.1 %
LUTs as Logic	1938	3.64%	3864	7.26%
LUTs as Memory	6	0.03%	4	0.02%
Slice Registers	1850	1.50%	3580	1.4%
RAM 36/18	0	0.00%	0	0.00%
DSP48	0	0.00%	0	0.00%

Table IV illustrates the result of our YCbCr to RGB system on hardware for color system conversion. Table IV present the resource usage and maximum frequency.

The resources utilization of the RGB to YCbCr system on Zynq xc7z020clg484-2 FPGA is as follows: 3602 slice registers and 4400 slice LUTs.

TABLE. IV. UTILIZATION AND MAXIMUM FREQUENCY COLOR SPACE CONVERSION

YCbCr to RGB			**RGB to YCbCr**	
Max frequency: 302 MHz			***Max frequency: 260 MHz***	
Resolution	***Pixel Per Frame***	***Maximum Frame Rate FPS***	***Pixel Per Frame***	***Maximum Frame Rate FPS***
1920x1080	2073600	143.1	2073600	125.5
1440x900	1296000	228.2	1296000	197.6
1024x1024	1048576	282.5	1048576	242.5
1280x720	921600	321.3	921600	376.2
1024x768	786432	373.4	786432	324.1
640x480	307200	955.5	307200	823.1
512x512	262144	1.125.6	262144	963.2

V. DISCUSSION AND CONCLUSION

The current paper suggests a VHLS method for image processing designs. This method gives a high abstraction level environment to the users, which can improve the development productivity by automating the MATLAB/Simulink-to-RTL synthesis process. We prototyped the suggested method by utilizing recently available Model-Based Design tools based on Simulink /HDL coder modeling. This method is then verified within two real-life applications. Experiments show that it can lead to the benefits of FPGA related to the tools of other kinds in the same abstraction level. It is worth noting that the findings of the study show that it will decrease the complexity of the algorithm behaviors depicted using MATLAB in routine level. This study also demonstrates the usefulness of heterogeneous Zynq SoC to establish an embedded vision system for a smart camera dedicated to traffic surveillance.

Resolving the following issues can facilitate reaching this goal; one of the needs of hardware-software architecture is to monitor the mastery of the HDMI video signal to the system the AXI bus-based communication between the FPGA and ARM processor. It shows that the suggested flow reduces FPGA prototyping time by up to 60% with MATLAB and HDL Coder. MATLAB and HDL Coder are used to eliminate the step of translating the initial algorithm to HDL by hand. HDL coder facilitates the improvements completed in hours, not weeks.

REFERENCES

[1] H. M. Abdelgawas, M. Safar, A. M. Wahba, “High Level Synthesis of Canny Edge Detection Algorithm on Zynq Platform,” International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:9, No:1, 2015, pp. 148-152.

[2] J. Monson, M. Wirthlin and B. L. Hutchings, "Implementing high-performance, lowpower FPGA-based optical flow accelerators in C," Application-Specific Systems, Architectures and Processors (ASAP), 2013 IEEE 24th International Conference on, Washington, DC, 2013, pp. 363-369.

[3] Ben Hamida, A., Koubaa, M., Nicolas, H., Amar, C.B.: Video surveillance system based on a scalable application-oriented architecture. Multimedia. Tools Appl. pp. 1–27 (2015). doi: 10.1007/s11042-015-2987-5.

[4] Fleck, S., Strasser, W.: Smart camera-based monitoring system and its application to assisted living. Proc. IEEE 96(10), 1698–1714 (2008). doi: 10.1109/JPROC.2008.928765.

[5] Huang, D.Y., Chen, C.H., Chen, T.Y., Hu, W.C., Chen, B.C.: Rapid detection of camera tampering and abnormal disturbance for video surveillance system. J. Vis. Commun. Image R. 25(2), 1865–1877 (2014).

[6] Chao Li, Yanjing Bi., Benezeth, Franck Marzani, Fan Yang: Fast FPGA prototyping for real-time image processing with very high-level synthesis. J. Real-Time Image Process. (2017). doi: 10.1007/s11554-017-0688-1.

[7] Henning Sahlbach, Daniel Thiele, Rolf Ernst: A system-level FPGA design methodology for video applications with weakly-programmable hardware components. J. Real-Time Image Process. (2017). doi: 10.1007/s11554-014-0403-4.

[8] Baklouti, M., Aydi, Y., Marquet, P., Dekeyser, J., Abid, M.: Scalablempnoc for massively parallel systems—design and implementation on FPGA. J. Syst. Archit. 56(7), 278 – 292 (2010). doi:10.1016/j.sysarc.2010.04.001. Special Issue on HW/ SW Co-Design: Systems and Networks on Chip.

[9] T. Han, G. W. Liu, H. Cai and B. Wang, "The face detection and location system based on Zynq," Fuzzy Systems and Knowledge Discovery (FSKD), 2014 11th International Conference on, Xiamen, 2014, pp. 835-839.

[10] P. K. Dash, S. S. Pujari and S. Nayak, "Implementation of edge detection using FPGA and Model-based approach", Proceedings of 2014 International Information Communication and Embedded Systems (ICICES), IEEE, (2014), pp. 1- 6.

[11] Sukhwani, B., Thoennes, M., Min, H., Dube, P., Brezzo, B., Asaad, S., Dillenberger, D.: A hardware/software approach for data base query acceleration with FPGAs. Int. J. Parallel Prog. 43(6), 1129–1159 (2015). doi:10.1007/s10766-014-0327-4.

[12] Jiang, J., Liu, C., Ling, S.: An FPGA implementation for real time edge detection. J. Real-Time Image Process. (2015). doi:10. 1007/s11554-015-0521-7.

[13] S. Sanchez-Solano, M. BroxJimenez, E. delToro, P. BroxJimenez and I. Baturone, "Model-based design methodology for rapid development of fuzzy controllers on FPGAs", IEEE Trans. Ind. Informat, vol. 9, no. 3, (2013), pp. 1361-1370.

[14] The MathWorksInc, (2014, May 17), "Optimizing HDL Code" [Online], Available:http:// http://www.mathworks.se/products/hdl-coder/description3 .html.

[15] The MathWorksInc, (2014, May 17), "Automating FPGA Design" [Online], Available:http:// www.mathworks.se/products/hdl-coder/description4.html.

[16] Bailey, D.G.: Design for Embedded Image Processing on FPGAs. Wiley (Asia) Pte Ltd, Singapore (2011).

[17] The MathWorksInc, (2014, May 18), "GenerarateVerilog and VHDL code for FPGA and ASIC designs." [Online], Available: http:// www.mathworks.se/products/hdlcoder/.

[18] The MathWorksInc, (2014, May 17). "HDL Coding standards" [Online], Available:http:// www.mathworks.se/products/hdl-coder/description7. html.

[19] The MathWorksInc, (2014, May 17), "Generating HDL Code" [Online], Available:http:// www.mathworks.se/products/hdl-coder/description2. Html.

[20] Kyo, S., Okazaki, S.: IMAPCAR: a 100 GOPS in-vehicle vision processor based on 128 ring connected four-way VLIW processing elements. J. Signal Process. Syst. 62, 5–16 (2011).

[21] Leu, A., Aiteanu, D., Graser, A.: A novel stereo camera-based collision warning system for automotive applications. In: IEEE International Symposium on AppliedComputational Intelligence and Informatics (SACI), pp. 409–414 (2011).

[22] Stein, G.P., Rushinek, E., Hayun, G., Shashua, A.: A computer vision system on a chip: a case study from the automotive domain. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPRW) (2005).

[23] Crockett, L.H.; Elliot, R.A.; Enderwitz, M.A.; Stewart, R.W. The Zynq Book: Embedded Processing with the Arm Cortex-A9 on the Xilinx Zynq-7000 All Programmable Soc; Strathclyde Academic Media: Glasgow, UK, 2014.

[24] Xilinx: Introduction to FPGA design with vivado high-level synthesis. Tech. Rep. UG998 (v1.0), Xilinx (2013).

[25] Coussy, P., and Morawiec, A.: 'High-Level Synthesis: from Algorithm to Digital Circuits', Berlin: Springer Science + Business Media, chapters 1, 4, 2008.

Design and Analysis of an Arithmetic and Logic Unit using Single Electron Transistor

Aarthy M[1], Sriadibhatla Sridevi[2]
Department of Micro and Nano Electronics
Vellore Institute of Technology, Vellore, India

***Abstract*—The demand for low power dissipation and increasing speed elicits numerous research efforts in the field of nano CMOS technology. The Arithmetic Logic Unit is the core of any central processing unit. In this paper, we designed a 4-bit Arithmetic and Logic Unit (ALU) using Single Electron Transistor (SET). Single-electron transistor (SET) is a new type of switching nanodevice that uses controlled single-electron tunneling to amplify the current. The single-electron transistor (SET) is highly scalable and possesses ultra-low power consumption when compared to conventional semiconductor devices. Reversible logic gates designed using SET are used for performing 4-bit arithmetic operations. We modelled symmetric single gate SET operating at room temperature using Verilog A code. The design is carried out in cadence simulation environment. The 4-bit SET based ALU design exhibits the power of 0.52 nW and delay of 350pS.**

Keywords—Single electron transistor; reversible logic gates; low power; speed

I. Introduction

Miniaturization has brought electronic devices close to the size where quantum phenomena play a significant role in altering the whole device properties. Nano-scale devices like Single Electron Transistor (SET), Resonance Tunneling Diode (RTD), Quantum Cellular Automata (QCA) and Carbon Nano Tube (CNT) can often perform the same tasks similar to microscale devices such as Field Effect Transistors, yet the working principles are different. The single-electron transistor (SET) is one of the fascinating nanodevices. SET can perform as a switch and exploits the quantum mechanical phenomenon of electron tunneling through the tunnel junctions and also control the transport of single electrons [1,2]. The different models used for simulation of SET are as follows: Uchida's model [3] is more feasible at higher temperatures and over a huge range of drain voltages. The MIB model [4] is more flexible than Uchida's model and fewer exponential terms exist, so simulation time is lesser. Inokawa's model [5] is an extension of Uchida's model for asymmetric SETs in which the source and drain resistance, capacitance are unequal. A new computation model has been proposed to perform the arithmetic functions by controlling the movement of electrons within the circuit. Using this model, they designed 4-bit Digital to Analog Converter, adder, and multiplier circuits. The quantitative and qualitative comparison in terms of delay, sensitivity to process variations by varying the SET parameters, temperature, bias current and drain voltage of SET based circuits had been analyzed using SPICE, SIMON simulator. A novel quasi analytical model has been developed and validated for single-electron transistors based logic circuit simulation in static and dynamic regimes as well as for hybrid-SET by combining SET with MOSFET had been reported [6-15]. Sharifi.M.J et al. [18] proposed speed enhancement and bit error rate reduction in SET based digital circuits by reducing tunneling wait time.

An ALU is a combinational circuit that can perform a set of basic arithmetic and logical operations. Modern central processing unit (CPUs) and graphics processing units (GPUs) contain very dynamic and complex ALUs acting as the fundamental building block. Several research works had been proposed on reversible gate based ALU design. Reversible Logic plays a major role in fields such as Nanotechnology, low power CMOS design, optical computing, low loss computing. Limited research work was carried on SET based arithmetic circuits. In this paper, we used reversible gates using SET based on the MIB model for performing arithmetic operations. The Feynman gate is the most basic reversible gate. It is the only 2x2 reversible gate mainly used for fan-out purposes. The Toffoli gate, Fredkin gate, New gate and Peres gate are 3x3 reversible gates that can be used to realize various Boolean functions. The TSG gate, MKG gate, HNG gate, PFAG gate are 4x4 reversible gates that are designed to implement reversible adders. Slimani Ayyoub et al. [16] designed ALU using double Peres gate reversible logic to reduce quantum cost, the number of garbage outputs and depth of the circuits. Bolhassani, A et al. [17] developed new reversible ALU using elementary quantum gates which can be used in the implementation of Quantum computers. Vandana Shukla et al. [19,22] proposed a novel design approach for a 2-bit ALU design using 8:1 MUX with the reversible logic and simulated using Modelsim tool. Shahram Babaie et al. [20] designed Quantum cellular automata multilayer ALU to perform arithmetic and logical operation using QCA designer tool. Aarthy et al. [21] developed binary multiplier using single gate SET, double gate SET and hybrid SET by combining SET with MOSFET and analyzed the performance in terms of area, power and delay. The research paper is organized as follows: Section II describes the working principle and characteristics of SET. Section III describes the working model of SET based 4-bit ALU. The outcome of the proposed model has been elaborated in Section IV. Finally, the paper concludes in Section V.

II. Set Framework

The SET consists of a metallic island, placed between the source and drain tunnel junctions and has a gate electrode similar to regular FET. The tunnel junctions are thin (<10 nm) oxide layer between the island and the gate electrodes. Quantum dots have been used as islands for the SET. Both tunnel junctions in the SET have intrinsic tunnel resistance (R_{ts} and R_{td}) and capacitance (C_{ts} and C_{td}) parallel to each other. The SET schematic is shown in Fig. 1.

The SET island is very small in the nanometric scale accommodating a vast number of electrons. We can add or subtract electrons from the island, either by charging it positively or negatively based on the electron tunneling. The number of excess electrons on the island is denoted as n. The value of n can also be negative, which means that electrons have been removed from the island, leaving a positive charge. The electrostatic energy of the system is affected by the presence of excess electrons, which depends on the charge of the island.

$$E_C = \frac{1}{2}\frac{Q_{island}^2}{C_\Sigma} = \frac{1}{2}\frac{n^2e^2}{C_\Sigma} \quad (1)$$

Where Q_{island} is the charge on the island, n the number of excess electrons, e is the charge of an electron and C_Σ the total capacitance which is equal to

$$C_\Sigma = C_{g1} + C_{g2} + C_{ts} + C_{td} \quad (2)$$

where C_{g1} , C_{g2} are gate capacitance and C_{ts} , C_{td} are the intrinsic source and drain tunnel junction capacitance.

The electrostatic energy of the system becomes

$$E_{electrostatic} = \frac{1}{2}\frac{Q^2}{C_\Sigma} = \frac{1}{2}\frac{(ne-Q_g)^2}{C_\Sigma} = \frac{1}{2}\frac{(ne-V_gC_g)^2}{C_\Sigma} \quad (3)$$

Where Q_g is the gate charge. This energy determines if electron tunneling through a junction is restricted or allowed. The addition of an excess electron on the island increases the energy of the system, then electron tunneling will be energetically prohibited, so no tunneling occurs through the junction known as the Coulomb blockade. The drain-source potential V_{ds} determines the energy of the electrons before the junction. Only if this energy is greater than the Coulomb blockade, the electrons will overcome the blockade and tunneling will occur. Mainly, the Coulomb blockade is based on the number of excess electrons on the island (n) and the gate charge (C_g).

A. Parameters that Improve the Performance of SET

The single-electron transistor performance is not merely determined by source voltage (V_s), drain voltage (V_d) and gate voltage (V_g) but also by other parameters such as external charges and temperature. Increasing the charging energy ($\frac{e^2}{2C_\Sigma}$) in the SET will provide the possibility of SET to operate at high temperatures, which is obtained by reducing the device capacitance to a very small value (in the order of 10^{-18} F) since the electrostatic energy is inversely proportional to it ($E_C = \frac{Q^2}{C_\Sigma}$). The presence of charges that are not on the SET island but nearby referred to as external charges' is one more important parameter that can cause a severe problem like an uncontrolled drift of threshold voltage of the transistors.

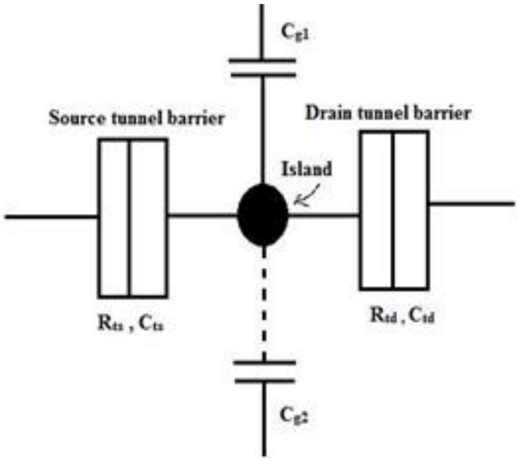


Fig. 1. SET Schematic.

The switching time is essential to operate SET as a switch. The switching time τ_T is based on the total device capacitance C_Σ and the tunneling resistance R_T.

$$\tau_T = R_T C_\Sigma \quad (4)$$

The switching time is short due to the Heisenberg principle that depends on the charging energy of the device E_C,

$$E_C \tau_T \geq h \quad (5)$$

The higher-order tunneling processes (like co-tunneling) can be stopped by higher tunneling resistance so-called von-Klitzing resistance.

$$R_T \gg \frac{2h}{Q^2} = 26K\Omega \quad (6)$$

B. SET Tunneling Mechanism

The tunnel junction in SET acts as opaque capacitors when no electron tunnel through them. Before any electron tunneling, the potential of the island is expressed as

$$V_{island} = \frac{C_g}{C_\Sigma}V_{gs} + \frac{C_{td}}{C_\Sigma}V_{ds} \quad (7)$$

The electron tunneling can take place only if $|V_{island}| > \frac{e}{2C_\Sigma}$ through the source tunnel barrier or if $|V_{ds} - V_{island}| > \frac{e}{2C_\Sigma}$ through the drain tunnel barrier. To understand the mechanism let us keep V_{ds} constant at $\frac{e}{2C_\Sigma}$ also, vary V_{gs} from zero to any higher positive value, we can observe the following mechanism represented in Fig. 2.

The perpendicular lines indicate the drain (D), island (I) and source (S) terminals and the parallel lines represent the corresponding voltages. The source (V_s) is grounded and the drain is connected to ∝ (V_D). The black dots denote the potential of the island before electron tunneling takes place. The white dots denote the potential of the island after electron tunneling. The solid arrows denote the electron tunneling and dotted arrows denote the changes in island potential. The numbers (1,2,3,...) denote the total current conduction sequences.

- When $V_{\mathrm{island}} < \frac{e}{2C_{\Sigma}}$ the potential drop across both source and drain tunnel junction is less than $\propto = \frac{e}{2C_{\Sigma}}$, so the device enters the Coulomb blockade, which is highlighted in Fig. 2(a).
- If V_{gs} is increased further (higher than $\propto$) then $V_{\mathrm{island}} > \frac{e}{2C_{\Sigma}}$, which allows one electron to tunnel in from source terminal to island. As a result island potential is reduced by the amount of $\frac{e}{C_{\Sigma}}$ consequently drain tunnel junction potential becomes higher than $\frac{e}{2C_{\Sigma}}$,which allows one electron to tunnel out from the island to the drain terminal that is highlighted in Fig. 2(b).
- If V_{gs} is increased further (higher than $2\propto$), once again the SET enters coulomb blockade. Initially when $V_{\mathrm{island}} > \frac{e}{2C_{\Sigma}}$ one-electron tunnels in from a source to the island which reduces the island potential by $\frac{e}{C_{\Sigma}}$. As a result, the potential drop across both source and drain junction becomes lower than $\frac{e}{2C_{\Sigma}}$ and that is highlighted in Fig. 2(c).
- If V_{gs} is increased further (higher than $3\propto$), as shown in Fig. 2(d) the SET comes out of Coulomb blockade as $V_{\mathrm{island}} > \frac{3e}{2C_{\Sigma}}$. When$V_{\mathrm{island}} > \frac{e}{2C_{\Sigma}}$, one-electron tunnels in from a source to the island that reduces the island potential by $\frac{e}{C_{\Sigma}}$. Since V_{island} is still higher than $\frac{e}{2C_{\Sigma}}$ one more electron can tunnel in from source to island hence the island potential reduces by the amount of $\frac{e}{C_{\Sigma}}$. The tunneling mechanism in step b will be resumed (5-->6-->3-->4-->5-->...) and a continuous current path from source to drain is re-established.

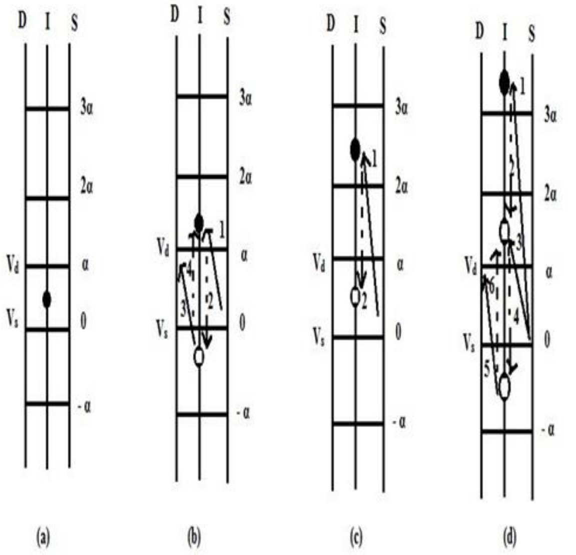


Fig. 2. Electron Tunneling Mechanism in a SET System (a-d).

III. Proposed Work

The important component of the central processing unit of a computer is an ALU, which performs arithmetic and logical operation. We proposed 4-bit SET based ALU which consists of SET based adder block, subtractor block, logical block, shifter block and a quadruple multiplexer (MUX). We designed a quadruple 4:1 MUX to select specific operation in ALU by using control signals S0 S1. We used reversible gates to perform the arithmetic operation to reduce power consumption. The proposed model of 4-bit ALU is shown in Fig. 3. Table I shows the operation table of 4-bit ALU.

We have designed the 4-bit SET based ALU which has the following features:

- Operates at a low voltage of 0.4V.
- Consumes low-power.
- Operates at high speed.
- Performs both arithmetic and logical operations.

A. SET based AND Gate

The AND gate design using single gate SET is presented in Fig. 4. The design consists of six SETs where four SETs for NAND gate design and two SETs for Inverter. Similar to static CMOS structure, the pull-up network and pull-down network are dual to each other for SET based logic gate design also. For each SET, the input is applied to gate1 and gate2 is grounded. The two gate capacitances C_{g1} and C_{g2}, drain tunnel junction capacitance C_{td} and source tunnel junction capacitance C_{ts}. We modelled symmetric SET by operating at room temperature by increasing the charging energy $E_C = \frac{e^2}{2C_{\Sigma}}$ which is achieved by lowering the gate capacitance and tunnel junction capacitance less than 1aF. So by increasing the temperature, the capacitance size reduces thereby feature size of island scaled-down, which narrows down the coulomb blockade region. The single gate controls the Coulomb blockade region, which regulates electron tunneling from source to drain terminal. The simulation parameters are C_{g1}=0.23aF, C_{g2}=0, C_{td}= C_{ts}=0.06aF, R_{ts}= R_{td}=1MΩ.

In a similar fashion, all logic gates have been designed using SET to design the arithmetic, logical and shifter blocks of ALU.

B. SET based Reversible Adder/Subtractor

In conventional gates, the inputs cannot be originated from the outputs, so there will be loss of one or more bit information which is dissipated as heat. Using reversible logic, the inputs can be retrieved from the outputs and vice-versa by which energy loss can be vanquished. Reversible logic gates commenced as a promising calibrating model for low power applications, quantum computing, quantum cellular automata, DNA computing and nanotechnology. To maintain the reversibility of the digital circuits, the reversible logic gate uses extra outputs known as garbage outputs.

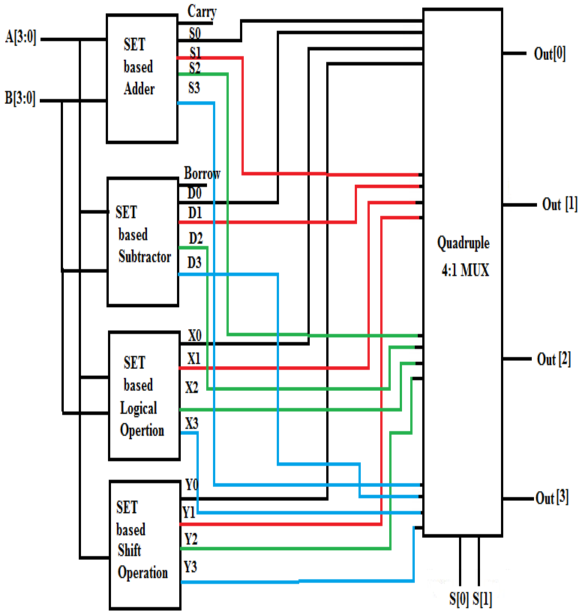


Fig. 3. SET based 4-Bit ALU Block Diagram.

TABLE. I. OPERATION TABLE OF 4-BIT ALU

S0	S1	Operation
0	0	4-bit addition
0	1	4-bit subtraction
1	0	Logical operation
1	1	Shift operation

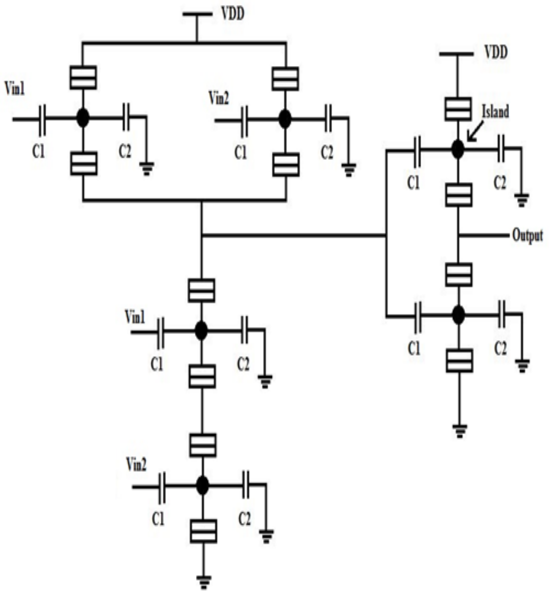


Fig. 4. SET based 2-Bit AND Gate.

In this paper, full adder and subtractor are designed using a reversible WG gate implemented using a single-electron transistor. The 4X4 WG gate design includes three inputs A, B, C and D is the control input, U, V, W and X act as outputs. By setting D=0, the circuit performs addition operation and if D=1 it performs subtraction operation. W acts as sum and X acts as carry in case of WG gate as a full adder. W acts as difference and X acts as borrow in case of WG gate as a full subtractor. The block diagram of reversible gate as full adder and subtractor are shown in Fig. 5, 6 and 7, respectively.

The circuit is simulated by using Cadence Virtuoso tool. Fig. 8 and Fig. 9 show the simulation result of SET based WG reversible full adder and full subtractor respectively. Fig. 8 shows that when the input is A=B=C=1, the circuit produces the Sum=1 Carry=1. From Fig. 9, it is analysed that when the input is A= 0 B= 1 C=1, the circuit produces the Difference= 0 Borrow=1.

C. 4-bit SET based Reversible Adder/Subtractor

We designed 4-bit reversible adder and subtractor using SET in which the SET based EXOR gate plays a significant role in determining which operation to be performed. Fig. 10 shows the block diagram of SET based reversible 4-bit adder and subtractor. When the control signal applied to the EXOR gate is set to one, it performs subtraction operation else addition operation. The circuit is simulated using 608 SETs to verify the functionality. The simulation results in Fig. 11 reveals that when the inputs A3=1 A2=0 A1=1 A0=0 and B3=0 B2=1 B1=0 B0=1 and when control is 0 the circuit performs addition operation exhibiting output as C4=1 S3=0 S2=0 S1=0 S0=0 else when control is 1 it performs subtraction operation displaying output as 10101.

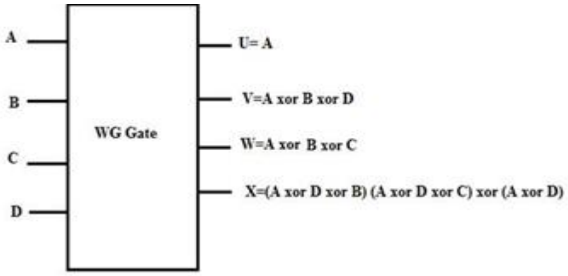


Fig. 5. WG Reversible Gate.

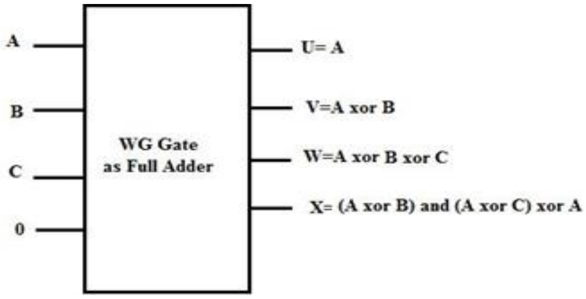


Fig. 6. WG Reversible Gate as Full Adder.

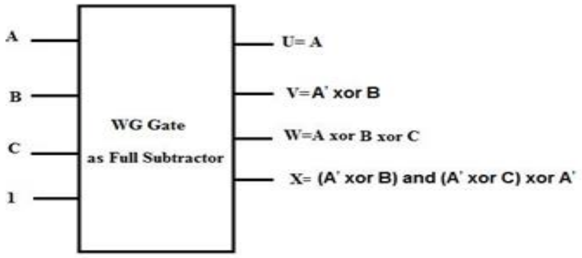

A
B
C
1
WG Gate as Full Subtractor
U= A
V=A' xor B
W=A xor B xor C
X= (A' xor B) and (A' xor C) xor A'

Fig. 7. WG Reversible Gate as Full Subtractor.

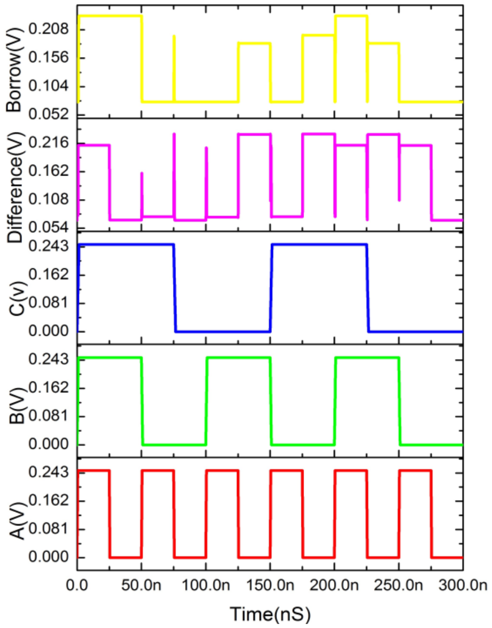


Fig. 8. Output Waveform of SET based WG Reversible Gate as Full Adder.

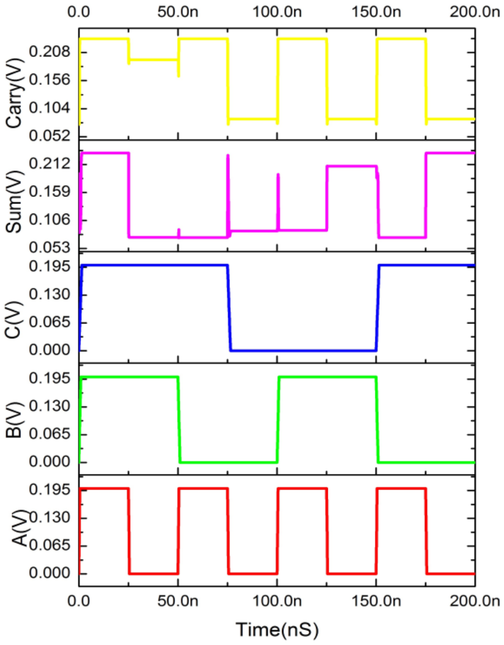


Fig. 9. Output Waveform of SET based WG Reversible Gate as Full Subtractor.

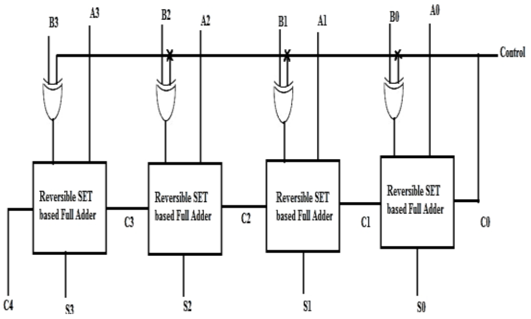

B3
A3
B2
A2
B1
A1
B0
A0
Control
Reversible SET based Full Adder
Reversible SET based Full Adder
Reversible SET based Full Adder
Reversible SET based Full Adder
C3
C2
C1
C0
C4
S3
S2
S1
S0

Fig. 10. 4-bit SET based Reversible Gate Full Adder/ Full Subtractor.

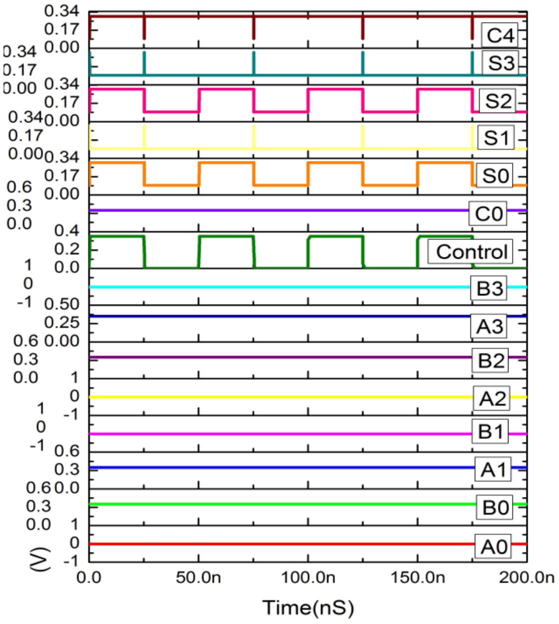


Fig. 11. Output Waveform of 4-bit SET based Reversible Gate Full Adder/ Full Subtractor.

D. 4:1 SET based Multiplexer

The 4:1 MUX is designed using SET based AND and OR gates to select the specific operation of the ALU based on the control signals S0, S1. The control signals, S0 and S1, are used to specify various actions, as given in Table II. Fig. 12 reveals the circuit is simulated using 54 SETs to verify the functionality. The block diagram of SET based 4:1 multiplexer is presented in Fig. 13.

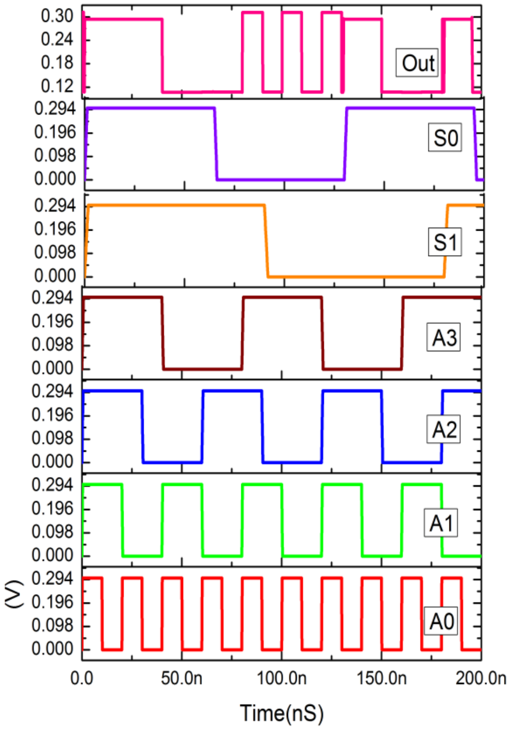


Fig. 12. Output Waveform of 4:1 SET based Multiplexer.

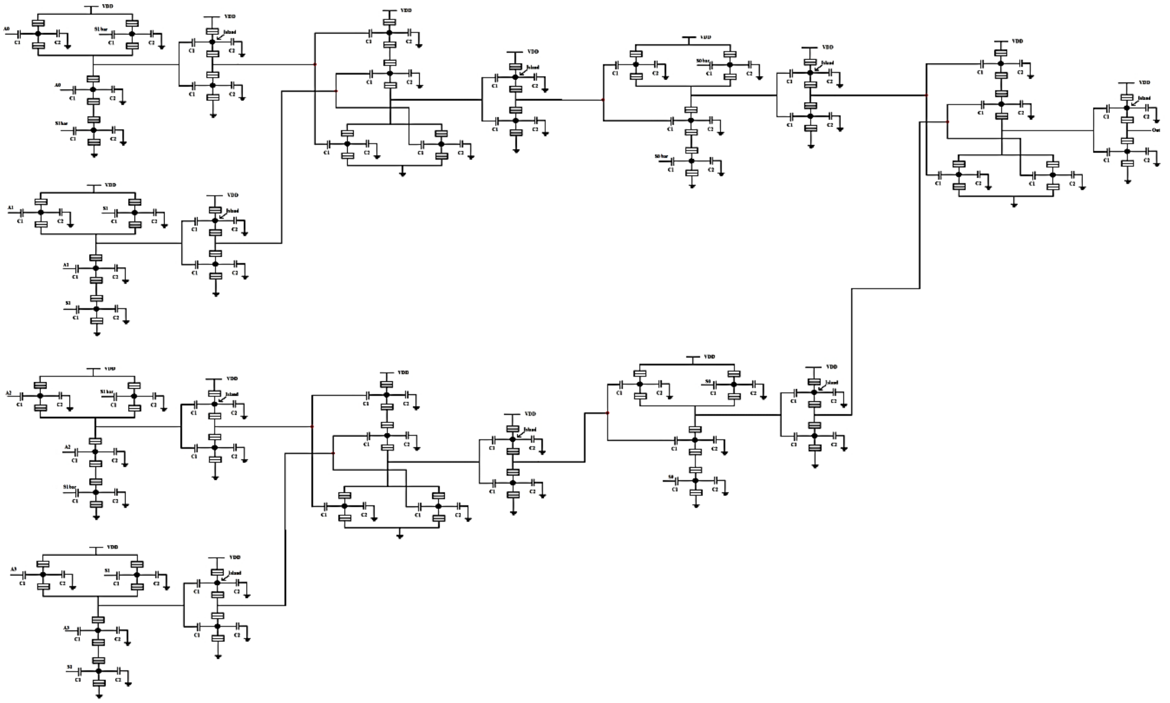


Fig. 13. 4:1 SET based Multiplexer.

TABLE. II. OPERATION TABLE OF 4:1 MULTIPLEXER

S0	S1	Out
0	0	A0
0	1	A1
1	0	A2
1	1	A3

E. 4-bit SET based Left/Right Shift Register

The shift registers are used for transfer or storage of binary data, which are generally used in computers or calculators to store binary data. A shifter is used to shift the data to the left or right side by a fixed number of positions. The vacant position is filled with zero. We designed a 4-bit logical shifter using SET based multiplexer. The selection signals, S0 and S1, are used to specify the various actions, as given in Table III. The block diagram of SET based Left/Right Shift register is shown in Fig. 14. We simulated the circuit using 204 SETs with a supply voltage of 400mV to verify the functionality. Fig. 15 shows that when the input A3=1 A2=1 A1=0 A0=1 and when the control signal S0S1=10, the input data is shifted right side providing the output as Out3 Out2 Out1 Out0 = 0110. When S0S1=11, it performs left shift operation providing the output as 1010.

F. 4-bit SET based Magnitude Comparator

A comparator compares two binary numbers of 4-bit size and generates three outputs, such as equal, greater and smaller. We have designed a magnitude comparator using SET based logic gates like INVERTER, EXNOR gate, AND gate and OR gate as shown in Fig. 16.

The condition of A>B in a 4-bit comparator can be possible in the following cases as shown in Fig. 17.

- If A3 B3 = 1 0
- If A3 B3= XX and A2 B2 = 1 0
- If A3 B3=XX, A2 B2 = XX and A1 B1 =1 0
- If A3 B3= XX, A2 B2= XX, A1 B1=XX and A0 B0 =1 0

In the same way A<B condition can be possible in the following cases:

- If A3 B3 = 0 1
- If A3 B3= XX and A2 B2 = 0 1
- If A3 B3=XX, A2 B2 = XX and A1 B1 =0 1
- If A3 B3= XX, A2 B2= XX, A1 B1=XX and A0 B0 =0 1

Where X can be either 0 or 1 treated as don’t care.

The A=B condition is applicable when all the independent bits match exactly with resemblant bits of other number. We simulated the circuit using 206 SETs to verify the functionality as shown in Fig. 18.

TABLE. III. LEFT / RIGHT SHIFT OPERATION TABLE

S0	S1	Operation	Out3	Out2	Out1	Out0
0	0	No change	A3	A2	A1	A0
0	1	No change	A3	A2	A1	A0
1	0	Right shift	0	A3	A2	A1
1	1	Left shift	A2	A1	A0	0

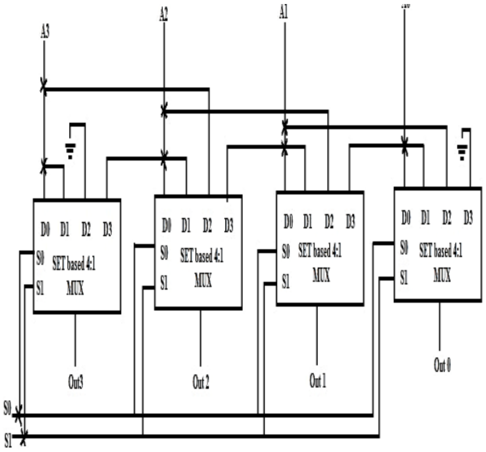


Fig. 14. 4-bit SET based Left/Right Shift Register.

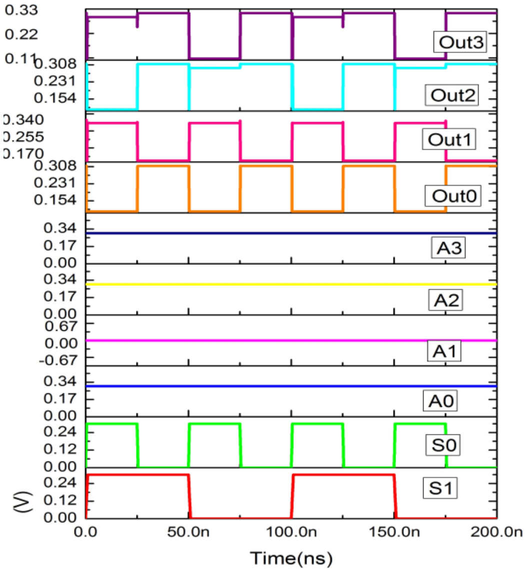


Fig. 15. Output Waveform of 4-bit SET based Left/Right Shift Register.

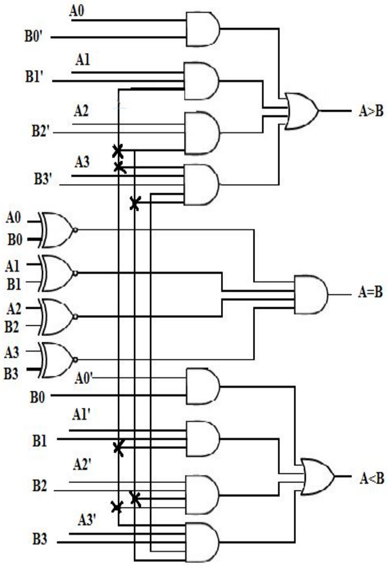
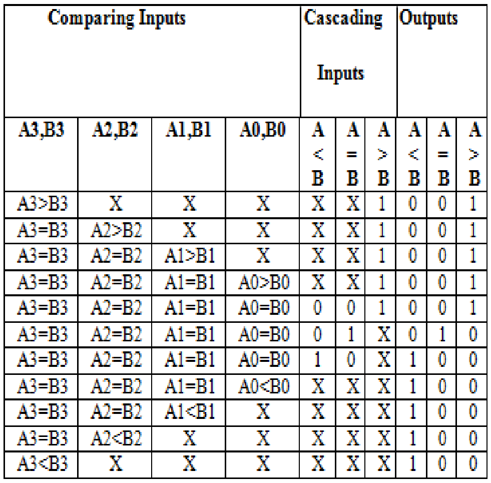


Fig. 16. 4-bit SET based Magnitude Comparator.



Comparing Inputs				Cascading Inputs			Outputs		
A3,B3	A2,B2	A1,B1	A0,B0	A < B	A = B	A > B	A < B	A = B	A > B
A3>B3	X	X	X	X	X	1	0	0	1
A3=B3	A2>B2	X	X	X	X	1	0	0	1
A3=B3	A2=B2	A1>B1	X	X	X	1	0	0	1
A3=B3	A2=B2	A1=B1	A0>B0	X	X	1	0	0	1
A3=B3	A2=B2	A1=B1	A0=B0	0	0	1	0	0	1
A3=B3	A2=B2	A1=B1	A0=B0	0	1	X	0	1	0
A3=B3	A2=B2	A1=B1	A0=B0	1	0	X	1	0	0
A3=B3	A2=B2	A1=B1	A0<B0	X	X	X	1	0	0
A3=B3	A2=B2	A1<B1	X	X	X	X	1	0	0
A3=B3	A2<B2	X	X	X	X	X	1	0	0
A3<B3	X	X	X	X	X	X	1	0	0

Fig. 17. 4-bit Magnitude Comparator Operation Table.

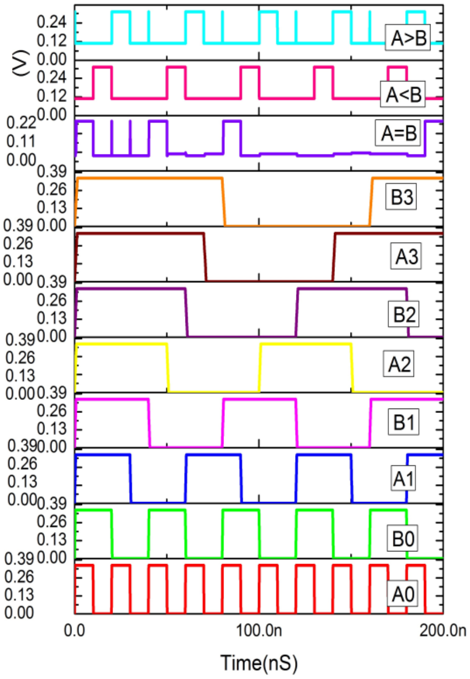


Fig. 18. Output Waveform of 4-bit SET based Magnitude Comparator.

IV. Results and Discussion

In this section, the functional simulation of SET based ALU is presented initially and later, the performance evaluation in terms of power and delay. We simulated the proposed ALU with 1554 SETs using MIB model. The control signals (S0, S1) are used to select one among various operations to determine the final output. Fig. 19 shows that when the inputs A3=1 A2=0 A1=1 A0=0 and B3=0 B2=1 B1=0 B0=1 and the control signal S0S1=00, the circuit performs addition operation providing output Carry=Out[3]= Out[2]= Out[1]= Out[0]=10000. When control signal S0S1=01, it acts as subtractor and provides output 10101. When the control signal S0S1=10, it is logical AND operation with output 0000. When control signal S0S1=11, the data is shifted left and the result is 0100. The performance evaluation of the proposed SET based ALU is shown in Table IV. The symmetric SET based 4-bit ALU operates at room temperature with a supply voltage of 400mV exhibits power of 0.52nW and propagation delay of 350pS.

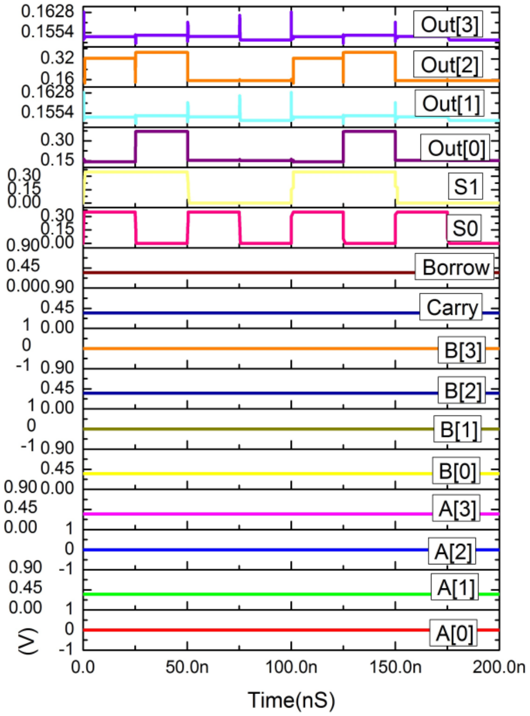


Fig. 19. Output Waveform of 4-bit SET based ALU.

TABLE. IV. PERFORMANCE ANALYSIS

Si.No	Parameters	Evaluated
1.	Number of SETs	
a.	4-Bit Adder	304
b.	4-Bit Subtractor	304
c.	4-Bit Left/Right Shifter	204
d.	4-Bit Logical Operator	24
e.	4-Bit ALU	1554
2.	Delay	350pS
3.	Power	0.52nW
4.	VDD	400mV

V. CONCLUSION

The nanodevices have unique properties such as small size and have the ability to operate at low voltage can be used for designing ultra-low-power digital circuits. Based on this property, we implemented logic circuits using the SET and developed 4-bit ALU. The proposed ALU can handle arithmetic and logical operations using two inputs of four-bit size and two control inputs to select a particular operation. The results show that the proposed ALU exhibits the power of 0.52nW and a delay of 350pS. The proposed ALU can be designed using double gate SET (DGSET) in which two gates control single electron tunneling which offers low power consumption. The proposed SET based ALUs can also be used in the implementation of quantum computers making significant improvements in the design of electronic circuits.

REFERENCES

[1] Taur, Y.; Buchanan, D.A.; Chen, W.; Frank, D.J.; Ismail, K.E.; Lo, S.-H.; Sai-Halasz, G.A.; Viswanathan, R.G.; Wann, H.-J.; Wind, S.J. , “CMOS scaling into the nanometer regime,” in Proc. IEEE 1997, 85, 486–504.

[2] Likharev, K.K , “Single-electron devices and their applications,” in Proc. IEEE 1999, 87, 606–632.

[3] Uchida, K.; Matsuzawa, K.; Koga, J.; Ohba, R.; Takagi, S.; Toriumi, A.; “Analytical single-electron transistor (SET) model for design and analysis of realistic SET circuits,” in Jpn. J. Appl. Phys. 2000, 39, 2321.

[4] Mahapatra, S.; Ionescu, A.M.; Banerjee, K. , “A quasi-analytical SET model for few-electron circuit simulation. ,” in IEEE Electron Device Lett. 2002, 23, 366–368.

[5] H. Inokawa and Y. Takahashi, ‘A compact analytical model for asymmetric single-electron tunneling transistors,” in IEEE Trans. Electron Devices. Vol. 50, no. 2, pp. 455–461, Feb. 2003.

[6] Cotofana, S.; Lageweg, C.; Vassiliadis, S., “Addition related arithmetic

[7] operations via controlled transport of charge,” in IEEE Trans. Comput. 2005, 54, 243–256.

[8] Wang, X.; Porod, W., “A Single-electron transistor analytic I–V model for SPICE simulations,” in Superlattices Microstruct. 2000,28,345–349.

[9] Cao, L.; Altomare, F.; Guo, H.; Feng, M.; Chang, A.M. , “Coulomb blockade correlations in a coupled single-electron device system,” in Solid State Commun. 2019, 296, 12–16.

[10] Dubuc, J.; Beauvais, J.; Drouin, D. , “Single-electron transistors with wide operating temperature range,” in Appl. Phys. Lett., vol. 90, no. 11,p. 113 104, Mar. 2007.

[11] Beaumont, A.; Dubuc, C.; Beauvais, J.; Droui, D.; , “Room Temperature Single-Electron Transistor Featuring Gate-Enhanced ON-State Current,” in IEEE Electron Device Letters vol. 30, no. 7, July 2009.

[12] TuckerJ.R., “Complementary digital logic based on the Coulomb blockade ,” in J. Appl. Phys., 1992, 72, (9), pp. 4399–4413.

[13] ShinS.J, LeeJ.J, KangH.J., “Room-temperature charge stability modulated by quantum effects in a nanoscale silicon Island ,” in Nano Lett., 2011, 11, (4), pp. 1591–1597.

[14] HuC, CotofanaS.D, Jiang.J, “Single-electron tunneling transistor implementation of periodic symmetric functions ,” in IEEE Trans. Circuits Syst. II, 2004, 51, (11), pp. 593–597.

[15] Miyaji.K, Saitoh.M, Hiramoto.T, “Compact analytical model for room temperature operating silicon single-electron transistors with discrete quantum levels ,” in IEEE Trans. Nanotechnol., 2006, 5,(3),pp.167–173.

[16] M. H. Sulieman and V. Beiu, “On single-electron technology full adders,” in IEEE Trans. Nanotechnol, vol.4,no.6,pp.669–680,Nov. 2005.

[17] Slimani, A.; Benslama, A, “Optimized 4-bit Quantum Reversible Arithmetic Logic Unit,” in International Journal of Theoretical Physics Aug2017, Vol. 56, Issue 8, p2686-2696.

[18] Bolhassani, A.; Haghparast, M, “Optimized designs of reversible arithmetic logic unit,” in Turkish Journal of Electrical Engineering and Computer Sciences Apr.2017, 25, 1137-1146.

[19] Sharifi, M.J; Ahmadian, M, “A Novel designs for digital gates based on single-electron devices to overcome the traditional limitation on speed and bit error rate ,” in Microelectronics Journal, 73 (2018) 12–17.

[20] Shukla, V.; Singh, O.P.; Mishra, G.R.; Tiwari R.K, “A Novel Approach to Design 2-bit Binary Arithmetic Logic Unit (ALU) Circuit Using Optimized 8:1 Multiplexer with Reversible logic, ” in Journal Of Communications Software and Systems, VOL. 11, NO. 2, June 2015.

[21] Babaie, S.; Sadoghifa, A.; Bahar, A.N, “Design of an Efficient Multilayer Arithmetic Logic Unit in Quantum-Dot Cellular Automata (QCA), ” in IEEE Transactions on Circuits and Systems—ii: express briefs, VOL. 66, No. 6, June 2019.

[22] Aarthy, M.; Sriadibhatla, Sridevi., “Design and Analysis of an Ultra Low Power Single Electron Transistor based Binary Multilpier,” in Journal of Advanced Research in Dynamical & Conttrol systems, VOL.11, No.8,2019.

[23] Weste, N.H.E., Harris, D, “CMOS VLSI Design, A Circuits And Systems Perspective, ” in Published by Person Education, Boston, Addison Wesley (2005) 3rd edn., pp. 715–738.

A Review and Development Methodology of a LightWeight Security Model for IoT-based Smart Devices

Mathuri Gurunathan[1], Moamin A. Mahmoud[2]
College of Computing and Informatics
Universiti Tenaga Nasional
Kajang, Malaysia

***Abstract*—Internet of Things (IoT) turns into another time of the Internet, which contains connected smart objects over the Internet. IoT has numerous applications, for example, smart city, smart home, smart grid and healthcare. In common, the IoT system comprises of heterogeneous devices that deliver then trade endless sums of safety-critical information, also as privacy-sensitive information. Nevertheless, connected devices can give your business a genuine lift, yet anything that is connected to the Internet can be vulnerable to cyberattacks. Most present IoT arrangements rely upon centralized architecture by associating with cloud servers through the Internet. The public cloud is described as computing services publicized by third-party suppliers over the Internet, making them accessible to anybody who needs to use or buy them. This solution gives magnificent flexible calculation and information the executives capacities, as IoT systems are developing increasingly mind-boggling; nonetheless, despite everything, it faces different of security issues. One of the weaknesses is that your information moving in IoT devices by means of public cloud could be in danger, despite the fact that the hacker was not explicitly focusing on you and with the public cloud you have insignificant authority over how rapidly you can grow the cloud. In this case, a secured protocol in IoT is vital to ensure optimum security to the information being traded between connected devices. To overcome the limitation, in this paper, we conduct a comprehensive review on existing security protocols and propose a development methodology of a blockchain-based lightweight security model that provides end to end security. By utilizing lightweight, an authenticated client can get to the information of IoT sensors remotely. The presentation investigation shows that lightweight offers better security, less overheads, and low communication.**

***Keywords*—Lightweight; security mode; IoT; smart devices**

I. Introduction

The Internet of Things (IoT) is a grouping of connected devices with the Internet. The selection of IoT-based advancements opens up new doors in different parts of our everyday lives, for example, smart home, smart transportation, and manufacturing [1]. Hence, IoT based applications become necessary things in our everyday life [2]. In a cloud-based IoT condition, the cloud stage is utilized to store the information got from the IoT sensors. Cloud innovations are pointed at provisioning steady and shared computational and storage assets to different clients and applications [3]. Public cloud is described as computing services publicized by third-party providers over the Internet, making them open to anyone who needs to use or get them [3] [4]. They may be free or sold on-demand, empowering customers to pay simply examine for the CPU cycles, the capacity they use or bandwidth [4]. Not at all like private clouds, public clouds can spare companies from the high costs of having to buy. Since certain applications in cloud-based IoT are critical bases, the data gathered and sent by IoT sensors must not be leaked during the user to device communication [5]. Be that as it may, public cloud is an increasingly attractive objective for hackers. Your information could be in danger, despite the fact that the hacker was not explicitly focusing on you. It has restricted adaptability in design and security and it isn't perfect for organizations or clients who utilize sensitive information [6]. In this manner, we require secure verification protocol for cloud-driven IoT-based applications in which a genuine client and an IoT sensor can commonly validate each other for secure communication yet in a recently published paper by Bogdan-Cosmin Chifor et.al [7] proposed a theft safe security scheme utilizing a keep-alive protocol that is executed intermittently and each time the client request a Fido verification via cloud platform. The Fido UAF model gives focal points over traditional verification mechanisms, such as strong authentication and a simplified registration and authentication method [8]. In any case, the Fido protocol too presents a few noteworthy limitations: i) the critical functionality of the Fido protocol regularly works in a customer platform such as a mobile device, which is vulnerable to a variety of attacks as malware and infections, its clients convey unsupervised computer program, and the deployed operating systems may be vulnerable to vulnerabilities; ii) the expense is additionally costly on account of the high foundation and support cost related with unified mists, huge server cultivates and organizing hardware. The sheer sum of communications that drive needs to be taken care of once IoT devices develop to the tens of billions will increment those costs significantly [64] [65]68.

To overcome these limitations, in the recent decade, a blockchain has drawn attention to improving reliability, auditability, security, and secrecy of the Internet of Things (IoT) where billions of gadgets are associated with the Internet to ease everyday life and offer customized services. The blockchain is a kind of appropriated record for keeping up an unchanging and deliberately structured record of value-based information [8]. A blockchain limit as a decentralized database is overseen by PCs having a place with a peer-to-peer (P2P) network. One of the advantages of blockchain is that it's open. Everyone taking part can see the blocks and the transactions stored in them. This doesn't mean everybody can see the real content of your exchange, such that's ensured by your private key.

Be that as it may, the current blockchain suffers from the main challenge which is overheads and scalability. In an average blockchain execution, all blocks are broadcast to and confirmed by all nodes. The bandwidth and memory which are restricted in the IoT devices and these highlights are wasteful to satisfy the complicated security issues and it prompts critical scalability issues since the broadcast traffic and preparing overheads would increase quadratically with the number of nodes in the network [9]. To accord with this, we require lightweight security schemes to verify the correspondence among taking an interest substance in the IoT condition. Thus, this study attempts to answer two questions (i) what are the existing security protocols that have been developed for IoT devices? And (ii) how a lightweight security model can be developed. To do so, we set the two objectives: (i) to study and analyze existing security protocols used in IoT applications, and ii) to propose a development methodology of a blockchain-based lightweight security model that provides end to end security for cloud-driven IoT-based Smart Devices.

II. Review of IoT-based Security Protocols in Smart Applications

A. *What is IoT*

The loT or the Internet of Things may be a basic concept of connection between electronic devices such as smart-phones, smart TVs, Tabs, Computers and actuators to the Internet. These devices are connected together in such a way that they will be empowering the user to perform an unused medium of communication between things and things additionally between people and things [1]. Within the world of IoT, everything genuine gets to be virtual, which implies that each individual and thing is locatable and addressable, and could be a readable object on the internet. IoT devices will indeed have "the capacity to sense, communicate, organize and create new data, turning into a necessary piece of the Internet.

The progression of the Internet of Things will revolutionize a number of segments, from smart home, smart city, smart grid and etc. IoT idea can likewise be inferred to make another framework and wide upgrade space for smart homes to pro ide smart, quality and to develop the quality of life. In this paper, we center on IoT based smart home applications. A general smart application of IoT environment is presented in Fig. 1.

B. *IoT Applications Involved In Security Issues*

Fig. 2 shows a comparison of the IoT applications percentage up to the present. We considered five IoT application spaces that include Smart City, Smart Home, Smart Health, and Smart Grid. Smart city and smart home have the highest portion.

Fig. 3 shows the Number of Security Vulnerabilities in Smart application IoT from 2010 – 2018. The result shows a steep increase in 2017 onward.

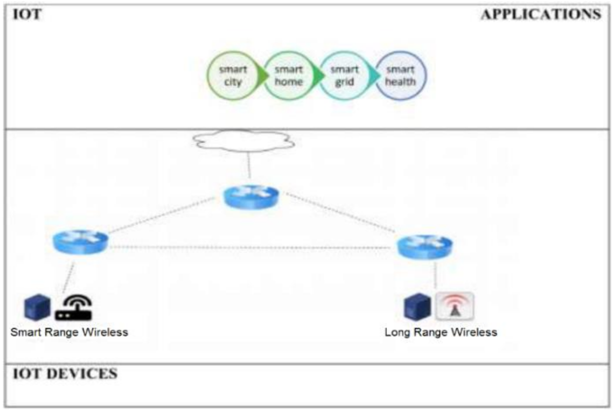


Fig. 1. An Example of IoT Application Environment.

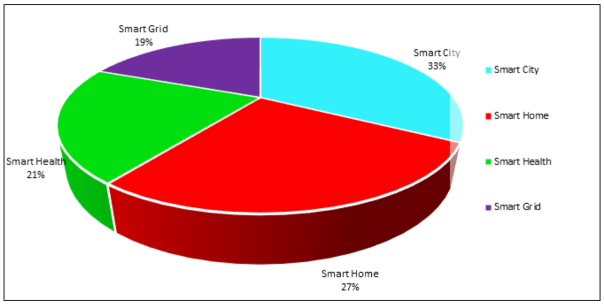


Fig. 2. Percentage of the showed IoT Applications.

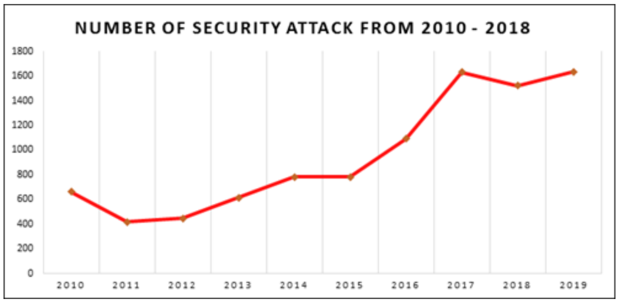


Fig. 3. Number of Security Vulnerabilities in Smart Application IoT.

Fig. 4 presents that the articles on the security protocol of IoT devices in a smart system that were included in this review hailed from 43 nations and nationalities. These articles, for the most part, include a research study conducted within the 43 countries.

In specific, the geological distribution of the picked articles on the security protocol of IoT devices in a smart system as far as numbers and rates show that the premier beneficial authors are from China, with 24 papers. This was taken after by India with 18 papers and USA with 15 study cases; Korea with 14 study cases; Saudi Arabia with 12; France and the Canada with 8 each; Germany, Malaysia, and UK with 6 each; Italy, New York, Pakistan, Singapore and Spain with 5 each; the Australia with 4; Brazil, Poland and Sweden with 3 each; and Indonesia, Japan, Jordon, Romania, Switzerland and Taiwan with 2 each; and Abu Dhabi, Beijing, Beirut Lebanon, Bosnia Herzegovina, Belgium, Egypt, Moscow, Sri Lanka, Thailand, Vietnam and etc. are with 1 paper each. Fig. 2 presented by the authors' nationality.

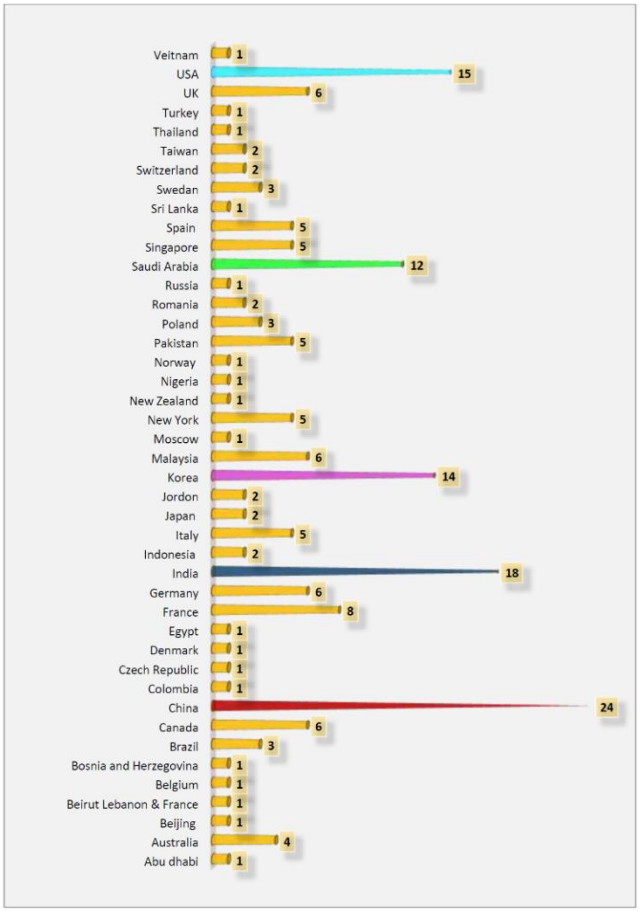


Fig. 4. Distribution by Authors' Nationality.

Smart Home: Smart Home is a term utilized to depict a home that contains a communication network that interfaces with various devices and enables them to be remotely controlled, observed, and got to by utilizing your smartphone applications, PC, and tablets [10]. It allows us to control our home appliances from a remote distance anytime and anywhere with Internet access [7]. The IoT smart home services are growing step by step, the technologies can viably communicate with one another using Internet Protocol (IP) addresses. All smart home devices are associated with the internet in a smart home environment. As the number of devices increases higher inside the smart home environment, the odds of malignant attacks also increase. [11]. One of the premiers touted focal points of a smart home is giving genuine feelings of serenity to house proprietors, empowering them to screen their homes remotely, countering dangers or threats, for example, an overlooked forgotten coffee maker left on or house door left open. In addition, smart home additionally assists consumers with improving capability [12]. Instead of leaving the air conditioning on all day, the home automation system can get familiar along with your behaviors and ensure the house is chilled off once you arrive home from work [13]. Notwithstanding security, various smart home users worry about data privacy. When developing solutions for smart home, security, and privacy are the best concerns [14]. Hence, breaking the security of a home automation system can lead to unauthorized get to access to private information. Existing vulnerabilities like poor setup and the use of default passwords are among the factors that can aid a hacker in compromising at the slightest one device in home automation [15]. Once a single device is compromised, hackers can take several actions based on the capabilities and functions of the device. Subsequently, it is very vital to distinguish all sorts of weaknesses and to address issues that can lead to unauthorized access to management of home automation and their information [16] [68].

Smart City: A Smart city is an urban zone that utilizations different sorts of electronic Internet of things (IoT) sensors to gather data and after that uses of this data to manage resources and resources effectively [17]. The data collected from people, devices and assets that will be processed and analyzed to monitor and manage information, traffic, water supply, hospital, power plants, video monitoring and so on. IoT Technology is making humans life better and easier. One of the ways will be secure wireless connectivity and IoT technology is changing traditional city life [18]-like streetlights into another era intelligent lighting stages with extended capabilities. It incorporates integrating solar power based power and associating with a cloud-based focal control system that will connect to other resources within the environment [19]. However, bringing insecure items into the smart city enormously broadens security and privacy risks. As many devices connected, vulnerabilities in a place get higher where attackers will have many loopholes to get the data [20]. In a general sense, each new device added to an IoT environment includes a new threat to the surface [21].

Smart grid: The smart grid is an electrical grid that is a combination of the electrical network and smart digital communication technology [22]. It has capable of giving electrical control from various and extensively circulated sources, as from wind turbines, solar-oriented control systems, and possibly surely module half and half electric vehicles [23]. A smart grid is a communication network on the beat of the power framework to assemble and analyze information from various parts of a power matrix to foresee power supply also, to predict which can be utilized for power managing [24]. Besides, a smart meter is one of the smart systems from the smart grid. It has installed at many organizations which to monitor the energy conception [25]. Besides, the smart grid has various characteristics such as data rate, time constraints, etc. This will be vulnerable to malicious cyber-attack of varying types that can severely obstruct its far deployment [26].

Smart Health: Smart health is defined by the technology that leads to greater treatment for patients, better diagnostic tools, and devices that can improve the quality of life for every individual. IoT changes the medical information into insights for smarter patient care [27]. Healthcare is now more technologically progressed and is all almost connecting devices together. In this manner, IoT is very important in the health system. Besides, by utilizing devices like connected sensors and other sorts of things that individuals can wear all that data can be placed within the cloud, and the doctor can effectively monitor the real-time data of the patient [28]. Nowadays, numerous healthcare devices operate all through the world which gets to be an issue because it can cause information loss and mistakes in diagnosis. To defeat this the information which is collected will be stored in the cloud. In security terms, IoT devices have constrained resources and these devices are associated with the internet. Hence, privacy and security are one of the enormous issues with IoT in healthcare [29] [30].

C. *Existing Security Protocols*

Numerous protocols have been proposed by the literature, Fig. 5 shows the number of security protocols articles from 2010 to 2019 and Fig. 6 shows the top highest number of protocols from 2010 to 2019.

However, in the following sections, we present the most recently used protocols which are, Zigbee, 6LoWPAN, Constrained Application Protocol (CoAP), Software-defined networking (SDN), and Blockchain.

Zigbee: ZigBee is a mesh network protocol. It is outlined to carry small information packets over brief distances whereas keeping up low control consumption. ZigBee is also an open-source wireless technology utilized in low-powered embedded devices (radio frameworks) to encourage productive [31]. It is more like an alternative to Bluetooth and Wi-Fi. ZigBee was based on the IEEE 802.15.4 standard detail and is made by a set of companies that shape the ZigBee Alliance. Whereas other wireless standards are concerned with exchanging huge sums of information, ZigBee is built for devices that have littler throughput needs [32]. Besides, the other driving components are low cost, high reliability, higher security, less battery usage, simplicity and interoperability with other ZigBee devices [33].

Moreover, one of the issues of the wireless sensors is that they require as well much power to operate properly however ZigBee gives long-lasting batteries with which they can remain lively for months or even a long time.

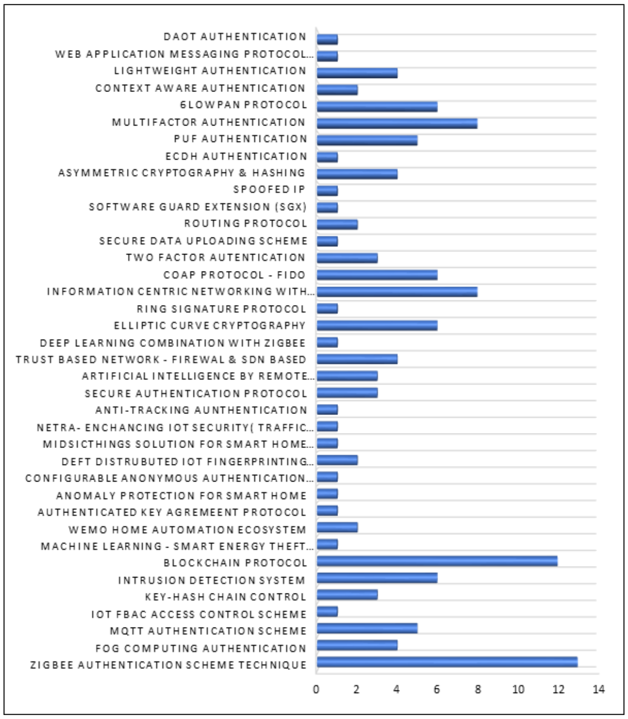


Fig. 5. Number of Security Protocols Articles from 2010 to 2019.

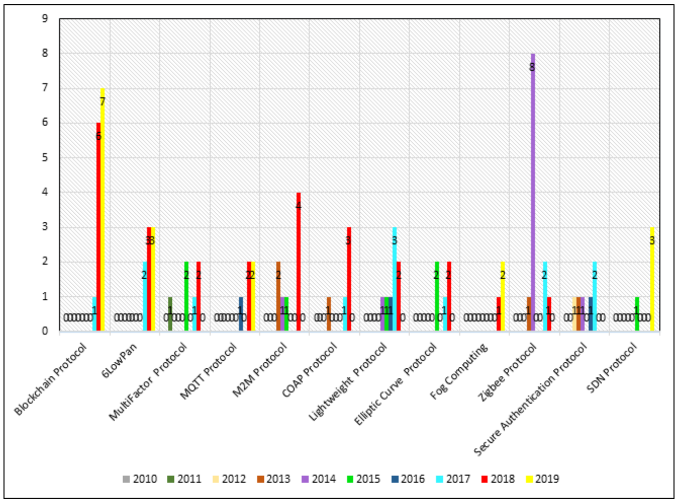


Fig. 6. The Top Highest Number of Protocol from 2010 to 2019.

Likewise, ZigBee devices are a lot less expensive than different devices. The low information rate will play an imperative part in the accomplishment of ZigBee in the future as the organizations will concentrate on the lost cost and low information rate solutions for their issues instead of costly ones [34]. Last but not least, ZigBee's capacity to support mesh networking implies it can boost information transmission extend and give more prominent stability (even when a single connected hub comes up short and doesn't work) but security isn't very well executed by the engineers in ZigBee [35]. The ZigBee network deployment in this study is displayed in Fig. 7.

6LoWPAN: 6LoWPAN is a direct low-cost communication organize that grants wireless network in applications with limited control and loosened up throughput prerequisites as it gives IPv6 organizing over IEEE 802.15.4 frameworks It is molded by devices that are reliable with the IEEE 802.15.4 standard and characterized by brief run, low bit rate, minimal effort, low control, and low memory usage

Exactly when a lower processing capacity sensor node in a 6LoWPAN or purported reduced capacity device (RFD) needs to send its information parcel to an IP-empowered device outside the 6LoWPAN, it at first sends the bundle to the higher preparing ability sensor node or so-called full function device (FFD) in a similar PAN. The FFDs which respond as a switch in 6LoWPAN will advance the information parcel bounce by a jump to the 6LoWPAN entryway. The 6LoWPAN gateway that connects with the 6LoWPAN with the IPv6 domain will at that point forward the packet to the destination IP-empowered device by utilizing the IP address.

A 6LoWPAN system consists of many embedded wireless remote devices that are perceived by the power constraint, low-information rate, and limited memory. The 6LoWPAN architecture is portrayed in Fig. 8 in which the end-to-end communication for interconnecting 6LoWPANs to the Internet is outlined. Each associated 6LoWPAN is an IPv6 stub network on the Internet, on the grounds that the IP packets can be gotten from or sent to it, however, there can't be a packet transit to other Internet systems.

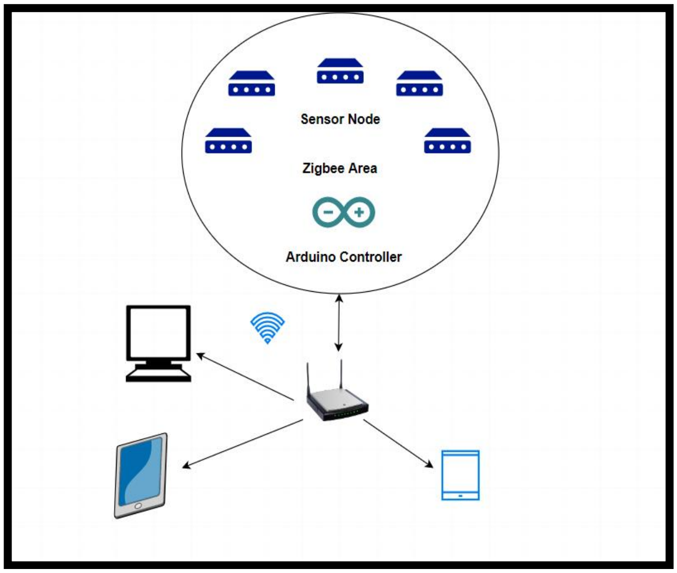


Fig. 7. Zigbee Network.

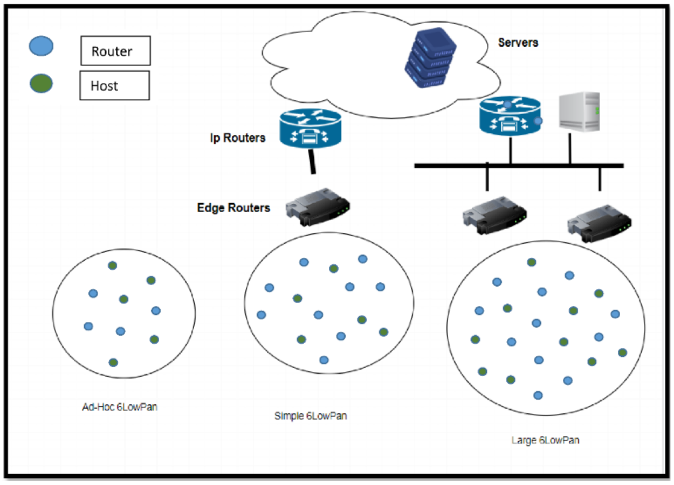


Fig. 8. 6LoWPan Network.

Constrained Application Protocol (CoAP): The Constrained Application Protocol (CoAP) is a web transfer protocol. It has been designed for a constraint environment and by the Internet Engineering Task Force (IETF). The main objective of CoAP is to restrict the required fragmentation by using small message overhead [36]. Also, this protocol appropriate for constrained systems such as 6LoWPAN which underpins the fragmentation of IPv6 packets into little outlines and the examples of the constrained devices is sensors, low power node, switches, and low power networks [37].

Constrained Application Protocol (CoAP) is a simplified version of HTTP and that is because the protocol looks more like a traditional website-based business [38] which gives the capacity to be compatible with an existing network that's web service-based [39]. Besides, this protocol has been created as a specialized web transfer protocol for utilizing constrained nodes and constrained like an example of low-power systems [40]. The CoAP protocol stack, where it utilizes Request/Response and a variant of Publish/Subscribe (Resource/Observe) architectures. The Request/Response model is like the Client/Server model in HTTP. CoAP is proposed to assume a comparable job as HTTP accomplishes for Web Internet and is being considered as a trade of HTTP for IoT networks and is turning into a standard protocol for some IoT solutions.

Software-defined networking (SDN): Software-defined networking (SDN) is a networking worldview that permits consistently centralized control of network switches and routers. SDN permits the probability of making new services and progressively productive applications dependent on the interaction with systems traffic, organize security usage, or quality of service [41]. In SDNs, most of the network capacities are actualized in applications [42]. The SDN controller keeps up a logical outline of the network and covers up the network complexity from applications through reflections.

SDN is a rising and promising innovation to make it occur. SDN decouples the control plane from the information plane. An SDN controller can take contributions from end structures applications and settle on choices to the information plane roughly what traffic can experience [43]. SDN can possibly benefit the security of IoT frameworks in at scarcest three viewpoints. In the first place, it can help shape a feedback-control loop from end IoT frameworks to the SDN controller which helps controls at least one programmable switches. Second, similar attacks to different exploited victims from a similar source inside a similar framework can be blocked and subsequently advantage the total IoT network in a cooperative manner [44]. Lastly, the attacking data can be shared among numerous peering controllers that oversee and control distinctive systems. As shown in Fig. 9, fruitful integration depended on the IoT system's utilization of key SDN highlights [12].

Blockchain: A blockchain is characterized as a distributed database that keeps up a changeless and tamper-proof record of value-based information. A blockchain is totally decentralized by depending on a peer-to-peer arrangement. More absolutely, each note of the arrangement keeps up a copy of the record to avoid a single point of failure. All duplicates are updated and approved at the same time. A block is an information structure that permits Blockchain to record the produced and traded exchanges and each block is connected to the chain by cryptography [45] The Blockchain is a distributed ledger that has three essential attributes: decentralized, transparent and recorded. All members keep and update a duplicate of a distributed ledger to check and approve transaction which makes Blockchain transparent and difficult to hack or lost any information [20]. Every transaction incorporates three principle segments, i.e., the information, the hash, and the hash of the previous block [21]. Each block in the system records the hash of the previous block. This prompts a chain of blocks with improved security. For instance, in Fig. 1, there is a chain of three blocks. Block 3 to block 2 and block 2 points to block 1 utilizing the hashes of previous block 1. On the off chance that hackers alter the second block information, the related block hashes changes. This makes the third block and every ensuing block invalid since they have not put away a legitimate hash of the previous block [21]. In Fig. 10, it presented how blockchain hash generally works.

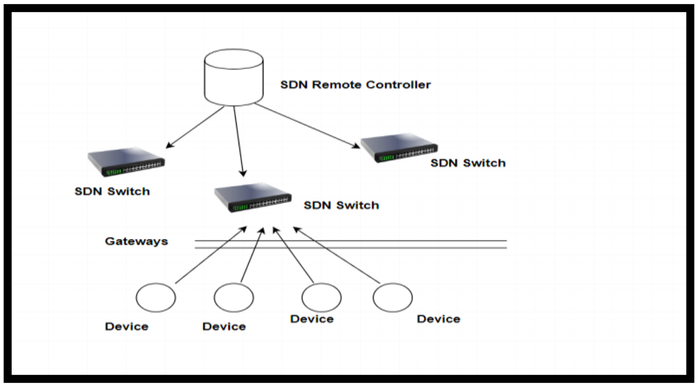


Fig. 9. SDN Network.

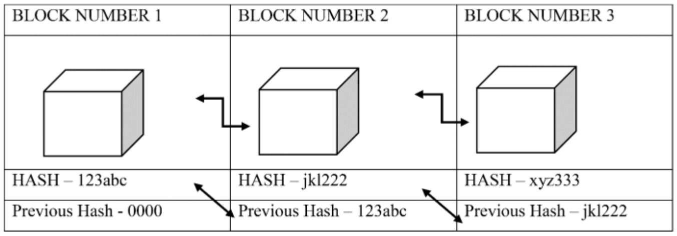


Fig. 10. How Blockchain Hash Works.

III. Why Blockchain Technology

A Blockchain is described as an “advanced, decentralized, and flowed record in which trades are logged and included sequential requests with the objective of making never-ending and tamperproof records" [46]. Basically, it is a novel instrument for storing, sharing information and securing between various nodes in a system [47]. Blockchain parts from the traditional unified by overseeing chain data over a dispersed and interlinked arrangement of nodes. The principle qualities of Blockchains are shared immutability, decentralization, recordkeeping, tamper evidence, distributed trust and tamper resistance [48]. The term 'Blockchain' picked up its prominence as the yield of a combination of configured advancement, methods and tools underpinning the digital currency Bitcoin. In itself, Bitcoin is a decentralized digital currency depending on an open system of a computer system and online communication protocols [49] and was the principal fruitful application based on an online Blockchain.

Using distributed technologies for IoT devices can understand security issues as well as include new features and lessen working expenses [66] [67]. Blockchain is an innovation that works with exchanges and gives communication in the system. It is incredible for monitoring processes in IoT. For instance, in light of the blockchain, you can bolster the identification and disclosure of gadgets, encourage microtransaction exchanges among them, and give proof of payment. In any case, blockchain technology stands to cause an immense effect on the Internet of Things. The blend of information that can't be adjusted however can be followed and verified from connected devices will drive the birth of exchanges among connected devices. With the intensity of blockchain technology, devices will have the option to network and direct trade as microtransactions utilizing a digital currency [50]. Blockchain technology will likewise upgrade security among the connected devices.

Blockchain technology can be utilized in following billions of connected devices, enable the handling of trades and coordination between devices; think about immense hold assets to IoT industry makers. This decentralized methodology would wipe out single points of failure, making a stronger ecosystem system for devices to run on. The record is tamper-proof and cannot be controlled by noxious actors since it doesn’t exist in any single area, and what's more, man-in-the-middle attacks can't be organized since there's no single string of communication that can be catching [51]. In an IoT system, the blockchain can keep a changeless record of the historical backdrop of smart devices. This include empowers the independent functioning of smart devices without the required for centralized specialist. Fig. 11 shows the advantages of blockchain.

The perspective on general security requirements just as necessities of security is given by blockchain. Fig. 12 presented the security requirement of blockchain and the following shows a detailed explanation of the security requirement of blockchain.

- Integrity: The point of the integrity of information is to keep up the consistency and accuracy of information all through the lifecycle of information [52].
- Authentication: Authentication is stated to a procedure inside which the credentials gave to get to a file are compared and the ones that are given in the database by the approved clients [53].
- Verifiable: Verifiable substances in the system are important to ensure the blockchain ledger against tampering [54].
- Confidentiality: Confidentiality of information is to protect it from unveiling to unauthorized parties [55],
- Trust: The fundamental job of producing trust is to guaranteeing trustworthiness, reliability or the capability of the individual system nodes based on the applied monitoring schemas [56] [57].
- Anonymity: Anonymity has applied consequently with the execution of the blockchain. In the blockchain, exchanges are performed with secret keys and the public [58].
- Immutability: Traditional relational databases give variable storage, as changes made to a particular record or table are supplanted in that document [59].
- Authorization: the authorization is a procedure wherein the get to a level of a previously authenticated client is allowed. In which it is resolved that activities can a client performed and for what he/she isn't permitted [60].
- Privacy: The blockchain-based IoT approaches in the writing who have kept up the privacy are [54] [61].

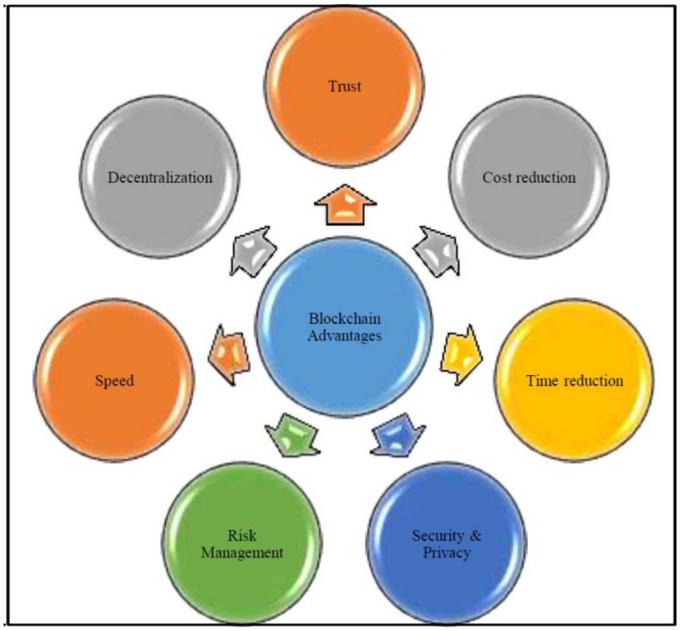


Fig. 11. The Advantages of Blockchain.

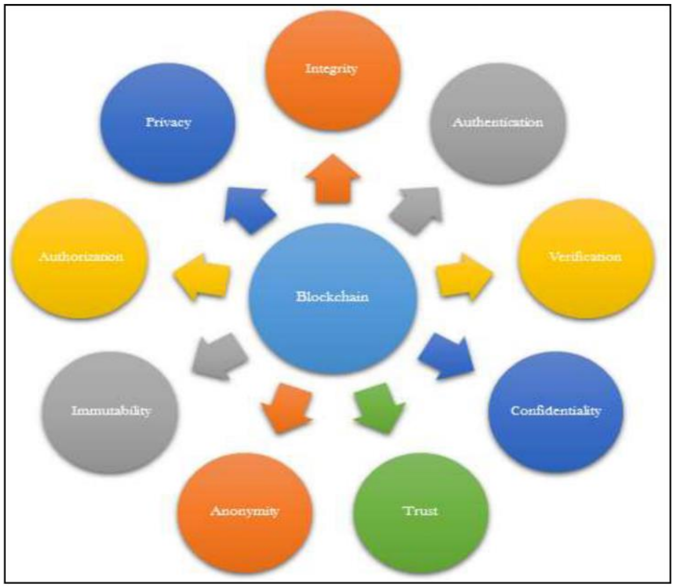


Fig. 12. Presented Security Requirement of Blockchain.

IV. Development Methodology

In research design, we conceive a collection of research activities that lead to the achievement of the research objectives. Fig. 13 shows the research activities that we compiled as the research design. It consists of activities that review the literature to develop the problem statement along with the research questions and objectives. Particularly, the study relies on the existing models in the literature to develop the conceptual model. The models are efficient Lightweight integrated Blockchain (ELIB) which is our baseline method; A Lightweight Scalable Blockchain (LSB), and a Universal Authentication System Protocol (UAF-FIDO).

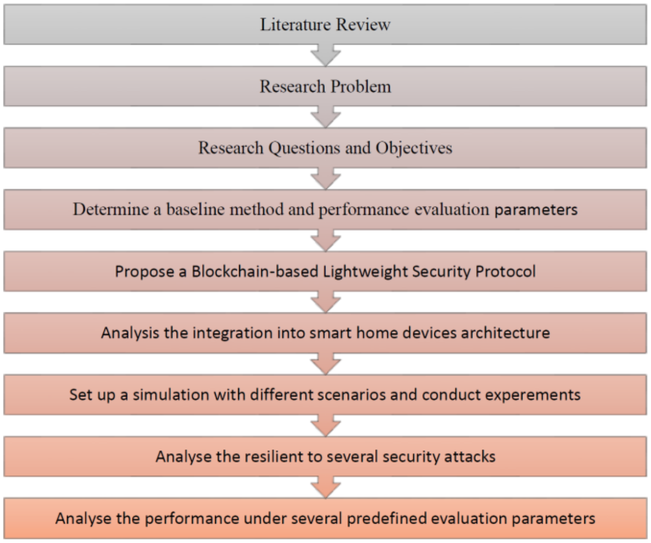

Literature Review
Research Problem
Research Questions and Objectives
Determine a baseline method and performance evaluation parameters
Propose a Blockchain-based Lightweight Security Protocol
Analysis the integration into smart home devices architecture
Set up a simulation with different scenarios and conduct experements
Analyse the resilient to several security attacks
Analyse the performance under several predefined evaluation parameters

Fig. 13. Development Methodology of a Blockchain-based Light Weight.

A. Determine a Baseline Method and Performance Evaluation Parameters

The efficient Lightweight integrated Blockchain (ELIB) model is created to meet the requirements of secure IoT. The ELIB accomplishes a sum of half sparing (50%) in dealing with time on contrasting with the baseline method with the base energy utilization. The experiment exhibited that the ELIB shows the most extraordinary performance under a couple of evaluation parameters.

The ELIB models need more time for packet handling with when just like the baseline which has the component of the extra-encryption and hashing limits. Within the destitute case for the inquiry-based store T, additional overhead brought about by ELIB that is incredibly low in absolute terms.

B. Propose a Blockchain-Based Lightweight Security Protocol

In spite of the fact that Blockchain is a viable technology for giving privacy and security in IoT, applications within the IoT context presents a few noteworthy challenges. IoT devices don't have adequate memory, control, for calculating or zone for executing a hardware module, since the device can perform a particular reason. Therefore, we use lightweight to overcome the restriction. Within the architecture of the smart home, we supplant fido protocol and integrate with blockchain-based lightweight to decrease packet overheads on IoT devices and to have secure communication with clients and gadgets. Consequently, IoT devices have constrained resources, hence, confirming all modern blocks and exchanges may be distant past their capabilities. To guarantee scalability and lessen processing and packet overhead on IoT devices, we accept that the blockchain is overseen by a subset of the overlay hubs. This decreases the packet overhead for information transactions. To guarantee information integrity, the exchange corresponding to the traded information between overlay nodes contains the hash of the information signed by the exchange generator [9].

C. Analysis of the Integration into Smart Home Devices Architecture

Within the architecture of the smart home, we supplant fido protocol and integrate with blockchain-based lightweight to diminish parcel overheads on IoT gadgets and to have secure communication with clients and devices. Consequently, IoT devices have restricted assets, in this way, confirming all modern blocks and exchanges may be distant past their capabilities. To ensure scalability and lessen handling and packet overhead on IoT devices, we acknowledge that the blockchain is overseen by a subset of the overlay nodes. Here, each person node within the overlay is called as PK. This decreases the packet overhead for information trade. To guarantee information integrity, the transaction compared to the exchanged information among overlay nodes contains the hash of the information signed by the exchange generator.

D. Set up a Simulation with different Scenarios and Conduct Experiment

Here we utilize two simulators as takes after to evaluate the performances of the proposed solution and Table I shows a brief summary of the simulator.

E. Analyze the Resilient to Several Security Attacks

The study comes up with several security attacks to which IoT systems or blockchains are especially vulnerable and layout how LSB secures against them.

Denial of Service (DOS) attack: The design of DOS attack has a few hierarchical security protection against this attack. It has few levels of a protective layer that can be credited to the reality that it would be impossible for an attacker to straightforwardly install malware on smart IoT devices since these devices are not legitimately accessible [62]. Let us for a minute expect that the attacker some way or another still figure out way of how to infect the devices. There it comes the LSB which is Lightweight Scalable Blockchain ensures against these attacks. The method of LSB is to defend the attack like the example of OBM. Overlay Block Manager (OBM) would not send an exchange to their cluster members except if they discover a match with a substance in their key list.

Dropping attack: The attacks influence the consumption of energy and packet drop parameters in the network [63]. From the attack, OBM is an entity responsible for the blockchain management. Therefore, OBM drops exchanges to or on the other hand from its cluster members to confine them from the overlay network. A segment of the exchanges is verified as the OBMs manufacture up trust in each other. To ensure along with the attack, a cluster member can change the OBM it is related to on the off chance that it sees that its exchanges are most certainly not being handled.

TABLE. I. SIMULATOR BRIEF EXPLANATION

Tool	Description
Cooja Simulator	Utilize Cooja to consider the performance of the smart home tier.
NS3 Network Simulator	Utilize NS3 to evaluate the overlay performance because it has been broadly utilized for analyzing peer-to-peer systems

Blockchain: If an attacker advertises a false record of blocks and makes it as the longest record. The proposed DTC will restraint the number of blocks each OBM can create in a time interval. This will constrain the number of noxious blocks that an OBM can affix.

F. Analyze the Performance under Several Predefined Evaluation Parameters

This phase gives a point by point validation of a distinctive view of the proposed protocol. To assess the performance, four evaluation parameters have been identified from the literature, processing time, energy usage, overhead, and Scalability.

V. CONCLUSION

Security is consistently and will be a critical perspective on the IoT network. In this paper, it has indicated how the security concerns in the IoT applications have developed. The methodical mapping process of this study discovers how the development has occurred, what sorts of concerns and solutions exist, and what gaps remain. Based on the review outcomes, the public cloud that shares computing services among different customers runs by third-party suppliers over the Internet, making devices accessible online. However, the public cloud is an increasingly attractive objective for hacker and information in the cloud is risky. Therefore, reliable protocol to securely communicate between IoT devices and the cloud is inevitable. On the other hand, Blockchain Can Be Game-Changer for IoT for now and the future. Despite the advantages of blockchain technology, several limitations have been addressed in this study such as High Complexity, Restricted Scalability, High Bandwidth Overhead, and Latency –Delay & overhead. To overcome these limitations, in our future work, we shall develop a lightweight security scheme to verify the correspondence among taking an interest substance in the IoT condition. More specifically, we shall propose a blockchain-based lightweight solution that offers better security and low communication.

ACKNOWLEDGMENT

This project is sponsored by Universiti Tenaga Nasional (UNITEN) under the Bold Research Grant Scheme No. 10346494/B/2019046.

REFERENCES

[1] Ahmad, S., Hang, L., & Kim, D. H. (2018). Design and implementation of cloud-centric configuration repository for DIY IoT applications. Sensors, 18 (2), 474.

[2] Höller J., “Introduction to a New Age of Intelligence,” Mach. to Internet Things, 2014.

[3] Sriram D., “Trust Based Security for Cloud Systems What needs to be done to solve this problem ? What has been done ? What can be done ?,” pp. 3–5.

[4] Delsing, J., Eliasson, J., van Deventer, J., Derhamy, H., & Varga, P. (2016, December). Enabling IoT automation using local clouds. In 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT) (pp. 502-507). IEEE.

[5] Wazid, Mohammad, et al. "LAM-CIoT: Lightweight authentication mechanism in cloud-based IoT environment." Journal of Network and Computer Applications 150 (2020): 102496.

[6] Eric Vanderburg, “Public Cloud Security Concerns Remain after Recent Study,” TCDI Blog, 2019.

[7] Chifor, B. C., Bica, I., Patriciu, V. V., & Pop, F. (2018). A security authorization scheme for smart home Internet of Things devices. Future Generation Computer Systems, 86, 740-749.

[8] Cooijmans, T., de Ruiter, J., & Poll, E. (2014, November). Analysis of secure key storage solutions on android. In Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices (pp. 11-20).

[9] Dorri, A., Kanhere, S. S., Jurdak, R., & Gauravaram, P. (2017). Lsb: A lightweight scalable blockchain for iot security and privacy. arXiv preprint arXiv:1712.02969.

[10] Marikyan, D., Papagiannidis, S., & Alamanos, E. (2019). A systematic review of the smart home literature: A user perspective. Technological Forecasting and Social Change, 138, 139-154.

[11] Yoon, S., Park, H., & Yoo, H. S. (2015). Security issues on smarthome in IoT environment. In Computer science and its applications (pp. 691-696). Springer, Berlin, Heidelberg.

[12] Jacobsson, A., Boldt, M., & Carlsson, B. (2016). A risk analysis of a smart home automation system. Future Generation Computer Systems, 56, 719-733.

[13] Lyu, Q., Zheng, N., Liu, H., Gao, C., Chen, S., & Liu, J. (2019). Remotely Access "My" Smart Home in Private: An Anti-Tracking Authentication and Key Agreement Scheme. IEEE Access, 7, 41835-41851.

[14] Schiefer, M. (2015, May). Smart home definition and security threats. In 2015 ninth international conference on IT security incident management & IT forensics (pp. 114-118). IEEE.

[15] Mohammad Z., Qattam T. A., and Saleh K., "Security Weaknesses and Attacks on the Internet of Things Applications," 2019 IEEE Jordan Int. Jt. Conf. Electr. Eng. Inf. Technol., pp. 431–436, 2019.

[16] Lin B. N. H, "IoT privacy and security challenges for smart home environments." 2016.

[17] Kumar A., Zeadally S., and He D., "Taxonomy and analysis of security protocols for Internet of Things," Futur. Gener. Comput. Syst., vol. 89, pp. 110–125, 2018.

[18] Airehrour D., Gutierrez J., and Kumar S., "Journal of Network and Computer Applications Secure routing for internet of things : A survey," vol. 66, pp. 198–213, 2016.

[19] Gemalto, "Secure, sustainable smart cities and the IoT," 2019. [Online]. Available: https://www.gemalto.com/iot/inspired/smart-cities.

[20] Ribagorda A., Alcaide A., and Palomar E., "Anonymous authentication for privacy-preserving IoT target-driven applications," vol. 7, 2013.

[21] El-hajj M., Fadlallah A., and Serhrouchni A., "Taxonomy of Authentication Techniques in Internet of Things (IoT)," 2017.

[22] Antonopoulos A. M., "Mastering Bitcoin: unlocking digital cryptocurrencies," O'Reilly Media, Inc., vol. 6, no. 5, pp. 900–917, 2014.

[23] Miloslavskaya N. and Tolstoy A., "Internet of Things : information security challenges and solutions," Cluster Comput., vol. 22, no. 1, pp. 103–119, 2020.

[24] Ghasempour, "Optimum Number of Aggregators based on Power Consumption, Cost, and Network Lifetime in Advanced Metering Infrastructure Architecture for Smart Grid Internet of Things.," Proc. IEEE Consum. Commun. Netw. Conf. (IEEE CCNC 2016), p. 2016, 2016.

[25] Sheik Dawood M. J. M. M., Abinaya P, "Improving the Network Lifetime and Energy Conservation using Target Trail in Cluster of Mobile Sensor Networks," Asian J. Res. Soc. Sci. Humanit., no. 18, pp. 430–447, 2016.

[26] Gupta V. A. B. B., "Security in Internet of Things : issues , challenges , taxonomy , and architecture," Telecommun. Syst., vol. 67, no. 3, pp. 423–441, 2018.

[27] Jing Q., A. Vasilakos V, and Wan J., "Security of the Internet of Things : perspectives and challenges," pp. 2481–2501, 2014.

[28] Neelam S., "Internet of Things in Healthcare," Computing at Blekinge Institute of Technology, 2017.

[29] Madakam S. T. S., Ramaswamy R., "Internet of Things (IoT): A literature review," J. Comput. Commun., p. 164, 2015.

[30] Suresh R. A. P., Daniel J. V., Parthasarathy V., "A state of the art review on the Internet of Things (IoT) history, technology and fields of deployment," Sci. Eng. Manag. Res. (ICSEMR), 2014 Int. Conf. on, 2014, pp. 1–8, 2014.

[31] Yiqi W., Lili H., Chengquan H., Yan G., and Zhangwei Z., "2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications A ZigBee-based smart home monitoring system," 2014 Fifth Int. Conf. Intell. Syst. Des. Eng. Appl., pp. 114–117, 2014.

[32] Gao L., Wang Z., Zhou J., and Zhang C., "Design of Smart Home System Based on ZigBee Technology and R & D for Application," no. January, pp. 13–22, 2016.

[33] Lewis E., Cook F. L. D. J. and Das S. K., "Wireless Sensor Networks," in Smart Environments: Technologies, Protocols and Applications," no. January, pp. 227–228, 2014.

[34] Talal M., "Smart Home-based IoT for Real-time and Secure Remote Health Monitoring of Triage and Priority System using Body Sensors : Multi-driven Systematic Review," 2019.

[35] Kulkarni S., "Considering Security For ZigBee Protocol Using Message Authentication Code," no. February, 2019.

[36] AliA. A. and Member H. I. S., "Constrained Application Protocol (CoAP) for the IoT," no. May, 2018.

[37] Randhawa R. H., Hameed A., and Mian A. N., "Energy efficient cross-layer approach for object security of CoAP for IoT devices," Ad Hoc Networks, no. xxxx, p. 101761, 2019.

[38] Lamichhane M., "CoAP for IOT," ITMO University, Russia, 2017.

[39] Jonathan Fries, "Why are IoT developers confused by MQTT and CoAP," 2017. [Online]. Available: techtarget.com.

[40] Jang S., Lim D., and Kang J., "An Efficient Device Authentication Protocol Without Certification Authority for Internet of Things," Wirel. Pers. Commun., vol. 91, no. 4, pp. 1681–1695, 2016.

[41] Olivier F., Carlos G., and Florent N., "New Security Architecture for IoT Network," Procedia - Procedia Comput. Sci., vol. 52, no. BigD2M, pp. 1028–1033, 2015.

[42] Open Network Foundation, "Software-Defined Networking: The New Norm for Networks.," 2017. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/ sdn-resources/white-papers/wp-sdn-newnorm.pdf.

[43] Sharma P. K., J. Park H., Jeong Y., and Park J. H., "SHSec : SDN based Secure Smart Home Network Architecture for Internet of Things," pp. 913–924, 2019.

[44] Ahmed A. W., "Software Defined Network (SDN) Based Internet of Things (IoT): A Road Ahead," no. July, 2017.

[45] Wüst K. and Gervais A., "›'Do you need a blockchain?,'" Crypto Val. Conf. Blockchain Technol. (CVCBT). IEEE, 2018, pp. 44–45, 2018.

[46] Treiblmaier H., "The impact of the blockchain on the supply chain: a theory-based research framework and a call for action," Supply Chain Manag., vol. 23, no. 6, pp. 545–559, 2018.

[47] Atlam H. F. and Wills G. B., Intersections between IoT and distributed ledger, 1st ed., vol. 115, no. January. Elsevier Inc., 2019.

[48] Rauchs M. et al., "Distributed Ledger Technology Systems: A Conceptual Framework," SSRN Electron. J., no. August, 2018.

[49] Fosso W. S., Kala Kamdjoug J. R., Epie Bawack R., and Keogh J. G., "Bitcoin, Blockchain, and FinTech: A Systematic Review and Case Studies in the Supply Chain Blockchain, and FinTech: A Systematic Review and Case Studies in the Supply Chain. Production Planning and Control, Forthcoming. *Corresponding author Bitcoin, Bloc," pp. 0–53, 2018.

[50] Zhang A., Zhong R. Y., Farooque M., Kang K., and Venkatesh V. G., "Blockchain-based life cycle assessment: An implementation framework and system architecture," Resour. Conserv. Recycl., vol. 152, no. May 2019, p. 104512, 2020.

[51] Treiblmaier H., "Toward More Rigorous Blockchain Research: Recommendations for Writing Blockchain Case Studies," Front. Blockchain, vol. 2, no. May, pp. 1–15, 2019.

[52] Apte S. and Petrovsky N., "Will blockchain technology revolutionize excipient supply chain management?," J. Excipients Food Chem., vol. 7, no. 3, pp. 76–78, 2016.

[53] Patil H. K. and Seshadri R., “Big data security and privacy issues in healthcare,” Proc. - 2014 IEEE Int. Congr. Big Data, BigData Congr. 2014, pp. 762–765, 2014.

[54] Zhang J., “A multi-transaction mode consortium blockchain,” Int. J. Performability Eng., vol. 14, no. 4, pp. 765–784, 2018.

[55] Hersh W. R. et al., “Health Information Dissemination from Hospital To Community Care : Current State And Next Steps In Ontario,” J. Med. Syst., vol. 63, no. 50, pp. 425–432, 2016.

[56] Sun Y., Han Z., and K. Liu J. R., “Defense of trust management vulnerabilities in distributed networks,” IEEE Commun. Mag., vol. 46, no. 2, pp. 112–119, 2008.

[57] Entrust, “The Concept of Trust in Network Security,” Entrust White Pap., no. August, pp. 1–7, 2000.

[58] Hodges E., “Blockchain is where anonymity meets transparency,” 2018.

[59] Morrison A., “The rise of immutable data stores,” 2018. [Online]. Available: http://usblogs.pwc.com/emerging-technology/the-riseof-immutable-data-stores/.

[60] Biswas K. and Muthukkumarasamy V., “Securing smart cities using blockchain technology,” Proc. - 18th IEEE Int. Conf. High Perform. Comput. Commun. 14th IEEE Int. Conf. Smart City 2nd IEEE Int. Conf. Data Sci. Syst. HPCC/SmartCity/DSS 2016, no. December, pp. 1392–1393, 2017.

[61] Kravitz J. C. D.W., “Securing user identity and transactions symbiotically: IoT meets blockchain,” Glob. Internet Things Summit, GIoTS, 2017.

[62] Dorri A., Kanhere S. S., Jurdak R., and Gauravaram P., “Blockchain for IoT security and privacy: The case study of a smart home,” 2017 IEEE Int. Conf. Pervasive Comput. Commun. Work. PerCom Work. 2017, no. October, pp. 618–623, 2017.

[63] Eastman D. and Kumar S. A. P., “A simulation study to detect attacks on internet of things,” Proc. - 2017 IEEE 15th Int. Conf. Dependable, Auton. Secur. Comput. 2017 IEEE 15th Int. Conf. Pervasive Intell. Comput. 2017 IEEE 3rd Int. Conf. Big Data Intell. Comput. 2017 IEEE Cyber Sci. Technol. Congr. DASC-PICom-DataCom-CyberSciTec 2017, vol. 2018-January, pp. 645–650, 2018.

[64] Al-Momani, A. M., Mahmoud, M. A., & Ahmad, M. S. (2018). Factors that influence the acceptance of internet of things services by customers of telecommunication companies in Jordan. Journal of Organizational and End User Computing (JOEUC), 30(4), 51-63.

[65] Al-Momani, A. M., Mahmoud, M. A., & Ahmad, M. S. (2019). A review of factors influencing customer acceptance of internet of things services. International Journal of Information Systems in the Service Sector (IJISSS), 11(1), 54-67.

[66] Al-Momani, A. M., Mahmoud, M. A., & Ahmad, M. S. (2018). Identification of Factors Influencing Customer Acceptance and Use of IoT Services. Advanced Science Letters, 24(10), 7428-7432.

[67] Al-Momani, A. M., Mahmoud, M. A., & Ahmad, M. S. (2016). Modeling the adoption of internet of things services: A conceptual framework. International Journal of Applied Research, 2(5), 361-367.

[68] Kumar, K., & Mahmoud, M. A. (2017). Monitoring and Controlling Tap Water Flow at Homes Using Android Mobile Application. American Journal of Software Engineering and Applications, 6(6), 128-136.

SentiFilter: A Personalized Filtering Model for Arabic Semi-Spam Content based on Sentimental and Behavioral Analysis

Mashael M. Alsulami[1], Arwa Yousef AL-Aama[2]
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—Unwanted content in online social network services is a substantial issue that is continuously growing and negatively affecting the user-browsing experience. Current practices do not provide personalized solutions that meet each individual's needs and preferences. Therefore, there is a potential demand to provide each user with a personalized level of protection against what he/she perceives as unwanted content. Thus, this paper proposes a personalized filtering model, which we named SentiFilter. It is a hybrid model that combines both sentimental and behavioral factors to detect unwanted content for each user towards pre-defined topics. An experiment involving 80,098 Twitter messages from 32 users was conducted to evaluate the effectiveness of the SentiFilter model. The effectiveness was measured in terms of the consistency between the implicit feedback derived from the SentiFilter model towards five selected topics and the explicit feedback collected explicitly from participants towards the same topics. Results reveal that commenting behavior is more effective than liking behavior to detect unwanted content because of its high consistency with users' explicit feedback. Findings also indicate that sentiment of users' comments does not reflect users' perception of unwanted content. The results of implicit feedback derived from the SentiFilter model accurately agree with users' explicit feedback by the indication of the low statistical significance difference between the two sets. The proposed model is expected to provide an effective automated solution for filtering semi-spam content in favor of personalized preferences.**

Keywords—Personalization; sentiment analysis; behavioral analysis; spam detection; recommendation systems

I. Introduction

Online Social Network (OSN) services provide online and instant communication in a large-scale manner. Despite the great social experience and communication benefits of these services, the vast usage of OSN services increases the amount of user-generated content, which brings several challenges and concerns regarding privacy, data management, information filtering, and content moderation. Users of such services are exposed to various kinds of content that can be unwanted or harmful [1].

Unwanted content can be defined as any electronic content, including text and multimedia that is not expected or welcomed by its final destination because of its disturbing or annoying nature. Unwanted content has been mostly considered and identified as spam content, which is received from undesired sources called spammers [2].

Solving the spam issue requires taking into consideration several aspects in order to propose solutions. These aspects include type of spam, where to detect spam, the form of spam, and how to detect it.

However, defining what spam is from users' personal perspectives needs further investigation. Personalization techniques could help to customize users' social space in OSN services and give the ability to recognize and detect what users really consider as spam messages to prevent them or block them from being received.

OSN services provide several interaction attributes that can be considered as indicators of users' perceptions of semi-spam content such as sharing/forwarding behavior, liking behavior, reporting behavior, and commenting behavior. Therefore, the authors of this paper assume that users use commenting behavior when they find a post they like or agree about or a post they do not want or disagree with. The aim of this work is to infer users' perception about a particular topic from detecting the sentiment of their comments to a post involving that topic combined with other behavioral factors.

In the context of our work, semi-spam content is defined as any electronic message that a particular user perceives as unwanted, unpleasant, annoying, or disturbing, based on his/her interaction behavior.

The work in this paper empirically assesses the impact of combining the sentimental factor of users' comments with liking behavior to detect semi-spam content for a particular user.

Accordingly, a personalized filtering model was designed and developed, which we named SentiFilter model, to filter out semi-spam content based on the sentiment polarity of users' comments combined with users' liking behavior. The effectiveness of using behavioral and sentimental factors in detecting semi-spam content was evaluated by comparing the implicit feedback derived from each behavioral and sentimental factor against users' explicit feedback about certain topics. More precisely, the work in this paper focuses mainly on Arabic messages, since there is very little research found in the literature on filtering Arabic spam content.

However, most of the existing studies concentrate on a particular definition of spam, such as inappropriate content, bullying content, racism, and hateful-speech content, without consideration of personalization or personal preferences. Another limitation in previous work is the focus on either sentiment or behavior as an indicator of users' preferences without making full use of both factors to detect semi-spam content for each individual. In this research, experimentation was carried out on the tweets dataset extracted from the timelines of 32 Twitter users to assess the impact of sentimental and behavioral factors in reflecting users' perceptions of a given topic. The research question that this paper aims to answer is as follows: Which behavioral or sentimental factors are more effective to detect semi-spam content in terms of the agreement between the implicit feedback, derived from each factor, and users' explicit feedback about a topic?

The main contributions of this paper are summarized as follows:

- Propose a personalized filtering model for semi-spam content, which we called SentiFilter that combines both sentimental and behavioral factors in detecting semi-spam content.
- Propose a personalized aggregate factorization algorithm, which we named the personalized aggregate factorization (PAF) algorithm that combines sentiment of users' comments with liking behavior to detect Arabic semi-spam content.
- Propose a list-based classifier that applies our proposed PAF algorithm to maintain users' blacklists, whitelists, and greylists.
- Compare the effectiveness of behavioral and sentimental factors in terms of the agreement between users' explicit feedback and the implicit feedback derived from the SentiFilter model.

The paper is organized as follows. Section II discusses the related work on personalized spam detection in the relevant literature. Then, an overview of the proposed SentiFilter model is demonstrated, including the proposed PAF algorithm, in Section III. Section IV explains the design of the experiment. Results and discussion are discussed in Section V. Finally, conclusion and future work are given in Section VI.

II. Related Work

In this section, we highlight previous work that proposed solutions to the problem of semi-spam messages from two points of view: spam detection solutions and personalization.

A. *Spam Detection*

Spam, as a term, has been used to define various types of unwanted content including spam emails, SMS spam, spam URLs, social spam, and web spam. Based on the definition of spam content, spam can be recognized by several forms such as malicious content, malware content, inappropriate content, not-safe-for-work content, or denial of service attacks [2]. Studies that involved users' perspectives in identifying spam content have used terms such as semi-spam [3] and grey spam [2]. Previous work in recommendation systems such as [4] and [5] benefited from the sentiment of customers' reviews to infer users' preferences and provide them with personalized recommendations.

Several interventions and techniques have been proposed in the literature to solve the problem of spam messages. Traditional techniques aim to block the source or distributor of inappropriate content while existing methods examine the content to extract features in order to recognize certain patterns and predict them using machine learning algorithms. Additionally, the definition of inappropriate content is varied in the literature. Some studies describe inappropriate content as spam content or not-safe-for-work content as in [6][7][8], where other studies focused on a specific type of content, such as text messages, and performed analysis processes to detect inappropriate content such as abusive language [9] or bullying behavior [10]. Spam URLs have also been investigated in [11] using behavioral analysis.

Based on existing literature in detecting unwanted content, this paper categorizes related work in this area of spam detection systems into list-based filtering and content-based filtering techniques. List-based spam detection methods aim to apply blacklisting techniques to detect the sources or the distributors of spam content and block them. On the other hand, content-based spam detection methods aim to extract features from the content itself to identify unwanted patterns.

1) List-based spam filtering: The concept behind list-based filtering is to create a blacklist of distributors of spam content. Those lists are blacklists, whitelists, and sometimes greylists. For instance, Tewari and Jangale [12] defined greylists by the sending pattern of the sender, where a greylist contains all unknown users who are initially rejected by the mail server. The mail server of the receiver will send a failure notification to the mail server of the sender. If the mail server of the sender sends the message again, the mail server of the receiver will accept it and move it from greylist to whitelist [12].

They classified senders by how many times they send the same message, relying on the fact that spam emails are usually sent in batches. On the other hand, Liu et al. [3] classified spam emails into two categories: complete spam and semi-spam emails. They considered complete spam emails as emails identified by all users as spam emails, while semi-spam emails are identified by crowdsourcing using trusted contacts. Therefore, a trust value needs to be assigned and computed for each contact [3].

O'Connor et al. [13] proposed a method to determine if a user is a source or a distributor of unwanted content by tracking his/her activities when a message was sent. They defined a set of metrics to decide if a certain user is a source of undesired electronic content. These metrics included message rate, block count, block rate, and message uniqueness. Their method can be considered as a collaborative method that collects information from users of a specific application to make a decision to ban or prevent users from that application or to add them to a watch list. The main goal of their method was to identify users who send such unwanted content. They defined patterns that indicated the distributor of unwanted content because those users usually change their accounts' information periodically [13].

Bodkhe et al. [14] proposed a filtering method called Filter Wall (FW) to filter unwanted content based on a message trust management method. In their approach, a trust value was assigned for each message to classify it as wanted or unwanted. The trust value was assigned by users who were using the same application. The classification was based on computing the trustworthiness of each sender. The authors computed this value by aggregating the trust values for each message that each sender had sent and got its average [14].

Ma and Yan [15] also addressed the problem of blocking sources of unwanted content using a trust management method, which is a method in the communication field that is used to control unwanted traffic [15]. They proposed a system called PSNController to manage unwanted content. The brief concept of this system was to assign a trust value for each user. Users who used the same application had the ability to see the trust value of each other. PSNController is a customizable system to monitor unwanted content and identify its sources. They categorized unwanted content as bad text attacks, distributed denial of service, spammed multimedia, and viruses. They evaluated the system in terms of accuracy, efficiency, and robustness [15].

2) Content-based spam filtering: Content-based filtering (CBF) is mainly performed by determining the correlation between the content of items and user's preferences [16]. Applying CBF requires analyzing the content of each item to represent it as a set of features or terms, which is an expensive process.

Detecting spam messages based on their content is applied to several applications such as emotion recognition and inappropriate content detection tools. To detect spam messages based on their content, several approaches have been proposed in the literature.

In terms of Arabic spam detection, Mubarak and Darwish [9] studied the problem of detecting abusive Arabic text in social media. They used Twitter to create an Arabic corpus that contained a list of obscene words. They used that list to classify Twitter users based on their use of these words.

Another approach that applies CBF is proposed by Zitouni et al. [16]. They proposed a semantic content-based filtering technique, which benefits from the Web of Data concept, which is a term that refers to using all interconnected knowledge about different domains in the World Wide Web as a global database [15]. They integrated linked data with friend-of-a-friend (FOAF) vocabulary to enhance the semantics of data that are extracted from the web [16].

B. Personalization

Personalization can be defined according to [17] as the process of customizing content with respect to users' needs and preferences to enhance user experience. Personalization is the basic foundation of several studies including web content personalization [18], recommendation systems [19], and personalizing social media pages [20].

Personalization mainly depends on the construction of user profiles to get insights of what users may like or dislike. The information included in user profiles can be gathered from different sources such as log files and human resources indicating both implicit and explicit feedback [21].

Several studies have considered interaction attributes as implicit feedback of users' preferences. Bhavithra and Saradha [22] proposed a case-based reasoning strategy to recommend web pages based on the searching history of a particular user. They considered several interaction factors to be added in a user profile such as time on page, time on site, exit rate, and others. Their main aim was to benefit from these attributes to recognize patterns and apply collaborative filtering [22]. Moreover, Stai et al. [21] developed a mechanism to effectively personalize the enriched multimedia content based on users' interests and needs. They considered some interaction attributes to infer users' preferences such as "share video on social media, click on enrichment, click on ads, and playtime of a main video." Singh and Sharma [23] developed a multi-agent context-aware framework to personalize the web. They designed a dynamic user profiling technique to keep track of changes in users' behavior, which influences their interests.

Nabil et al. [24] proposed a sentiment-aware approach for article recommendation systems. They used consumers' reviews to detect feelings and infer preferences. They integrated both content-based and collaborative-based approaches to develop a hybrid recommendation system. Furthermore, a sentiment factor was considered by [25] to detect spammers. They found in their exploratory study that there were substantial differences between sentiments of spammers and sentiments of normal users. Hu et al. [25] incorporated a sentiment factor to a spammers detection framework to enhance the detection rate using sentiment analysis. A recent work by [26] proposed a protocol-based architecture model to predict direct and indirect interests of a user using the semantic relatedness concept.

III. An Overview of the Proposed SentiFilter Model

The proposed SentiFilter model was designed to utilize textual-based human emotional feedback through sentiment analysis to detect Arabic semi-spam content for a particular user. The results of the SentiFilter model are assessed through comparing the impact of liking behavior, commenting behavior, and the combination of liking behavior with sentiment of users' comments in effectively detecting semi-spam content for individuals.

To the best of our knowledge, the proposed SentiFilter model is the first to combine a sentimental factor of users' comments with other behavioral factors to detect semi-spam content for individuals. In our work, three factors are considered to model the proposed filter. Those are defined as follow:

- Liking behavior is a user's act that reflects liking reaction to a message, and it occurs when a user clicks on the like button.

- Commenting behavior is a reply act to a specific message.
- Sentiment factor represents the textual opinion/point of view of a user about a topic. It can be negative, positive, or neutral.

The SentiFilter model consists of three components and two classifiers. The three components are Extractor, Data preprocessor, and Detector. Each of the three components contains several functional modules. They are implemented as several python and R scripts. The two classifiers are a sentiment-based classifier, which is the core component in the SentiFilter model, and a list-based classifier, which is a data mining rule-based classifier. An overview of the structure of the SentiFilter model is shown in Fig. 1.

- Extractor: The extractor component is responsible for fetching and collecting implicit information about users' reactions to construct a user profile for each user. The Extractor consists of four modules. The workflow of the SentiFilter model starts by collecting the timeline of the incoming data stream from the user's social space, using the user's timeline Extractor module. Then, the user's comment Extractor module extracts all replies or comments from the user's timeline and prepares them to be passed to the data preprocessor component. All messages that a user has liked are collected, using the user's Like Extractor module. Each extracted comment is associated with its original message using the original message's Extractor module.
- Data Preprocessor: This component is responsible for cleaning Arabic text that was extracted, such as comments and tweets from the Extractor component. It involves removing meaningless and stop words.
- Arabic Topic Detector: This module is responsible for discovering the domain, subject, or topic that represents a particular message. Each domain *d* is represented by a set of Arabic keywords *T*, where $T_d = \{w_1, w_2, w_3, \ldots, w_n\}$. Arabic keywords for each topic are specified by crawling OSN messages of certain hashtags representing that domain. Then a preprocessing and tokenization over the collected set is carried out to extract the most frequent terms appearing in these messages. We propose to use hashtags, key phrases, and keywords to determine the topic of a particular message.
- Sentiment-based Classifier: This module is responsible for analyzing messages that are replies to other messages, with the goal to infer a user's attitude toward the original messages. The sentiment classification module is a predictive model that uses a supervised machine learning classification algorithm [27] to predict the polarity of a comment by examining its text.

The Term Frequency feature engineering method TF-IDF [28] was used to create a lexicon-based dictionary to identify positive, negative, and neutral keywords. This method aims at finding the most frequent words in a document for search purposes. The comments containing those keywords are manually labeled as positive, negative, or neutral messages. The labeled keywords are used by human annotators to train the sentiment classifier. The outcome of this module is a set of labeled messages that are passed to the next module. The workflow of this module is illustrated in Fig. 2.

- List-based Classifier: The Sentiment Classifier module and user's likes Extractor module work simultaneously to generate a rule-based classification model using a proposed personalized aggregate factorization (PAF) algorithm as shown in Fig. 3.

The proposed algorithm takes into consideration all previously mentioned factors to produce a personalized blacklist, whitelist, and greylist for each individual. There are seven cases in this algorithm, considering that I_m is the interaction behavior of a message m (i.e., a user likes a message), R_m is a reply to a message m, and d_m is a topic of a message m:

- Case 1: if I_m is null AND R_m is positive, THEN Whitelist (d_m).
- Case 2: if I_m is null AND R_m is neutral, THEN Greylist (d_m).
- Case 3: if I_m is null AND R_m is negative, THEN Blacklist (d_m).
- Case 4: if I_m is not null AND R_m is positive, THEN Whitelist (d_m).
- Case 5: if I_m is not null AND R_m is neutral, THEN Whitelist (d_m).
- Case 6: if I_m is not null AND R_m is negative, THEN Greylist (d_m).
- Case 7: if I_m is null AND R_m is null, THEN Blacklist (d_m).

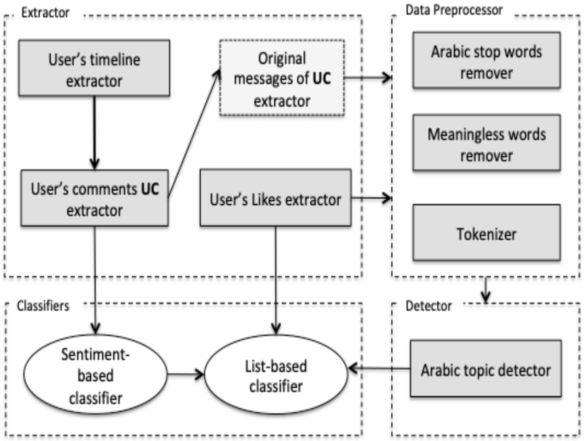

Extractor
User's timeline extractor
Original messages of UC extractor
User's comments UC extractor
User's Likes extractor
Data Preprocessor
Arabic stop words remover
Meaningless words remover
Tokenizer
Classifiers
Sentiment-based classifier
List-based classifier
Detector
Arabic topic detector

Fig. 1. The Structure of the Senti Filter Model.

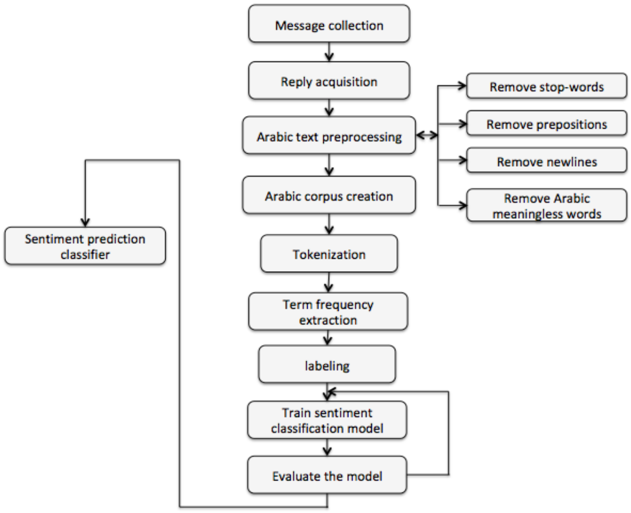

Message collection
Reply acquisition
Arabic text preprocessing
Remove stop-words
Remove prepositions
Remove newlines
Remove Arabic meaningless words
Arabic corpus creation
Sentiment prediction classifier
Tokenization
Term frequency extraction
labeling
Train sentiment classification model
Evaluate the model

Fig. 2. The Workflow of the Sentiment-based Classifier.

Algorithm: Personalized Aggregate Factorization PAF algorithm

Input:
I_m is an interaction behavior of a message *m*, ***R_m*** is a reply to a message *m*, ***count*** is the number of times that a topic has been classified into a list, ***thres*** is a predefined threshold score that determines if a topic is definitely belongs to a particular list, ***sent*** is a sentiment polarity of a reply R for a message m, ***B*** is a blacklist for user *u*, ***W*** is a whitelist for user *u*, and ***G*** is a greylist for user *u*.
Output: classifying a message m into a list (labeling a message)
Procedure:

```
for each m in user social space do
{
    Perform a topic detection for m → $d_m$
    for all d in user profile do
        if ($d_m$ is found) then
            if (count ($d_m$) < thres) then
                monitor()
            else monitor()
} // End for
void monitor()
{
    count($d_m$) ++
    if ($I_m$ is NULL AND $R_m$ is Null) then
        add $d_m$ to B
    else if ($I_m$ is not NULL AND $R_m$ is Null) then
        add $d_m$ to G
    else
        if ($I_m$ is not NULL AND $R_m$ is not Null) then
            if (sent($R_m$) is negative) then
                add $d_m$ to G
            else add $d_m$ to W
        else if ($I_m$ is NULL and $R_m$ is not Null) then
            if (sent($R_m$) is positive) then
                add $d_m$ to W
            else if (sent($R_m$) is negative) then
                add $d_m$ to B
            else add $d_m$ to G
} // End of monitor
```

Fig. 3. The Proposed Personalized Aggregate Factorization Algorithm (PAF).

IV. Experiment Design

An experiment was conducted to empirically evaluate the impact of combining a sentimental factor with users' liking behavior to improve the filtering process of semi-spam content in terms of the agreement between the implicit feedback derived from the SentiFilter model and users' explicit feedback.

The experiment consisted of two main phases: 1) quantitative analysis of users' behavioral and sentimental factors, and 2) an analysis of users' explicit feedback towards certain topics using an online user survey instrument.

The aim of the first phase was to collect and analyze users' liking and commenting behaviors and to detect sentiment of users' comments. The objective of this phase was to examine the effectiveness of the proposed sentiment-based classifier by considering its accuracy using standard machine learning classification algorithms to select the most accurate one to be used in our analysis. In the second phase, a user survey was created asking users to rate several testing tweets to collect users' explicit feedback. The users' explicit feedback was used as a measurement to determine which behavioral and sentimental factors are closer to users' expectations.

A. Data Collection and Twitter API

Initially, a total of 80,098 Arabic tweets were collected using a Twitter API [29] for 32 users. The users in our sample were selected based on their number of tweets, comments, and likes to ensure their active status on Twitter. The mother language of all users was Arabic. They were selected from different backgrounds and interests. The gender distribution of the selected sample consisted of 21 females (65.6%) and 11 males (34.3%). The average number of comments for men was 52%, while it was 47.9% for women. The average number of likes for women was 51.75%, while it was 48.24% for men. The timeline of each user was crawled, including posts that he/she created or posts that were comments to other posts. Since the focus of this work is on personalization, we chose to collect timelines of users instead of collecting tweets for certain chosen topics and to select users who interacted with these topics as done in [4] and [20].

Furthermore, tweets that a user has liked were crawled in the same period of time. In order to demonstrate our methodology, Twitter is selected as a source of our datasets because Twitter API [29] is a freely available API for developers who wish to explore with real time data, unlike other OSN services such as WhatsApp messenger that have strict privacy constraints that will not enable us to collect users' conversations and their activities.

B. Sentiment Analysis

Inferring users' emotional feedback from their text-based comments is a critical task to determine individually semi-spam content. After collecting the 80,098 Arabic tweets, a reply acquisition process was performed to extract comments and exclude self-posts. We ended up with 6,307 comments with an average of 197.09 comments per user. Then, the following steps listed in the next subsections were performed:

1) Data preprocessing: Preprocessing Arabic text is a challenging task because of the sophisticated structure of the Arabic language. The SentiFilter model performed data preprocessing that included removing stop-words and Arabic prepositions. Meaningless Arabic words that did not add any meaning to the context of the text were removed. Some examples of those words are shown in Table I.

2) Tokenization: Basically, an Arabic corpus was created for each user representing his/her comments after the preprocessing phase. Then, each comment was tokenized to several tokens. A TF-IDF method was selected to extract terms that were used to label tweets. Human annotators were used to label tweets as positive, negative, or neutral.

3) Annotation and predictive classifier: The predictive classifier in this work is a sentiment-based classifier that uses supervised machine learning classification algorithms to predict the sentiment polarity of a comment and then produces a pair consisting of the sentiment polarity of the comment and the topic of its original tweet.

A support vector machine (SVM) supervised machine learning algorithm [30] was selected based on the performance analysis. It used to perform a multi-class classification and classify replies as positive, negative, or neutral.

In the training phase, we listed some terms that appeared in users' comments after the tokenization process and grouped them to three categories: positive, negative, and neutral. The two annotators were asked to use these categories to label comments.

Each reply was classified as positive, negative, or neutral based on the sentiment polarity of its words that were derived from the three categories.

C. 4.3. Arabic Topic Detection Model

This model is a rule-based model responsible for detecting the main topic for each extracted tweet. We defined five general topics, and for each topic, we created an Arabic domain-specific dictionary that contains the 20 most used keywords in that topic. We collected these keywords by crawling Arabic tweets using the Twitter API for hashtags related to that topic and filtered them manually. We then performed the same preprocessing and tokenization to get the 20 most frequent terms in each defined topic. Table II shows examples of keywords for each topic.

TABLE. I. EXAMPLES OF ARABIC MEANINGLESS WORDS

Arabic meaningless words	Transliterated words	Corresponding English words
من	mn	From
يا	Ya	Oh
على	3la	On
في	Fy	In/inside
ما	Ma	What
مع	M3	With
الف	alf	A thousand

TABLE. II. EXAMPLES OF KEYWORDS USED IN THE PROPOSED ARABIC TOPIC DETECTION MODEL

Topics (in English)	Topics (in Arabic)	Examples of keywords (in Arabic)	Examples of keywords (in English)
T1: Health	صحة	صحية – التهاب –طبية الاطباء – غذائية - الجراحي	i.e., (Health; medical; inflammation; surgery; your doctor)
T2: Sport	رياضـ ة	الدوري – مباراة – الاتحاد – الاهلي – الهلال – دوري المحترفين	i.e., (League; match; (names of football teams); professional league)
T3 : Technology	تقنية	انترنت الاشياء – الذكاء الاصطناعي – تقنية – علم البيانات- تكنولوجيا	i.e., (IoT; AI; technology; data science; computer)
T4: Politics	سياسة	وزارة الداخلية – وزارة الخارجية – داعش – ايران - المرابطين	i.e., (Politics; Urdu; interior Minister; Minister of Foreign Affair; Nuclear power; armed forces)
T5: Social content	محتو ى اجتماعي	مبروك – مبارك – شكرا – الشكر – عظم الله -	i.e., (Thanks, thank you, congrats; congratulation; you deserve it)

D. Interaction-based Detection

Interaction factors were defined by the mean of liking and commenting behaviors. After collecting users' timelines, 66,576 tweets that users liked with an average of 2,080.5 tweets per user were collected. Then, the same preprocessing steps to the text of these tweets were performed.

E. User Survey

An online user survey instrument was developed and used as a web-based application. The goal of the survey was to collect users' explicit feedback towards topics considered in the evaluation of the proposed filtering model.

The survey was distributed among the same 32 users who we considered in the first phase of the study. The survey consisted of two parts. In the first part, several tweets representing the five topics described in Table II were displayed and shown to each user as a Likert scale-based question. Each user was asked to rate each of the shown tweets based on how likely he/she would be to like/reply to/re-tweet them if they appeared in their social space. A scale from 1 to 5 was presented where 1 represented the not likely attitude, and 5 represented the extremely likely attitude. In the second part, users were asked to order the five topics based on their interests from the most interesting topics to them to their least-preferred topics. A user–topic factorization matrix of users' ratings was constructed to evaluate the effectiveness of the SentiFilter model in terms of the agreement of the implicit feedback derived from it with users' explicit feedback.

V. RESULTS AND DISCUSSION

The results obtained from the two phases of the experiment are demonstrated in two forms: the effectiveness of the sentiment-based classifier, and the comparison between the implicit feedbacks derived from of the SentiFilter model against the users' explicit feedback. In order to demonstrate our results, Table III shows statistical details about our datasets.

TABLE. III. DATASETS DETAILS

Number of crawled tweets	80,098
Number of users	32
Average number of tweets per user	2,503.063
Total number of comments	6,307
Average number of comments per user	197.0938
Number of likes	66,576
Average number of likes per user	2,080.5
Number of positive comments	610
Average number of positive per user	19.06
Number of negative comments	234
Average number of negative per user	7.3
Number of neutral comments	5,457
Average number of neutral per user	170.53
Number of intersected tweets between users' comments and likes	595
Average number of intersected tweets between users' comments and likes per user	18.59

The datasets distribution for the sentiment-based classifier is illustrated in Fig. 4, where each dataset represents a user.

The results from Fig. 4 revealed that most of the users' comments were neutral. Our definition of neutral comments is those comments that do not have any positive or negative keywords. The majority of neutral comments were basically replies such as personal congratulations, thanks, and good wishes. In the SentiFilter model, only 5% of the total comments were detected as negative comments, which indicates that negative comments are rarely posted in user social space in our datasets sample unless the content of the original tweet was negative.

We evaluated the effectiveness of the sentiment-based classifier using well-known performance measures for classification, which are accuracy, precision, and recall. Accuracy is a ratio of correct classified comments to the total number of comments of a user, while precision is a measure that determines how precise our model is in terms of the percentage of correct classified comments. On the other hand, recall is a ratio of total classified comments to total comments of a user [31]. An SVM classifier was selected as a base algorithm for our sentiment classifier because it produced the best results among other algorithms on average with an average accuracy of 90.89% as shown in Fig. 5.

From the statistical analysis of our collected datasets shown in Table III, the results indicated that 10.4% of liked tweets had positive comments and 87.5% of them had neutral comments, while only 2% of the liked tweets had negative comments. Thus, sentiment polarity of users' comments as a stand-alone factor cannot be used to reflect users' perception of semi-spam content, which means that negative comments do not indicate disliking attitude towards the topic under discussion. However, there is a relationship between commenting on a topic and detecting semi-spam content by the consideration of the silent reaction that was represented as case 7 in our proposed PAF algorithm.

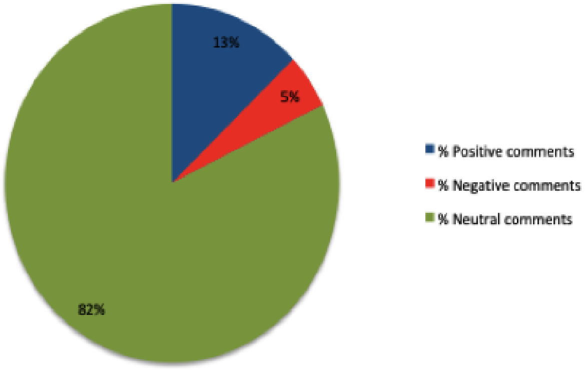


Fig. 4. The Datasets Distribution for the Sentiment-based Classifier.

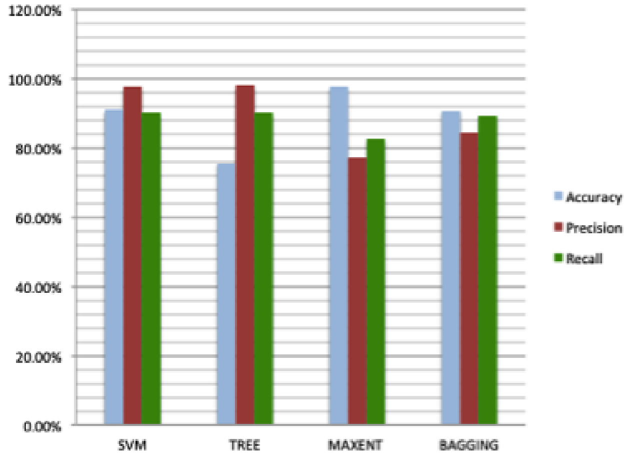


Fig. 5. The Classification Results of Sentiment-based Spam Classifiers.

The second phase of the evaluation was aimed to assess the effectiveness of the SentiFilter model in terms of the agreement between users' explicit feedback and the implicit feedback derived from behavioral and sentiment analysis. In this respect, a total of 30 users completed the survey. Users' explicit feedback was compared to the implicit feedback derived from liking behavior, commenting behavior, and the combination of sentimental factor with liking behavior by the mean indication. The statistical significance difference between each behavioral and sentiment factor with users' explicit feedback was computed and compared as shown in Table IV.

Results from Table IV show that there is no statistical significance difference between the implicit feedback derived from commenting behavior and users' explicit feedback (only 0.02 difference was found), which indicates that commenting behavior is the most effective behavioral indicator to significantly detect semi-spam content for an individual. On the other hand, the results of combining sentimental factor with liking behavior in the SentiFilter model are more effective to detect semi-spam content than considering liking behavior alone because the implicit feedback derived from the proposed model produces a high agreement with users' explicit feedback by the mean indication shown in Table IV.

TABLE. IV. THE STATISTICAL SIGNIFICANCE DIFFERENCES BETWEEN EACH FACTOR AND USERS' EXPLICIT FEEDBACK USING MEAN INDICATION (THE MEAN VALUE OF USERS' EXPLICIT FEEDBACK = 0.40)

Factor	Mean	Sig. Diff
Liking behavior	0.80	0.39
Commenting behavior	0.38	0.02
Liking behavior + Sentiment factor (SentiFilter model)	0.65	0.24

We measured the effectiveness of each factor by comparing the implicit feedback derived from each factor against users' explicit feedback by finding the significance difference between their mean values. The comparative analysis shown in Fig. 6 shows that there is only 0.24 statistical significance difference between implicit feedback derived from the SentiFilter model and users' explicit feedback, which indicates that the SentiFilter model accurately agrees with users' explicit feedback. However, implicit feedback derived from commenting behavior reports no statistical significance difference (only 0.02 difference) with users' explicit feedback.

The results also reveal that combining sentiment of users' comments with behavioral factors is a positive indicator to infer users' preferences, while negative comments do not reflect users' attitude towards semi-spam content.

The results and performance of the SentiFilter model might be varied if the knowledge is increased regarding the collected datasets.

Another constraint that can be encountered in the SentiFilter model is the social implication of using OSN services. We believe that users interact differently on different OSN services towards the same situations. For example, some users reply to others for the sake of courtesy in formal OSN services such as Twitter, while their reactions could be different if they were using an informal OSN service such as WhatsApp. Therefore, the SentiFilter model might report different findings when another OSN service is considered.

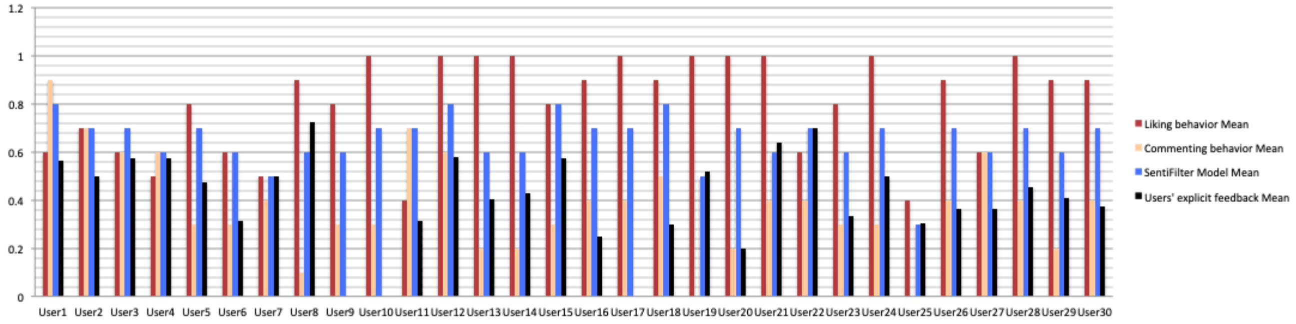


Fig. 6. Comparative Analysis of Sentiment and behavioral Factors with users' Explicit Feedback.

VI. CONCLUSION AND FUTURE WORKS

Personalization is a fundamental task in most aspects of our daily life. It provides users with a better experience and gives them more control over what they really want. Employing this concept in spam detection systems can help in enhancing users' browsing experience by automatically generating personalized filters that are able to recognize and control what users see or receive in their social spaces. Thus, this paper introduces a new factor that can be combined with other behavioral factors to be used as an indicator to detect semi-spam messages, which is the sentiment factor. The proposed SentiFilter model empirically assesses the impact of combining the sentimental factor with behavioral factors to filter semi-spam messages. Our results showed that commenting behavior is more effective than other behavioral factors in detecting semi-spam content by the high agreement between the implicit feedback derived from it and users' explicit feedback.

We have evaluated the effectiveness of the SentiFilter model in terms of the agreement between its implicit feedback and users' explicit feedback. The implicit feedback derived from the SentiFilter model accurately agrees with users' explicit feedback by the indication of the statistical significance difference between the two sets.

In our future work, we will concentrate on overcoming the knowledge limitation by increasing the sample size to produce more general results. We will apply the same model to English messages to find out if any different observations can be drawn. We will also work on designing and developing blocking mechanisms that maintain users' blacklists in effective ways. Furthermore, we will plan to identify some platform-specific behaviors that differentiate how users perceive semi-spam content. We encourage researchers in both HCI and information security fields to incorporate our findings in their design decisions to effectively maintain users' blacklists.

REFERENCES

[1] M. M. Alsulami and A. Y. Al-Aama, "Exploring User's Perception of Storage Management Features in Instant Messaging Applications: A Case on WhatsApp Messenger," Proc. 2019 2nd Int. Conf. Comput. Appl. Inf. Secur., pp. 1–6.

[2] K. S. Bajaj, "A multi-layer model to detect spam email at client side," Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng., vol. 198 LNICST, pp. 334–349, 2017.

[3] X. Liu et al., "CPSFS: A credible personalized spam filtering scheme by crowdsourcing," Wirel. Commun. Mob. Comput., vol. 2017, 2017.

[4] R. Harakawa, D. Takehara, T. Ogawa, and M. Haseyama, "Sentiment-aware personalized tweet recommendation through multimodal FFM," Multimed. Tools Appl., pp. 18741–18759, 2018.

[5] M. M. Alsulami and R. Mehmood, "Sentiment Analysis Model for Arabic Tweets to Detect Users' Opinions about Government Services in

Saudi Arabia: Ministry of Education as a case study," ALYamamah Inf. Commun. Technol. Conf.

[6] B. K. Narayanan, R. B. M, S. M. J, and M. Nirmala, "Adult content filtering : Restricting minor audience from accessing inappropriate internet content," Educ Inf Technol, 2018.

[7] G. Karyono, A. Ahmad, and S. A. Asmai, "Survey on Nudity Detection : Opportunities and Challenges based on ' Awrah Concept in Islamic Shari ' A," J Ther Appl Inf Technolo, vol. 95, no. 15, pp. 3450–3460, 2017.

[8] F. Aiwan and Y. Zhaofeng, "Image spam filtering using convolutional neural networks," Pers Ubiquit Comput, no. 22, pp. 5–6, 2018.

[9] H. Mubarak and K. Darwish, "Abusive Language Detection on Arabic Social Media," Proc. First Work. Abus. Lang. Online, pp. 52–56, 2017.

[10] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean Birds: Detecting Aggression and Bullying on Twitter," Proc. 2017 ACM Web Sci. Conf., pp. 13–22.

[11] C. Cao and J. Caverlee, "Detecting Spam URLs in Social Media via Behavioral Analysis," Eur. Conf. Inf. Retr., pp. 703–714.

[12] A. Tewari and S. Jangale, "Spam Filtering Methods and machine Learning Algorithm - A Survey," Int. J. Comput. Appl., vol. 154, no. 6, pp. 8–12, 2016.

[13] B. D. O'Connor, "System And Method For Detecting Unwanted Content," US9948588B2, 2018.

[14] T. G. and V. J. Renushree Bodkhe, "A Novel Methodology to Filter Out Unwanted Messages from OSN User's Wall Using Trust Value Calculation," Proc. Second Int. Conf. Comput. Commun. Technol., pp. 755–764, 2016.

[15] Z. Yan, "PSNController- An Unwanted Content Control System in Pervasive Social Networking Based on Trust Management," ACM Trans Multimed Comput Commun Appl, vol. 12, no. 1, p. 17, 2015.

[16] H. Zitouni, S. Meshoul, and K. Taouche, Enhancing Content Based Filtering Using Web of Data, vol. 1. Springer International Publishing, 2018.

[17] P. Germanakos and M. Belk, Human- Centred Web Adaptation and Personalization From Theory to Practice. Switzerland: Springer International Publishing, 2016.

[18] S. Ferretti, S. Mirri, C. Prandi, and P. Salomoni, "The Journal of Systems and Software Automatic web content personalization through reinforcement learning," J. Syst. Softw., vol. 121, pp. 157–169, 2016.

[19] D. F. Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti, "iSCUR : Interest and Sentiment-Based Community Detection for User Recommendation on Twitter," Switz. Springer Int. Publ., pp. 314–319, 2014.

[20] U. K. Wiil, "Emotion-Based Content Personalization in Social Networks," vol. 42, no. 1, pp. 1–16, 2018.

[21] E. Stai, S. Kafetzoglou, E. E. Tsiropoulou, and S. Papavassiliou, "A holistic approach for personalization, relevance feedback & recommendation in enriched multimedia content," Multimed. Tools Appl., vol. 77, no. 1, pp. 283–326, 2018.

[22] J. Bhavithra and A. Saradha, "Personalized web page recommendation using case-based clustering and weighted association rule mining," Cluster Comput., vol. 0123456789, pp. 1–12, 2018.

[23] A. Singh and A. Sharma, A Multi-agent Framework for Context-Aware Dynamic User Profiling for Web Personalization. Springer Nature Singapore Pte Ltd. 2019.

[24] S. Nabil, J. Elbouhdidi, and M. Yassin, "Recommendation system based on data analysis- Application on tweets sentiment analysis," 2018 IEEE 5th Int. Congr. Inf. Sci. Technol., pp. 155–160, 2018.

[25] X. Hu, J. Tang, H. Gao, and H. Liu, "Social Spammer Detection with Sentiment Information," IEEE Access, 2014.

[26] S. Goel and R. Kumar, "SoTaRePo: Society-Tag Relationship Protocol based architecture for UIP construction," Expert Syst. Appl., vol. 141, p. 112955, 2020.

[27] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," Heliyon, vol. 5, no. 6, 2019.

[28] J. Guo, Y. Mu, M. Xiong, Y. Liu, J. Gu, and J. Garcia-Rodriguez, "Activity Feature Solving Based on TF-IDF for Activity Recognition in Smart Homes," Complexity, vol. 2019, 2019.

[29] "TwitterAPI Documentation," 2017.

[30] P. W. Wang and C. J. Lin, "Support vector machines," Data Classif. Algorithms Appl., no. ii, pp. 187–204, 2014.

[31] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, "Precision and recall for time series," Adv. Neural Inf. Process. Syst., vol. 2018-Decem, no. NeurIPS, pp. 1920–1930, 2018.

Conceptual Framework for Developing an ERP Module for Quality Management and Academic Accreditation at Higher Education Institutions: The Case of Saudi Arabia

Mohammad Samir Abdel-Haq
College of Business / Management Information System Department
Dar Al Uloom University, Riyadh, Kingdom of Saudi Arabia

Abstract—**As a result of the high priority given by universities in Saudi Arabia to the implementation of quality systems and achieve international and local academic accreditation, especially NCAAA, as well as, academic accreditation standards require the provision of a set of data, documents, reports and evidence distributed among the various departments of the university (academic and non-academic), which must be provided periodically and annually. These universities need to provide an integrated system between these departments to cover the requirements of quality and academic accreditation. On the other hand, there are ERP systems that suit the organizational environment of the university, but these systems did not implement a module to cover quality assurance requirements or academic accreditation. This study proposed a framework includes ERP module for quality requirements and academic accreditation to facilitate the collection of required data from various sources within the university and helps to provide the necessary reports and statistics, where the module organizes the necessary processes for quality and accreditation by providing a special database and through linking with the other subsystems at the university.**

Keywords—Quality assurance; academic accreditation framework; ERP module

I. INTRODUCTION

Enterprise resource planning (ERP) systems are widely used by public and private organization. ERP consists of integrated application software modules aims to integrate all business processes and functions in a central database to improve productivity, quality and competitiveness for the organization, in addition to the ability of the ERP system to deal with a rapidly changing environment and overcome the limitations of legacy systems [1]. Recently, universities around the world have begun to move to ERP instead of having multiple management and organizational information systems at the university that may not be well interconnected, hindering inter-departmental processes [2]. On the other hand, current information systems may not provide strategic or historical analysis about students, courses and staff [3]. ERP system for universities is defined as “an information technology solution that integrates and automates recruitment, admissions, financial aid, student records, and most academic and administrative services” [4].

There are many reasons encourage the universities to adopt ERP system, such as global trends and government directives for universities to create an effective learning environment and to improve their performance and efficiency [5] especially in the Kingdom of Saudi Arabia, one of the most important trend which increase the demand of adoption ERP system is to meet the quality requirements and the academic accreditation standards (institutional or academic program level) of the Education Evaluation Commission - National Center for Academic Accreditation and Assessment (NCAAA), which aims to meet the expectations of stakeholders such as students, government and the labor market by developing competitive educational environments in those universities and ensuring continuous improvement in them. Quality assurance processes in HEI in KSA should cover all sectors of the institution to assure the effective integration with the administrative, planning processes and produce database and evidence-based performance indicator relating to NCAAA standards [6].

The accreditation is not easy process from the perspective of most faculty members as they have the misconception that it is an exhausting time consuming, not to say complex and unnecessary process [7], and there are some challenges, such as how to collect, aggregate, analyze data and collect evidences, then developing corrective action and future plans as needed based on NCAAA accreditation standards such as Strategic Planning indicators, Governance Systems, Teaching and Learning process, Student Records & achievements, Faculty & Staff Records, Institutional Resources management (Financial Resources, Information Technology, Facilities and Equipment and Safety and Risk Management), in addition to Research and Innovation and Community Partnership [7].

Little research has been conducted on ERP implementations in universities compared to other environments [1] [5]. Previous research has addressed the characteristics and components of ERP systems that are appropriate to the environment of educational organizations, especially universities, such as the use of ERP systems in the process of learning, teaching and classroom organization [8].

ERP systems vendors for HEIs such as Oracle, SAP, People Soft and Jenzabar did not implement a module to cover quality assurance requirements or academic accreditation [9], so these systems need to contain a special module to meet the requirements of quality and academic accreditation, including plans, reports, performance indicators and a survey of academic and non-academic services at the university, the difficulty in collecting such data hinders quality processes and delays academic accreditation and negatively affects the performance of the quality team and faculty to meet those requirements [6] [7]. The existence of a ERP module for quality requirements and academic accreditation greatly facilitates the collection of this data from various sources within the university and helps to provide the necessary reports and statistics for workers on this type of projects, where the module organizes the necessary processes for quality and accreditation by providing a special database and through linking with the other subsystems at the university.

However, quality and accreditation requirements call for research topics that attempt to add improvement to ERP systems to consider the collection, processing, storage and retrieval of data related to evidence, performance indicators, user surveys on services, learning resources, student results, statistics related to scientific research and community service, as well as follow-up implementation. Administrative and academic processes in line with the requirements of quality and academic accreditation, especially as these data are distributed across several departments in the university, and need an integrated system such as ERP system works to link these processes with each other through the development of a new module to fulfill the requirements of quality and academic accreditation processes that help decision makers at the university to study the strengths, weaknesses, opportunities and threats to build improvements plans and follow-up implementation.

This study proposes a conceptual framework to merge a new module to handle quality requirements and accreditation process as an attempt to enrich research topics in ERP in the higher education sector especially in Saudi Arabia. This study will review the modules of the ERP system that are suitable for the university environment to determine the common ERP Modules between ERP providers that are required for academic and administrative processes. This study also proposes a module for quality and academic accreditation Called QAAM (Quality Assurance and Accreditation Module) and linking this module with common ERP modules to help the university meet the standards imposed by the accreditation bodies.

II. Research Background

A. ERP Systems in Higher Education Institutions

ERP “is an information system software that aims to integrate all business processes and functions in a central database. This boosts the management of business resources (finance, production, human resource, materials, etc.)” [10]. Higher education institutions as one of important sectors around the worlds have been influenced by global trends to adopt new technologies aim to improve the performance and efficiency [1]. Universities have relied heavily on information systems for several years, while increased attention has been given to ERP systems to cover organizational activities and processes and ensure their integration, with the aim of achieving significant cost savings as well as improving performance and results and an opportunity to update procedures and align them with perceived 'best practices' [2]. Integrated information system such as ERP helps the university to provide high quality services to the stakeholders measured in terms of ease of access, complete coverage of all needs and availability of information, allow university context to be more flexible to match the continuous evolutions in higher education sector. On the other hand, University ERP system provides the instruments to support the governance processes, providing the data and analysis necessary for strategic planning and control which enable the strong development area for the university to represent high delivery of services for students, teachers, researchers and community demands [3]. University ERP systems can provide academic and non-academic entities including colleges and departments with completely functional applications [1] [11] which enable users to access students’ information, academic records, and other data needed to complete their daily work [12], this accessibility and availability will improve business processes and services provided to the faculty, students, and employees [13].

In Kingdom of Saudi Arabia there are (43) universities, (29) government universities and (14) private universities are sited on the ministry of education website (www.moe.gov.sa) three of them have an experience in ERP systems (1) King Abdul Aziz University, ERP System is ODES plus. (2) King Saud University, ERP System is MADAR. (3) King Fahd University, ERP System is Moraslat. These three systems cover the following administrative sectors: (1) Human resource. (2) Financial management. (3) Procurement management. (4) Where house (store) management. (5) Student registration management. (6) Library management [14]. The most known ERP vendors in Saudi Arabia Saudi Arabia are SAP, Oracle, PeopleSoft and JD Edward [14].

Table I shows the most important requirements of what (functional requirements) higher education ERP system should do, or how will do (non-functional requirements: performance characteristic of the system).

B. Quality and Academic Accreditation in HEIs in Saudi Arabia

In recent decades, governments in various countries have called on institutions of higher education to pay more attention to the quality of education and to improve learning outcomes, to develop the country's social and economic competencies and promote them among nations [15]. The International Network for Quality Assurance Agencies in Higher Education (INQAAHE) considered the quality assurance in HEIs as a process of establishing stakeholder confidence that provision fulfils expectations or measures up to threshold minimum requirements (Good Practice), which embraces input, process and outcomes [16]. In general, quality in higher education is a continuous development process based on quality policies, procedures to achieve the institution mission, values and stakeholder needs and expectations [17]. Quality assurance and Accreditation in HEIs is considered as one of the major processes in the development of higher education which is connected to the components of the educational process such as (students, teaching staff, programs, teaching methods, facilities and equipment, etc.) [18]. Accreditation is defined as "a process whereby officially appointed external regulatory bodies, accountable at government level, evaluate educational institutions sing established criteria, standards and procedures" [19]. The National Center for Academic accreditation and Assessment (NCAAA) in Saudi Arabia is established to improve the quality of higher education and to help universities to achieve international quality standards in the education sector in the provision of high-quality teaching, learning, research and community service [20]. All universities and academic programs in Saudi Arabia should achieve NCAAA academic accreditation [19] through well-defined processes based on quality standards and set of KPIs. Higher education institutions in Saudi Arabia are required to establish internal quality assurance systems that ensure high levels of quality across eight standards at institutional level and six standards at academic program level (NCAAA, Standards for Institutional Accreditation, 2018, Standards for Program Accreditation, 2018). Developing internal quality system in higher education in the Kingdom of Saudi Arabia is considered as one of the main challenges to HEIs [21]. Quality in universities and higher education institutes like any other systems has three main dimensions; inputs, the process and the outputs, also includes the interactions with the actors and the other systems to achieve the core functions and producing the major outcomes or outputs of teaching/ learning (graduates), research, and community service as shown in Fig. 1 [21].

TABLE. I. MOST IMPORTANT REQUIREMENTS FOR THE HIGHER EDUCATION ERP SYSTEM. SOURCE: [14] [23]

Functional Requirements		Non- Functional Requirements	
1	Institution profile	1	Accessibility
2	Staff profile	2	Documentation
3	Student profile	3	Efficiency
4	Curriculum	4	Effectiveness
5	Performance analysis	5	Extensibility
6	Attendance (Staff / Students)	6	Fault tolerance
7	Online examination	7	Interoperability
8	Online assignment	8	Privacy
9	Admission	9	Quality of Output
10	Academic advising	10	Response time
11	Timetable	11	Scalability
12	Internal messaging	12	Security
13	Alumni management	13	Stability
14	Library management	14	Supportability
15	Payroll	15	Testability
16	Accounting		
17	Fees management		
18	Asset Management		
19	Ad hock reporting		
20	Job / requirements analysis		
21	Labor market demand		
22	Education / labor market observatory		

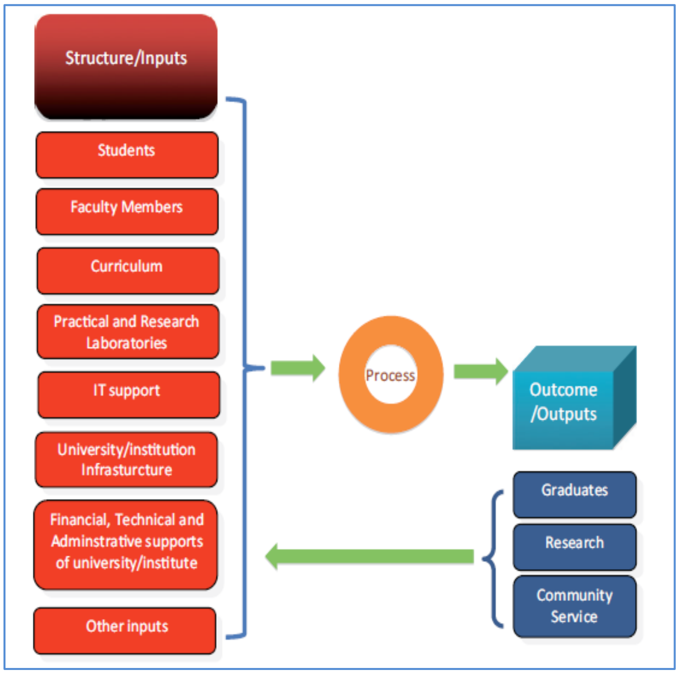

Structure/Inputs
Students
Faculty Members
Curriculum
Practical and Research Laboratories
IT support
University/institution Infrasturcture
Financial, Technical and Adminstrative supports of university/institute
Other inputs
Process
Outcome /Outputs
Graduates
Research
Community Service

Fig. 1. Dimensions of Quality in Higher Education. Source: [21].

To develop the quality system and implement its core processes as shown in Fig. 1, the university administration must adhere to quality policies including set of processes (quality planning, quality assurance and quality control) and the active participation of all faculty, administrative staff and students. This requires the provision of financial and technical support [21], by the provision of an information system that assists all actors of the quality system in collecting, processing, storing and retrieving data to obtain performance indicator data and to produce periodic reports to meet quality and academic accreditation requirements, Academic accreditation requires data collection on various aspects of the institution and decision-making in compliance with standards [19]. The accessibility to data, information and records related to the quality processes and requirements for all staff through the most effective channels are very important [22]. The main problems of implementing quality system in HEIs are related to collect accurate information from several sources, working with huge information and documents, then analyzing this information to generate reports and KPIs results [9]. The implementation of information systems such as (ERP) will support the integration of quality assurance processes in HEIs and produce assessable information to meet quality requirements. ERP systems vendors for HEIs such as Oracle, SAP, People Soft and Jenzabar do not develop an ERP solution to cover all HEIs demands [9], mainly quality assurance requirements or even academic accreditation. The literature review supports the objectives of this research, as these literature reflect the widespread use of ERP systems in higher education institutions, including universities in Saudi Arabia, but these systems need to be developed to support processes related to quality and academic accreditation in terms of linking these processes with other components in Enterprise. As well as the collection, storage and retrieval of data related to academic accreditation processes, this calls for the concept of developing a special unit of quality and academic accreditation in the ERP system to help meet the new needs of higher education institutions.

III. Research Methodology

The research methodology based on content analysis method which includes (1) analyzing the content of research papers related to ERP system, Quality and Academic accreditation in HEIs. (2) semi-structured interviews with five ERP systems vendors in KSA and (3) NCAAA quality requirements and academic accreditation standards and documents. (4) Analysis of user requirements for those who are responsible for quality assurance in Dar Al Uloom University. This methodology aims to achieve the following research objectives:

- Determine the common ERP modules in higher education institutions.
- Propose a new module for quality and academic accreditation to collect accurate information from several sources, working with huge information and documents, then analyzing this information to generate reports and KPIs results related to quality and accreditation process.
- Link the proposed module with the ERP modules and other systems at the University to form a proposed conceptual framework for all ERP components in higher education institutions.

The conceptual framework proposed in this study consists of two main interrelated components, The first component contains the main modules of the ERP system that are suited to the work environment and organizational context in HEIs such as strategic planning, governance systems, human resources, e-learning systems and student information systems, enterprise resources (finance, supply chain, warehouses, etc.), and Customers Relationship Management CRM. The identification of these modules will be relied upon studying the common functional requirements of ERP systems which used in HEIs in Saudi Arabia. This component will cover the first objective of the study, which is Determine the common ERP modules in higher education institutions. The second component will cover the second and third objective of the study, which is a proposed module dealing with the requirements of quality and academic accreditation (QAAM: Quality Assurance and Accreditation Module) aims to collect data, documents and forms preparation, data aggregation, data analysis, collecting evidences and reports generation based on Accreditation bodies standards, specially NCAAA. This module is linked with ERP modules in the first Component of the proposed framework through the data warehouse of the institution to provide data, statistics and reports related to quality processes and academic accreditation. QAAM also has interfaces that enable users to feed the system with the required data or to obtain data, statistics and reports that help them meet the requirements of quality and academic accreditation. Fig. 2 shows the high-level context of the proposed framework which explains the logical connectivity between the proposed framework components and the institution legacy systems, Learning Management System (LMS), Student Information System (SIS) and Data Warehouse.

A. Component One: main Modules of the ERP System

By reviewing the literature and the interviews with a group of service providers in Saudi Arabia such (Oracle, SAP, Infor, etc.) and (ODES, MADAR and Moraslat), the functionality of main modules of ERP are common and almost similar in ERP providers solutions. To give more focus on these functionalities which related to HEI, ERP modules should cover the functional and non-functional requirements which are mentioned in Table I. The proposed framework will consider all these functionalities.

B. Component Two: QAAM: Quality Assurance and Accreditation Module

Component two (QAAM) is the main module of the proposed framework, which aims to facilitate many processes related to quality and accreditation requirements. Fig. 3 shows the main components of QAAM, which is cover all NCAAA standards and requirements.

1) NCAAA standards interfaces: QAAM provides an interface for each standard of NCAAA standards, these interfaces will help the users to feed the systems with the required data and evidence. Also, the user can access the databases and reports related to QAAM.

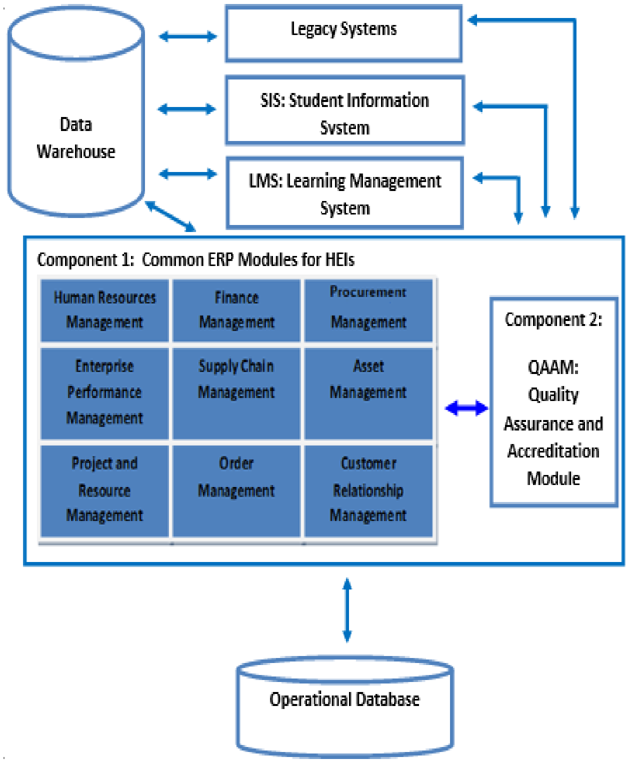

Data Warehouse
Legacy Systems
SIS: Student Information System
LMS: Learning Management System
Component 1: Common ERP Modules for HEIs
Human Resources Management
Finance Management
Procurement Management
Enterprise Performance Management
Supply Chain Management
Asset Management
Project and Resource Management
Order Management
Customer Relationship Management
Component 2:
QAAM: Quality Assurance and Accreditation Module
Operational Database

Fig. 2. Shows the High-Level Context of the Proposed Framework.

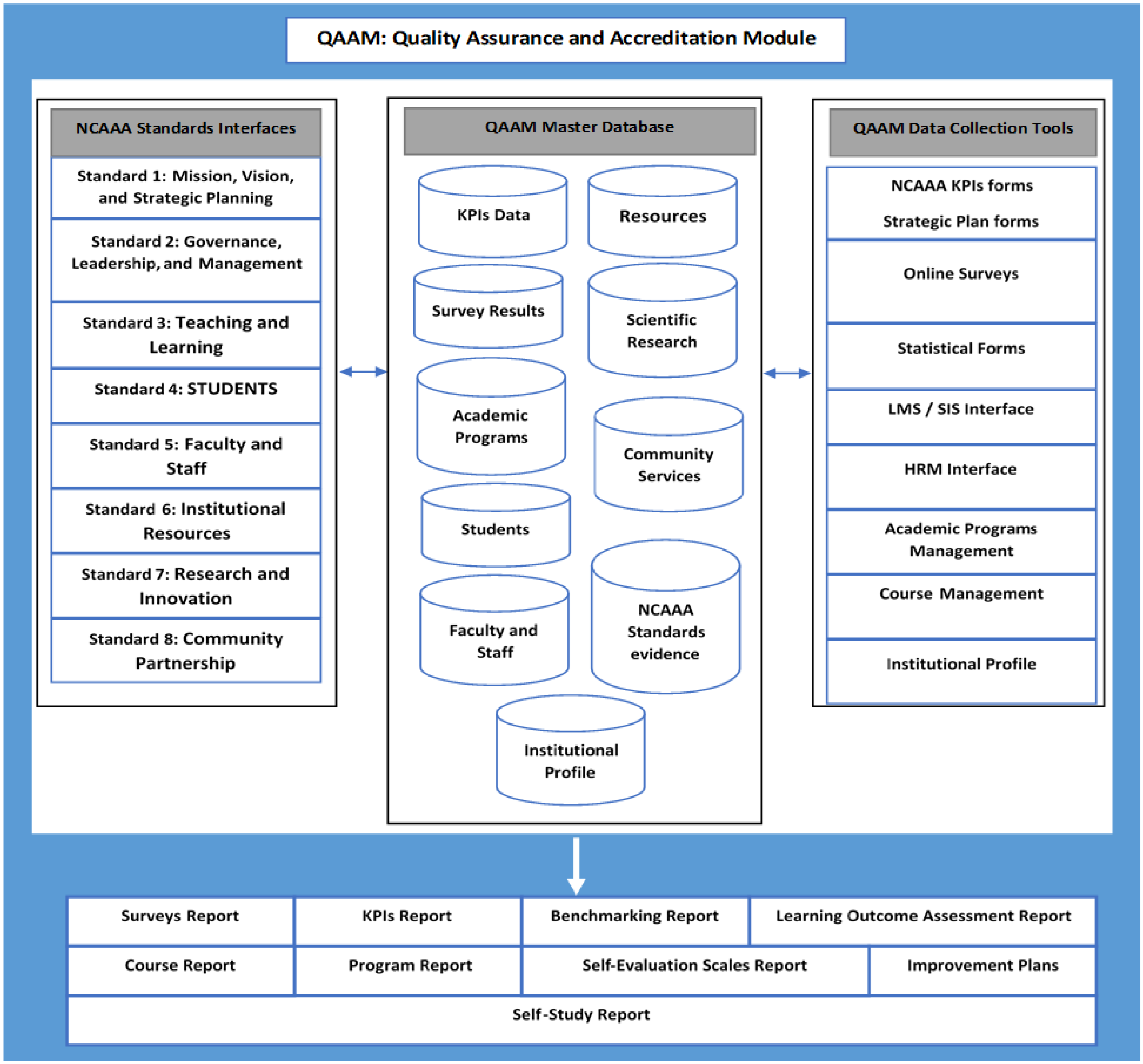

QAAM: Quality Assurance and Accreditation Module
NCAAA Standards Interfaces
Standard 1: Mission, Vision, and Strategic Planning
Standard 2: Governance, Leadership, and Management
Standard 3: Teaching and Learning
Standard 4: STUDENTS
Standard 5: Faculty and Staff
Standard 6: Institutional Resources
Standard 7: Research and Innovation
Standard 8: Community Partnership
QAAM Master Database
KPIs Data
Resources
Survey Results
Scientific Research
Academic Programs
Community Services
Students
Faculty and Staff
NCAAA Standards evidence
Institutional Profile
QAAM Data Collection Tools
NCAAA KPIs forms
Strategic Plan forms
Online Surveys
Statistical Forms
LMS / SIS Interface
HRM Interface
Academic Programs Management
Course Management
Institutional Profile
Surveys Report
KPIs Report
Benchmarking Report
Learning Outcome Assessment Report
Course Report
Program Report
Self-Evaluation Scales Report
Improvement Plans
Self-Study Report

Fig. 3. Shows the Components of the Proposed Quality Module QAAM.

2) QAAM Master Database: The main purpose of QAAM master database is to collect all data of quality assurance and accreditation in one single database, this database is integrated with the other sub-systems of the proposed framework to exchange data and connected with NCAAA standards interfaces to allow users to read and write on specific data fields. Also, QAAM Data Collection Tools will feed this database with the required data and evidence. Table II shows a description for each component of QAAM master database.

3) QAAM Data Collection Tools: QAAM Master Database is depended on set of tools to collect the requited data related to the functionalities of this proposed modules. Table II mapped the components of QAAM Master Database with these tools. Table III shows a description for each tool of data collection tools.

4) Main QAAM Reports: QAAM Master Database and Data collection tools are explained in Tables II and III which aim to provide set of reports on quality requirements and academic accreditation to achieve the primary goal of the proposed framework that system users expect to obtain as basic products on which decisions and improvement plans are based and which will be presented to the accreditation bodies. Table IV shows a description for each of the reports could by generated by the system.

TABLE. II. COMPONENTS OF QAAM MASTER DATABASE

No.	Database Component	The Stored Data	Data Source	Related Standards
1	KPIs Data	Used to store all NCAAA KPIs and Strategic Plan KPIs.	• KPIs Forms. • Strategic Plan Forms.	All
2	Survey Results	Used to store all surveys feedbacks and results.	Online surveys.	All
3	Academic Programs	Used to store all data related to each academic program (Program Specifications, Learning Outcomes Assessments, Courses, etc.)	• Statistical Forms. • LMS/ SIS Interfaces. • Academic Program Management. • Course Management.	Standard 3
4	Students	Used to store all data related to the students.	• Statistical Forms. • LMS/ SIS Interfaces. • Academic Program Management. • Course Management.	Standard 4
5	Faculty and Staff	Used to store all data related to Faculty and Staff.	HRM Interface	Standard 5
6	Resources	Used to store all data related to Institutional Resources includes related to Financial Resources and Budget, Information Technology, Facilities and Equipment	• Connected with ERP Modules (Finance Management, Project and Resource Management, Asset Management)	Standard 6
7	Scientific Research	Used to store all data related to Scientific Research at institutional and program levels.	• Statistical Forms.	Standard 7
8	Community Services	Used to store all data related to Community Services at institutional and program levels.	• Statistical Forms.	Standard 8
9	NCAAA Standards Evidence	Store all evidence of NCAAA standards and practices (Documents Based)	• All NCAAA Standards Interfaces	All
10	Institutional Profile	Used to store all data related to institutional profile which includes general information and statistics about the institution.	• Statistical Forms. • LMS/ SIS Interfaces. • institutional profile tool.	All

IV. DISCUSSION

The proposed a framework (Fig. 2) aims to facilitate the collection of required data from various sources within the university and helps to provide the necessary reports and statistics related to quality and accreditation. A set of quality and accreditation standards linked to human, financial resources and institution asset which could be covered by Component 1: Common ERP Modules for HEIs. ERP modules are used in universities and provided by the most known ERP system providers such as SAP, Oracle, PeopleSoft, also provided by ODES, MADAR and Moraslat. While Component 2, QAAM: Quality Assurance and Accreditation Module is the proposed module which plays a fundamental role in collecting data and generating reports directly related to all accreditation standards through four parts. The first part is NCAAA Standards Interfaces, which is the link between accreditation standards and users. This part relates to the database of the Quality and Accreditation Module. QAAM Master Database (second part of QAAM) store all evidence, data, performance indicators and statistics for each of the accreditation standards, which makes it easier for the user to access them easily and effectively without wasting time searching for them separately and also allows the user to amend and improve on that data and write the necessary comments for each standard in a participatory way between users which reflect those comments on the reports issued by the system, and it also unifies the mechanisms of work among the users without any conflict or inconsistency in the data, which helps in issuing the reports necessary for academic accreditation, Table II explains the role for each component of QAAM Master Database. The third part of the proposed module (QAAM Data Collection Tools) is extremely important as it represents the tools used to collect data and statistics necessary for each academic accreditation standard and store them in the database. The importance of this part comes from the fact that data, statistics, evidence and performance indicators are collected from a variety of sources that are difficult for the user to collect in a traditional way, for example, students' opinion on the quality of teaching courses, with many courses available, there is a need for tools to collect and analyze, so it is used Online survey tool to do this complex process. Each accreditation standard also has a set of performance indicators and statistics that are periodically compiled. Therefore, the data collection tools proposed in this model work to collect a large amount of data, organize and store it in an accurate way to be used by all users of the system, and this is one of the most important services that the user is satisfied with. Table III explains the purpose of each data collection tool. To create a business intelligence environment in the proposed framework, Data Warehouse is used to support QAAM Data Collection Tools by extracting raw data from various business systems and data sources in order to reveal meaningful knowledge, Data Warehouse a large relational database that combines pertinent data in an aggregate, summarized form suitable for enterprise wide data analysis, reporting, and management decision making, Data Mart is subset of a data warehouse that is usually designed for a specific set of users to provide them a specific data. The final product of the proposed framework and the proposed quality module QAAM is the extraction of the final reports of the quality and academic accreditation (Part Four) which depend on the integration of all other parts of the proposed framework, as the process of preparing these reports is extremely difficult due to the lack of evidence, data, indicators and statistics associated with each standard. This framework facilitates the process of issuing these reports in an automated way as they relate to the databases of QAAM as well as through comments provided by the user through NCAAA Standards Interfaces.

TABLE. III. QAAM DATA COLLECTION TOOLS

No.	**Data Collection Tools**	**Description**	**Integrated with**
1	NCAAA KPIs Forms.	Used to collect: • NCAAA KPIs (23 KPIs at Institutional Level and 17 at Program Level). • Strategic Plan KPIs. • Other Performance Indicators. • Input data by the users.	• ERP Modules. • LMS/ SIS Interfaces.
2	Online Surveys.	Used to collect client satisfaction in different domains: • NCAAA Surveys (Course Evaluation Survey, Program Evaluation Survey, Student Experience Survey, Employer Evaluation Survey and Alumni Evaluation Survey). • Institutional Surveys (Climate Survey, Faculty Members Satisfaction Survey, Staff Satisfaction Survey, Vision & Mission Surveys, All Services Clients Satisfaction Surveys)	• LMS/ SIS Interfaces. • HRM Interface. • Academic Programs Management. • Course Management.
3	Statistical Forms.	Used to collect the statistical Data to feed all NCAAA standards and KPIs such as: • All statistical Data related to the Students (registered, withdraw, delayed, denied, transferred, failed, successful, excellent, participating in various activities, etc.). • All statistical Data related to Alumni. • All statistical Data related to Employers • All statistical Data related to HR. • All statistical Data related to Libraries and Learning Resources. • All statistical Data related to Scientific Research. • All statistical Data related to Professional Development Processes. • All statistical Data related to Community Services. • All statistical Data related to Institutional Resources includes related to Financial Resources and Budget, Information Technology, Facilities and Equipment.	• ERP Modules. • LMS/ SIS Interfaces. • HRM Interface. • institutional profile.
4	LMS/ SIS Interfaces.	Used as an interface to link the proposed QAAM components with the LMS and SIS systems to obtain student-related information, academic program management, and courses management.	• QAAM Components. • SIS: Students Information System. • LMS: Learning Management System.
5	HRM Interface.	Used as an interface to link the proposed QAAM components HRM Module in ERP system.	• QAAM Components. • NCAAA KPIs Forms. • Statistical Forms.
6	Academic Programs Management.	This tool plays important role in managing academic programs such as: • Program Design and Development Processes. • Curriculum. • Course Mapping. • Program Specification. • Program Reports. • Graduate Attributes and Learning Outcomes. • Learning Resources. • Program Evaluation Survey. • Student Experience Survey. • Program KPIs. • Quality of Teaching and Students' Assessment. • Statistical Data of Program.	• QAAM Components. • LMS/ SIS Interfaces. • ERP Modules. • Online Surveys. • NCAAA KPIs Forms. • Statistical Forms.
7	Course Management	Used to manage all courses for each program: • Course Specifications. • Course Report. • Course Materials. • Course Evaluation Survey. • Students Attendance Reports. • Assessment Instructor Material. • Course Learning Outcomes Assessment. • Final grade reports.	• QAAM Components. • LMS/ SIS Interfaces. • ERP Modules. • Online Surveys. • NCAAA KPIs Forms. • Statistical Forms.

8	Institutional Profile	This tool is used to manage the Institutional Profile which is contains: • Summary of the Institution History. • Institution's Academic Units. • List of the Institution's Achievements, Awards, and Significant Accomplishments. • Institutional Performance Indicators. • Program Data. • Students Statistics. • Teaching Staff Statistics. • Graduates Statistics.	• QAAM Components. • LMS/ SIS Interfaces. • ERP Modules. • NCAAA KPIs Forms. • Statistical Forms.

TABLE. IV. MAIN QAAM REPORTS

No.	Main QAAM Reports	Description
1	Surveys Report	Includes analysis of each NCAAA Surveys and Institutional Surveys distributed by Online Survey Tool.
2	KPIs Report	Three types of KPIs analysis reports. First report is related to NCAAA KPIs at institutional level (23 KPIs). Second report is related to NCAAA KPIs at program level (17 KPIs). Third report is related to strategic Plan.
3	Benchmarking Report	Four types of Benchmarking Reports. Two reports at institutional level (Internal and External Benchmarking Reports). Two reports at program level (Internal and External Benchmarking Reports).
4	Learning Outcome Assessment Report	Comprehensive Learning Outcome Assessment Report includes Course / Program / Institution Learning Outcomes Assessment Report. This report. This report compares the current and previous targets for learning outcomes and analyzes the results to develop improvement plans and define new targets for learning outcomes.
5	Course Report	The course report is one of the most important requirements of NCAAA Requirements to monitor the quality of teaching courses and ensure continuous improvement. This report includes information about students in the course and their results, checking the appropriateness of teaching strategies and assessment methods for the course, results of measuring learning outcomes for the course, assessing the quality of the course by beneficiaries and management. Development of course improvement plans
6	Program Report	Based on NCAAA requirement, each academic program must issue an annual report on the quality of the program at the end of the academic year (Annual Program Report). This Report includes Program Statistics, Results of Program Learning Outcomes Assessment, Result Analysis of Course Reports, Program Activities, Program Evaluation and Program Improvement Plan.
7	Self-Evaluation Scales Report	Two types of Self-Evaluation Scales Report, one at institutional level and the other at program level. This report aims to help those responsible for quality assurance at institutional / Program Levels to conduct evaluation in an objective manner based on the institutional/programmatic quality assurance standards prepared by NCAAA. This document can also be used in the field of planning, internal auditing, and supporting strategies to improve institutional quality and the quality of academic programs in Institutions of higher education.
8	Improvement Plans	All previous reports help quality officials and decision makers develop improvement plans. This report is concerned with following up on what has been achieved of improvement initiatives at various levels of the institution.
9	Self-Study Report	The self-study report is the most important products of the proposed framework, which depends on all previous reports and on all proposed databases. The self-study report is an extensive examination of the quality of the educational institution or academic program in accordance with the quality assurance and accreditation standards set by NCAAA, linking them to the evidence, statistics and analyzes associated with them. Data warehouse of the proposed framework will provide set of data mart to feed the Self-Study Report.

This study provided a conceptual framework that is proposed to add a special module for quality and academic accreditation QAAM that is associated with the ERP system in the educational institution and other systems to facilitate the process of collecting, organizing and storing data, evidence and performance indicators for academic accreditation and issuing final reports. This framework aims to achieve the study objectives and to cover the functional and non-functional requirements in Table I as well as dimensions of quality in higher educational institutions in Fig. 1, for the purpose of covering the quality requirements and academic accreditation of National Center for Academic Accreditation and Assessment (NCAAA).

V. CONCLUSION AND FUTURE WORK

Higher education institutions have relied on information systems for carrying out their operations and managing their data for several decades. However, the structure of these systems is evolving as the needs of these institutions evolve. HEIs have started adopting ERP systems to ensures the integration of all the operations and data of the organization. One of the modern needs of higher education institutions is compliance with local and international quality and accreditation standards. As this has become one of the real challenges in collecting, linking and analyzing data to extract the necessary reports for this institution to be eligible to progress and obtain academic accreditation through the application of an effective quality system that ensures continuous improvement in academic processes and services provided to beneficiaries. Due to the diversity of data sources and challenges faced by those responsible for academic quality requirements in preparing evidence, files, statistics and performance indicators in accordance with academic accreditation standards, the main purpose of this study is to provide a conceptual framework that links the institution's systems with each other. The framework provides Quality Assurance and Accreditation Module (QAAM) as part of the ERP system and it undertakes a task Facilitate the work of collecting and analyzing all data related to quality and academic accreditation and issuing their reports such as performance indicator reports, program reports, academic courses and self-evaluation report, which helps in development of improvement plans and finally generating the self-study report, through an interactive interfaces between the system and the users that enable them to obtain data, evidence and reports and allow the users to feed the system with appropriate information and comments.

Future work leads us to implement this proposed framework, and to study its ability to achieve the goals for which it was created. It is also proposed to apply it widely among higher education institutions and to benefit from automated benchmarking between these institutions based on the performance indicators reports issued by the system, as the external benchmarking is also a requirement of quality and academic accreditation.

References

[1] Abugabah A., Sanzogni L. and Alfarraj O., (2015), "Evaluating the impact of ERP systems in higher education". The International Journal of Information and Learning Technology Vol. 32 No. 1, 2015 pp. 45-64 ©Emerald Group Publishing Limited 2056-4880. DOI 10.1108/IJILT-10-2013-0058.

[2] Pollock and Cornford (2004), "ERP systems and the university as a “unique” organisation". Information Technology & People Vol. 17 No. 1, 2004 pp. 31-52. Emerald Group Publishing Limited 0959-3845 DOI 10.1108/09593840410522161.

[3] SABAU G., MUNTEN M., BOLOGA A., BOLOGA R. and SURCEL T (2009), "An Evaluation Framework for Higher Education ERP Systems". WSEAS TRANSACTIONS on COMPUTERS. Issue 11, Volume 8, November 2009. ISSN: 1109-2750.

[4] Rico, D. F. (2004). "ERP in higher education". Retrieved 6 September, 2019, from: http://davidfrico.com/rico04f.pdf.

[5] Abugabah A. and Sanzogni L. (2010), "Enterprise Resource Planning (ERP) System in Higher Education: A literature Review and Implications". World Academy of Science, Engineering and Technology International Journal of Computer and Systems Engineering Vol:4, No:11, 2010.

[6] Hamdatu M., Siddiek A. and Al-Olyan F. (2013), "Application of Quality Assurance & Accreditation in the Institutes of Higher Education in the Arab World (Descriptive & Analytical Survey)". American International Journal of Contemporary Research Vol. 3 No. 4; April 2013.

[7] Abou-Zeid A. and Taha M., (2014), "Accreditation Process for Engineering Programs in Saudi Arabia: Challenges and Lessons Learned". IEEE Global Engineering Education Conference (EDUCON). 3-5 April 2014, Military Museum and Cultural Center, Harbiye, Istanbul, Turkey.

[8] Yvonne Lederer, A., Gail, C., Glenn, S. and Albert, L.H. (2004), “Enterprise systems education: where are we? Where are we going?”, Journal of Information Systems Education, Vol. 15 No. 3, pp. 227-233.

[9] Kahveci T., Uygun, Yurtsever, Ulaş and Sinan (2012), "Quality Assurance in Higher Education Institutions Using Strategic Information Systems". In 3rd. International Conference on New Horizons in Education - INTE 2012, Procedia - Social and Behavioral Sciences. 5 October 2012 55:161-167 DOI: 10.1016/j.sbspro.2012.09.490, Database: ScienceDirect.

[10] Sowan I., Tahboub R. and Khamayseh (2017), "University ERP Preparation Analysis: A PPU Case Study". (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 11, 2017.

[11] Rani (2019), "A Review of ERP Implementation in Higher Education Institutions". International Journal of Advanced Research in Computer Science and Software Engineering. Volume 6, Issue 6, June 2016 ISSN: 2277 128X.

[12] Davis, M. and Huang, Z., (2007), “ERP in higher education: a case study of SAP and campus management”, Issues in Information Systems, Vol. 8 No. 1, pp. 120-126.

[13] Kvavik, R., Katz, R., Beecher, K., Caruso, J. and King, P. (2002), “The promise and performance of enterprise systems for higher education”, Educause, Vol. 4 No. 1, pp. 5-123.

[14] Noaman and Ahmed (2015), "ERP Systems Functionalities in Higher Education," Procedia Computer Science, vol. 65, pp. 385-395, 2015.

[15] Materu, P. (2007), "Higher Education Quality Assurance in Sub-Sahara Africa - Status, Challenges, Opportunities and Promising Practices". Washington, D.C: World Bank working paper No. 124, Africa Region Human Development Department, The World Bank.

[16] INQAAHE. (2011). "Operating an External Quality Agency".

[17] Islam, Ali and Islam (2017), "Quality Assurance and Accreditation Mechanisms of Higher Education Institutions: Policy Issues and Challenges in Bangladesh". European Journal of Education Studies ISSN: 2501 - 1111 ISSN-L: 2501 - 1111. doi: 10.5281/zenodo.495792.

[18] Abu Jaber and Al Batsh (2016), "Jordanian Experience in Accreditation and Quality Assurance in HEIs". US-China Foreign Language, April 2016, Vol. 14, No. 4, 312-327 doi:10.17265/1539-8080/2016.04.007.

[19] Al Mohaimeed A., Midhet F., Barrimah I. and Saleh M., (2012), "Academic Accreditation Process: Experience of a Medical College in Saudi Arabia". International Journal of Health Sciences, Qassim University, Vol. 6, No. 1 (Jan 2012/ Safar 1433H).

[20] Alalfy H., Al-Aodah I. and Shalaby E., (2013), "Role of Development and Accreditation Deanship for Qualification of Hail Faculties, Saudi Arabia for Local Accreditation". Greener Journal of Educational Research ISSN: 2276-7789 Vol. 3 (3), pp. 123-133, May 2013.

[21] Al-shafei A., Abdulrahman K., Al-Qumaizi K. and El-Mardi A. (2015), "Developing a generic model for total quality management in higher education in Saudi Arabia". Medical Teacher. 37:sup1, S1-S4, DOI:10.3109/0142159X.2015.1006607.

[22] Mahbub (2017), "Quality Assurance for Higher Education: Challenges in Sustaining Continuous Quality Improvement for Malaysian Universities". Proceedings ff Inted2017 Conference 6th – 8th March 2017, Valencia, Spain, ISBN: 978-84-617-8491-2.

[23] Soliman and Karia (2015), "Enterprise Resource Planning Systems in Higher Education Context: Functionalities and Characteristics". International Journal of Innovative Research in Science, Engineering and Technology. Vol. 4, Issue 11, November 2015. ISSN (Online): 2319-8753. ISSN (Print): 2347-6710.

IoT and Blockchain in the Development of Smart Cities

Laith T. Khrais
Department of Business Administration, College of Applied Studies and Community Services
Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia

***Abstract*—With the advent and proliferation of the internet, the fourth industrial revolution is in full swing. As a result, different technologies have the potential to impact the course of human development. In other words, worldwide populations are moving towards growing urban centers and as a result, smart cities are emerging as the integration of human activities and technologies. These smart cities are built on top of different technologies such as blockchain and the Internet of Things (IoT). Consequently, the applications of these technologies in current and future smart cities will not only change the nature of human interaction and governance but also how business is conducted. This paper proposes an experimental study (qualitative and quantitative) that will determine the impact of blockchain and IoT technologies on the development of smart cities. It aims to derive insight from questions such as how current business models are preparing themselves for this disruption, the challenges they will face, and the potential contributions the two technologies will have on business development. The study's outcomes will provide the rationale for why businesses should start paying attention to these technologies and start on an early adoption plan that will slowly transform their business models as smart cities mature.**

Keywords—Internet of things; blockchain technologies; smart cities; emerging markets; electronic commerce

I. Introduction

Powered by technologies such as the internet, the fourth industrial revolution is coming along exponentially as the world transforms into a global village. As a result, new markets are emerging, such as smart cities that use technologies such as online payment, internet connectivity, and blockchain, among others, to perform transactions in a fast and secure manner [1]. Besides, according to Schwab and Davis [2], the fourth industrial revolution is powered by data and, therefore, any technology that is data intensive such as the Internet of Things (IoT), cloud computing, machine learning and, artificial intelligence will not only grow in value but be applied widely. It should be noted that IoT in this paper will not only include embedded devices with network connectivity but also other portable computing platforms such as tablets, laptops, and smartphones.

As a result, connectivity is a very important aspect of smart cities where wireless connectivity via the Wi-Fi protocol is not only available from the devices people use in their daily lives but also provided by internet and network service providers [3]. Additionally, people use this connectivity to connect as well as shop online and do their banking at the convenience of their homes or workplaces. Other online to offline interactions also occur where, for example, an individual can purchase a CCTV surveillance camera for their home or business and get after-sale services such as installation and maintenance for free depending on the vendor and location.

These smart cities, however, are built around centralized infrastructures and, as a result, a vulnerable to a single point of failure attack. A bank, for example, can charge very high tariffs for every transaction at the cost of the convenience. Additionally, the bank may not always be online or might be breached thus losing customer data to malicious attackers, as evidenced by the recent Capital One data breach [4]. Additionally, with the growing demand for personal data such as credit card information, passport details, and medical history on the dark web [5], smart cities expose their residents to undue risks.

To solve these and other problems facing smart cities, blockchain technology, derived from the bitcoin paper by Satoshi Nakamoto [6], is needed. Not only are blockchains tolerant to attacks, but they are also tolerant of failure (Byzantine fault tolerance), decentralized, secure, and peer to peer atomic transactions that eliminate the man in the middle (and their tariffs) [7]. When the decentralized nature of blockchains is coupled with the versatility of IoT, smart cities get the potential to grow further while shoring up their weaknesses.

There are drawbacks to using blockchain technologies in smart cities, however. As decentralized systems, blockchains exist outside the regulation and control of governments and their regulatory bodies. As a result, some countries have banned blockchains such as bitcoin while they view other decentralized applications (DApp) as threats [8]. Therefore, this paper is a review and exposition of blockchain and IoT technologies in the development of smart cities. As a result, the paper is divided into three major sections. The first section will review all background information, such as the emerging markets in smart cities, their business models, and their impact. The second section will conduct a qualitative and quantitative study on the development of blockchain and IoT technologies smart cities while the last section collects the results and discusses the findings.

II. Literature Review

A. Survey of Emerging Markets and Smart Cities

The 90s ushered in a period of unprecedented development with the invention and proliferation of electrical and electronic household and communications devices such as computers, modems, cable tv, and the earliest predecessor of the modern internet powered by the world wide web (and HTTP) [3]. This trend has developed and created new markets such as e-commerce, social networking, online shopping, and advertisements. These technologies have enabled companies to derive over 60% of their revenue from online activities such as advertisements [9] [10]. In 2014, for example, e-commerce sales ballooned from 6 billion dollars of previous years to over 25 trillion dollars and counting [1]. Therefore, the potential and opportunities for the growth and development of smart cities already exist. All that is missing are the underlying technologies and their associated business models to get the ball rolling.

B. Preview of the Two Business Models

This paper looks into the application of two business models, namely the Internet of Things (IoT) and Blockchain, and their interaction with the models of operations of e-commerce in smart cities. Besides, while blockchain and IoT can be implemented independently, their integration gives a product that is more than the sum of the individual technologies. This section outlines the existing business model built on top of IoT and Blockchain while citing successful examples.

1) Internet of Things (IoT) business models: In a nutshell, IoT is the systematic method of attaching sensors to an existing system and getting data out of the system. As of 2015, the IoT industry was valued at 11 trillion dollars, but despite this, three out of four IoT businesses fail [11]. Despite this, IoT devices proliferate cities and homes from smart alarms to refrigerators and TVs [12]. The reason behind this high failure rate is that most individuals and companies who apply IoT in their businesses get access to the data but do not know how to turn it into a source of competitive advantage or revenue [12]. Similar to how data is the currency for the fourth industrial revolution [2], data is the backbone of IoT and its associated business models. The following paragraphs present the different business models built on top of IoT.

Industries such as aerospace, oil, and gas extraction are heavily regulated due to the high risks, costs of accidents occurring, and the widespread consequences the accidents would have on the social, political, environmental, and economic aspects of commercial operation. The traditional method of ensuring safety was to set compliance standards, and having human inspectors regularly inspect the equipment and plants [13]. With IoT, however, businesses ensure compliance by attaching sensors where plant monitoring can be done in real-time [14]. As a result, companies can be proactive about safety without incurring additional costs. This practice has become common in the aerospace industry, where jet engines are equipped sensors that transmit engine health data to the manufacturer for real-time monitoring, performance evaluation, and warning [15].

The second business model applying IoT is preventative maintenance, where cheap IoT platforms can be used to monitor systems and plants in operation continuously and warn when malfunctions occur due to unfavorable operating environments [16]. Alternatively, equipment or plants can be scheduled for regular maintenance schedules where they are shut down. However, the economic costs of these downtimes can be prevented by using preventative maintenance policies, especially in equipment and plants whose operation and maintenance are complex [16]. General Electric has successfully utilized preventative maintenance with IoT enabled sensors in the wind turbine blades to not only change the pitch angle depending on the local airflow but also schedule downtimes for maintaining the turbines when there is no wind blowing [12].

Last but not least, IoT sensors and systems have found widespread application in the field of remote diagnostics, especially in the medical and agriculture industries. Greenhouse operators, for example, use IoT sensors to remotely monitor and control favorable plant growing conditions such as humidity, light, temperature, and $CO_{2,}$ among others [17]. In the medical industry, IoT platforms have found application in patient condition monitoring (using wearable devices), treatment (such as insulin pumps), and fitness tracking [18].

These IoT business models assume that the users and businesses rely on their expertise to design, develop, and produce these IoT platforms (both hardware and software). The truth, however, is that different business models have emerged that are catered to providing technical know-how where they develop the products the end users build their businesses on [19]. An IoT device developer and provider might, for example, lease out their products (especially software that runs on dedicated cloud servers) on a subscription model and make support services available 24/7.

Similar to the Software as a Service (SaaS) cloud architecture and business model, the subscription model has the advantage of sustainable revenue instead of selling the product in a one-time sale then having the customers pay for future upgrades [20]. Alternatively, an IoT developer and manufacturer might opt not to sell their product but sell the desired outcome. This outcome-based business model focuses on meeting the customer’s needs instead of making a product that a customer might choose to fit their needs. Rolls Royce is such an example where they sell the service (24/7 engine health and performance monitoring for every engine they sell, and it is in service) instead of simply selling engines based on the aircraft manufacturer’s or operator’s requirements [15].

2) Blockchain business model: The downside of smart cities is that there can be too much connectivity that instead of being a boon or extra feature, it starts to become an anchor. Take the internet; for example, targeted advertisement has become so common that companies create profiles of their users and sell the data to other companies. Additionally, user and browser fingerprinting have become advanced that it is now entering ethical grey areas [21]. In IoT, different devices and sensor networks collect private, confidential, and business secrets that are vulnerable to exposure through attacks such as a man in the middle attack [22]. Besides, even trust in the devices themselves is not guaranteed, as evidenced by the actions of the mobile phone company Huawei and the Chinese government [23].

As a result, the integration of blockchain and IoT will shore up the deficiencies of smart cities. However, what is blockchain? A blockchain is a distributed ledger whose entries are time-stamped and cryptographically signed to ensure their immutability [24]. A business model built on top of blockchain technology will not only become decentralized (exhibiting Byzantine fault tolerance) but also peer to peer transaction in a network where trust is enforced by the business model [25].

As a result, decentralization, immutability, and transparency form the cornerstones of all blockchain-based business models [26]. These business models utilize the blockchain infrastructure in three primary ways. First, they store all their data on the blockchain to ensure that the data is tamper-proof [27]. Secondly, they apply the transparency feature of blockchains to enhance the functionality and utility of existing infrastructures such as supply chains [28]. Lastly, the more advanced business models build their artificial intelligence systems on top of blockchains to build a decentralized AI system [29]. The following paragraphs describe the different business models inspired by or built on top of blockchains.

The easiest business model for a non-technical company that wants to integrate blockchain technology into their existing infrastructure and business model is the Blockchain as a Service business model. This is because the ecosystem allows the business to create their products, experiment, and release their product while abstracting away low-level infrastructure details [30]. One example is the Ethereum Blockchain as a Service developed and managed by a joint venture between Microsoft and ConsenSys [31].

A utility token business model, on the other hand, is based on the use and exchange of tokens that have inherent value within the blockchain. This business model, also known as token economics, performs a similar function to traditional banks only that they are decentralized, and the creation, utilization, and destruction of tokens are not under the control of a central authority such as government [32]. Besides, the tokens themselves are valuable because the users deem them valuable and not because they are backed by gold reserves, for example. As a result, utility token business models become profitable only when the value of the individual tokens increase.

In contrast, a securities business model is a recent blockchain business model where a company sells tokens known as securities that are expected to gain value when individuals buy them. In other words, a blockchain security token is synonymous with possessing legal ownership to an asset [33]. The role of the blockchain is, therefore, to verify the real owners of the tokens as well as regulate the creation and consumption of security tokens. This business model also allows verified owners not only to trade their tokens but also use them as collateral. As a consequence, the securities business model has the potential to not only redefine the traditional concept of ownership but also provide a means of redistributing wealth among a large population without devaluing the tokens [34].

C. Impact of the Business Models

It is projected that the IoT industry will grow to encompass over 20 billion devices by 2020, 4.5 billion of them being from Europe [35]. Additionally, over 65% of the worldwide human population is projected to have moved to urban cities by 2040 [35]. As a result, the emerging smart cities will not only be relying on traditional wireless networks to provide connectivity but also local area networks powered by the streetlight to form a mesh network or a wide area network (WAN). Besides, the exponential growth in the number of connected IoT devices in a smart city increases the risk of unauthorized access to private and sensitive data transmitted over the wide-area networks. It, therefore, becomes imperative that blockchain technology is used to secure these communications while maintaining transparency and data integrity. As a result, the following are how the integration of blockchain and IoT technologies would impact the development of smart cities.

IoT security is a concern that all developers and some end users recognize. This is because there have been demonstrations of how the security of IoT devices, especially the data they collect and transfer can be easily compromised. Such systems use traditional cryptographic methods of securing the data where private keys are exchanged and data transmitted over the network in the form of ciphertext. However, considering the global cost of cybercrime [36], the potential damage to smart cities would be exponential and can even destabilize a city. As a consequence, a different security architecture is needed where instead of an attacker breaking the security on singular devices, they will be forced to break the entire network (at least 51% in case of a blockchain-based IoT network). Blockchain networks are very hard for a malicious attacker to gain control because they have an enormous price tag attached to such efforts. As a result, by securing IoT networks with blockchain technology, the developers and users can be safe from most attacks, except individuals or institutions that are willing to pay the price [37].

There is a catch, however. In its current state, blockchain technology is slow and not scalable [38]. For instance, on the bitcoin network, transactions are stored and verified in blocks before they are permanently added to the blockchain. On average, verifying transactions in a single block takes 10 minutes. For a city with billions of devices, the speed of the IoT network built on top of the blockchain infrastructure will drop. This will also impact business models using blockchain. As a result, businesses might be expected to do a cost-benefit analysis and decide to do away with the extra security that blockchain affords. Security-conscious users and developers, however, might find the speed as a worthwhile trade-off. Efforts are underway, however, to improve the speed of blockchain and make it scalable [39] [40].

Despite the drawbacks, blockchain and IoT will be the cornerstone of current and future smart cities. This is because the integration of the two platforms is highly versatile, as well as easy to implement. One such impact these two technologies will have in smart cities is how they will change public transit. Public transit ridership has been on the rise in the last two decades, where millions of workers and city residents use it to move from point A to B (such as home to work) [40]. The number is expected to rise given the growing trend of climate awareness and the efforts to minimize further pollution by reducing emissions from burning fossil fuel. The Maltese government, for example, has contracted the private company Omnitude to develop and provide solutions to the challenges facing its public transport sector using blockchain technology [41]. This system is expected to be a payment point for the different public transport systems to improve transparency in the operations and sector.

Besides, another element of smart cities to be impacted by IoT and blockchain is e-commerce. Built on these two technologies, smart cities of the future will enhance the provision of local products and services to residents close to their homes or workplaces. Such drive for encouraging local uniqueness will encourage residents to create products and services using blockchain and IoT, thus eschewing national and overly capitalistic brands in favor of local brands that serve the local population. For this to be successful, however, the different blockchain applications and IoT platforms will have to be interoperable and compatible with each other [42]. This challenge is currently being solved by companies such as Chain of Things that are attempting to make IoT (software and hardware) and blockchain infrastructures interoperable [43].

III. Methodology

This section presents the methodology and research design aimed to determine the impact of IoT and blockchain technologies in the development of smart cities. As a result, the research design will both qualitative and quantitative. While the quantitative aspect of the research will give insight into the statistical aspects of the study, the quantitative study will provide additional information, especially objective and subjective perspectives the research participants have on the subject matter.

A. Qualitative Approach

The qualitative approach will entail collecting secondary information on model companies and assessing it to address the research problem. Businesses that adopt the Internet of Things and blockchain models will constitute the population for this study. A majority of the companies to be sampled for the study is in the world's smart cities and upcoming markets. A random sampling technique will be used to prevent sampling and information bias in the findings of the study. The sampling method would take into consideration the different outplays that cannot be left out, including a review of the previous studies about and related to the subject matter.

The qualitative approach has the limitation that technologies are constantly evolving. As a result, the findings will not necessarily represent the current state of affairs in the industry. For this reason, secondary data is not adequately updated or, at times, is left unattended as brand new models eclipse the previous ones. For instance, smart cities and emerging markets show that the shoppers and bankers feature a high use of technology in the form of two business models (IoT, blockchain or a combination of both). The two models define ways of making the length of processing transactions shorter and more flexible and transparent. Charts and graphical representation will be applied to understand the concepts and derive the impacts on e-commerce while using such business models.

B. Quantitative Approach

Besides, the study will employ a quantitative approach to address the research problem. The approach will be useful in collecting information on the perspective of industry experts on the various developments and dynamics of blockchain technologies and the Internet of Things.

1) Target population: The study will target company managers and intellectuals from the industry, such as university professors, pioneers of IoT, and blockchain technologies. The participants will be drawn from various companies as well as from different sectors. Consequently, the study will recruit 60 participants to contribute to the required information. Although a sample size of 60 is not enough to substantiate the validity of the findings of the study, it is appropriate for a pilot study on a fact-finding mission.

2) Research instrumentation: The study will utilize interviews to collect quantitative data. The interviews will constitute both closed and open-ended questions. The closed-ended questions will gather statistical data that will enable the researcher to measure opinion and the popularity of ideas on the subject matter. On the other hand, the open-ended questions will help obtain detailed information on the questions through the explanations given by the participants.

3) Data analysis: The collected data will be recorded in tables and analyzed using Excel software. Using the Excel program, the researcher will derive graphical illustrations and charts from demonstrating the outcome of the study.

C. Ethical Considerations

This study will adhere to the conventional ethical considerations that govern research work. These include obtaining informed consent from the sampled and accepted participants of the study before starting the study while respecting their decision not to take part in the study for various reasons. Secondly, the participants of the interviews were briefed on the purpose of the study, and hence, were given a chance to exercise free-will to give the required information. Considering this, no participant was coerced to engage in the study. Finally, the participants were accorded due to confidentiality and anonymity, as their details such as names were not recorded anywhere in the course of the study where codes (alphanumeric) were used.

IV. Results and Discussion

After data analysis, the findings of the research will be discussed in this section. The following are the different topics (aspects) of the subject matter that will be investigated from the qualitative and quantitative research designs.

A. Contributions of IoT and Blockchain Technologies to Business Development

Fig. 1 shows the ratings for how different users would use the combination of IoT and blockchain, and what they preferred or would like improved about the integration of the two technologies as they are used to develop smart cities.

These include faster, more scalable transactions, and transparency in the whole infrastructure (five-star ratings). On the other hand, IoT and blockchain technologies incorporate radio frequency identification to check on product quality, conditions and individual requirements. For instance, shoppers with an interest in chilled red meats or frozen dairy products can use such models to indicate their interests when placing orders. IoT, in combination with blockchain, helps product assessment to be conducted continuously.

Additionally, the end-to-end conversation between sellers and shoppers assists in real-time rectification of mistakes in product specifications, and the shoppers are, therefore, assured of high-quality products through the use of IoT and blockchain platforms. Besides, they are supplied with periodic updates on the service improvements and changes in the websites or devices to keep me up with the industry trends. Once again, at the backdrop of cutting short intermediaries, there can be the creation of jobs in the courier and warehouse services, in case deliverables are collected expressly.

These are among the multiple impacts of IoT and blockchain on the development of smart cities that will be substantiated by the data collected in this study. In this context, Fig. 2 and Fig. 3 summarize the fields impacted by blockchain and IoT as well as how their application is projected to integrate with communities living in smart cities.

However, there is a concern that such a research design is flawed and would be susceptible to confirmation bias. The conductors of this study have, however, considered this and prevented such an event by using statistical methods such as working by positing a null hypothesis to be confirmed or rejected by the collected data.

B. The Readiness of Modern Businesses in Embracing IoT and Blockchain Technologies

The interview questions will ask the participants whether their businesses have the capacity to adopt the Internet of Things and blockchain technology. Additionally, they will be asked if they are equipped with the resources necessary to not only be competitive enough in the emerging smart cities but also be profitable. This will help to understand questions on the readiness of the current business models to evolved and adapt to the changing tides. Out of 60 respondents, 25 noted that their companies were ready to fully pursue the transition. However, 35 of the respondents stated that the full embracement of the new changes requires undergoing structural and human resources adjustments in their organizations.

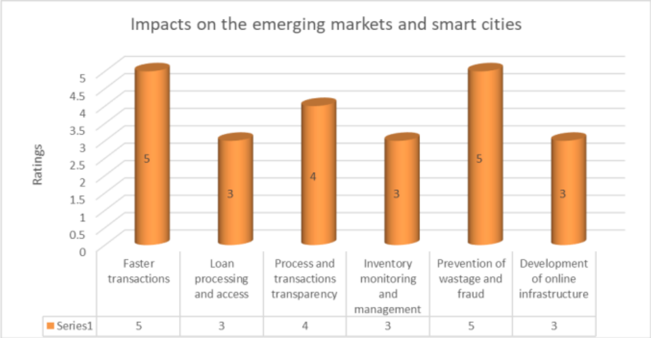


Fig. 1. Business Models and their Implications in Emerging Markets and Smart Cities.

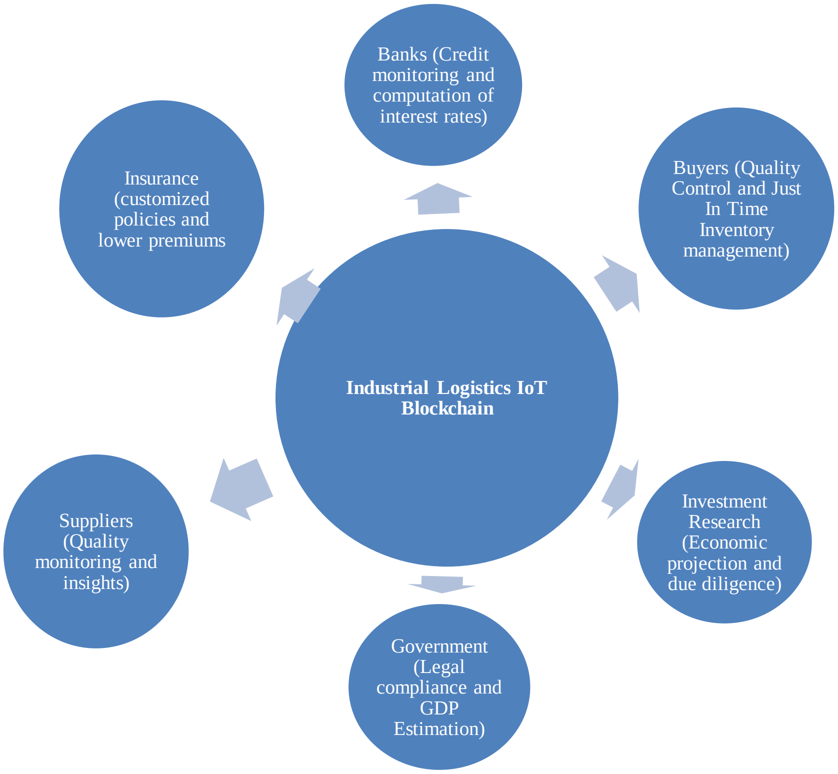

Banks (Credit monitoring and computation of interest rates)
Insurance (customized policies and lower premiums
Buyers (Quality Control and Just In Time Inventory management)
Industrial Logistics IoT Blockchain
Suppliers (Quality monitoring and insights)
Investment Research (Economic projection and due diligence)
Government (Legal compliance and GDP Estimation)

Fig. 2. The Analysis of the IoT Impacts on Business Commerce.

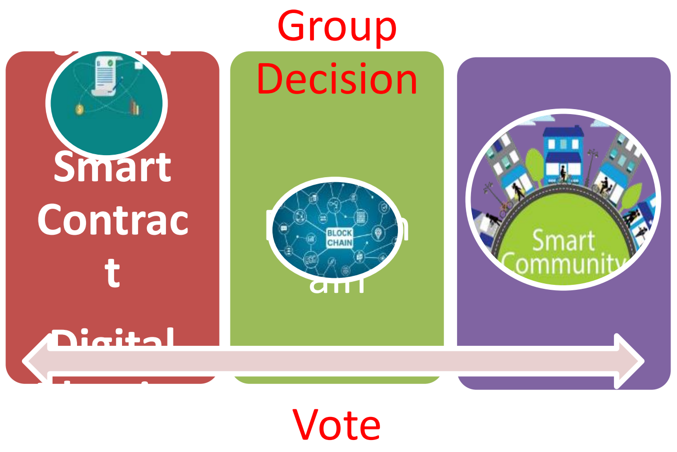


Fig. 3. Application of Blockchain in Emerging Markets.

C. Challenges Encountered when Adopting IoT and Blockchain Technology

The participants will also be prompted to indicate the challenges their companies encountered or were likely to encounter during enhancing of the IoT and blockchain technologies. This will help to determine which resources, such as personnel and different mindsets, are required in the transition and operation of the business models that integrate blockchain and IoT technologies. Additionally, the study aims to determine these challenges because they are also business opportunities for current and future entrepreneurs to exploit. These are summarized as follows in Fig. 4.

D. Theoretical and Practical Contributions

The outcome of this study will be imperative in giving information on the benefits of IoT and blockchain technology in businesses. Hence, this study establishes the rationale for companies to adopt the technologies. Besides, the study points out the challenges encountered by organizations in adopting the technology, thus, forming a foundation for further studies on the elimination of these hindrances to facilitate business development.

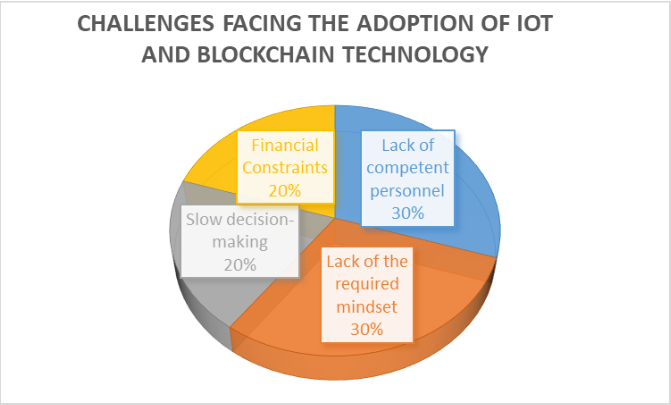


Fig. 4. Challenges Facing IoT and Blockchain.

V. Conclusion

The next step of development is smart cities, evidenced by the worldwide rural to urban migration, blockchain and IoT promise to deliver unprecedented development in all aspects of human society. Besides, the complementary nature of blockchain and IoT promises to solve most problems businesses face, especially in their adoption of new and untested business models. This paper proposed a study aimed to determine the developmental impact blockchain and IoT technologies will have on smart cities. Once complete, the research will reveal insight into issues such as the contributions of these two technologies as well as the challenges individuals and businesses will face when transitioning business models among others. Future studies will then build on this paper while expanding on the subject matter as well as larger sample sizes. Besides, time and financial constraints were very important factors in determining the study's sample size, despite its limitations.

References

[1] L. G. Anthopoulos and P. Fitsili, "Understanding smart city business models: a comparison," in Proc. of the 24th International Conference on WorldWide Web. Florence, 2015. doi: 10.1145/2740908.2743908.

[2] Schwab, Klaus, and Nicholas Davis. Shaping the future of the fourth industrial revolution. Currency, 2018.

[3] J. Ruan and Y. Shi. Monitoring and assessing fruit freshness in IOT-based e-commerce delivery using scenario analysis and interval number approaches, Information Sciences, 2016,373(1), 557-570.

[4] Lu, Jack. "Assessing The Cost, Legal Fallout Of Capital One Data Breach." Legal Fallout Of Capital One Data Breach (August 15, 2019) (2019).

[5] Stack, Brian. "Here's how much your personal information is selling for on the dark web." (2018).

[6] Stifter, Nicholas, et al. "Agreement with Satoshi–on the formalization of Nakamoto consensus." (2018).

[7] O. Sohaib, L. Haiyan and W. Hussain. "Internet of Things (IoT) in e-commerce: For people with disabilities," in 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA). Siem Reap, Cambodia, 2017. doi: 10.1109/ICIEA.2017.8282881.

[8] Kewell, Beth, and Peter Michael Ward. "Blockchain futures: With or without Bitcoin?." Strategic Change 26.5 (2017): 491-498.

[9] Liu, Stephanie Q., and Anna S. Mattila. "Airbnb: Online targeted advertising, sense of power, and consumer decisions." International Journal of Hospitality Management 60 (2017): 33-41.

[10] Aslam, Bilal, and Heikki Karjaluoto. "Digital advertising around paid spaces, E-advertising industry's revenue engine: A review and research agenda." Telematics and Informatics 34.8 (2017): 1650-1662.

[11] Manyika, James, et al. "Unlocking the Potential of the Internet of Things." McKinsey Global Institute (2015).

[12] Blanding, Michael. "The Internet Of Things Needs A Business Model. Here It Is". HBS Working Knowledge, 2019, https://hbswk.hbs.edu/item/the-internet-of-things-needs-a-business-model-here-it-is.

[13] Hopkins, Andrew. "Beyond compliance monitoring: new strategies for safety regulators." Law & Policy 29.2 (2007): 210-225.

[14] Reza Akhondi, Mohammad, et al. "Applications of wireless sensor networks in the oil, gas and resources industries." 2010 24th IEEE International Conference on Advanced Information Networking and Applications. IEEE, 2010.

[15] Royce, Rolls. The jet engine. John Wiley & Sons, 2015.

[16] Chaudhuri, Arindam. "Predictive maintenance for industrial iot of vehicle fleets using hierarchical modified fuzzy support vector machine." arXiv preprint arXiv:1806.09612 (2018).

[17] Yu, Jinying, and Wei Zhang. "Study on agricultural condition monitoring and diagnosing of integrated platform based on the internet of things." International Conference on Computer and Computing Technologies in Agriculture. Springer, Berlin, Heidelberg, 2012.

[18] Almotiri, Sultan H., Murtaza A. Khan, and Mohammed A. Alghamdi. "Mobile health (m-health) system in the context of IoT." 2016 IEEE 4th international conference on future internet of things and cloud workshops (FiCloudW). IEEE, 2016.

[19] Fleisch, Elgar, Markus Weinberger, and Felix Wortmann. "Business models and the internet of things." Interoperability and Open-Source Solutions for the Internet of Things. Springer, Cham, 2015. 6-10.

[20] Tzuo, Tien, and Gabe Weisert. Subscribed: Why the Subscription Model Will be Your Company's Future-and what to Do about it. Penguin, 2018.

[21] Lerner, Adam, et al. "Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016." 25th {USENIX} Security Symposium ({USENIX} Security 16). 2016.

[22] Mahmoud, Rwan, et al. "Internet of things (IoT) security: Current status, challenges and prospective measures." 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST). IEEE, 2015.

[23] Kwan, Martin. "Can Huawei Sue the US Government for Defamation? A Study on the Threshold of Foreign State Immunity from a Comparative Perspective." A Study on the Threshold of Foreign State Immunity from a Comparative Perspective (May 20, 2019) (2019).

[24] Puthal, Deepak, et al. "The blockchain as a decentralized security framework [future directions]." IEEE Consumer Electronics Magazine 7.2 (2018): 18-21.

[25] Hawlitschek, Florian, Benedikt Notheisen, and Timm Teubner. "The limits of trust-free systems: A literature review on blockchain technology and trust in the sharing economy." Electronic commerce research and applications 29 (2018): 50-63.

[26] Nowiński, Witold, and Miklós Kozma. "How can blockchain technology disrupt the existing business models?." Entrepreneurial Business and Economics Review 5.3 (2017): 173-188.

[27] Hwang, Junyeon, et al. "Energy prosumer business model using blockchain system to ensure transparency and safety." Energy Procedia 141 (2017): 194-198.

[28] Emmadi, Nitesh, and Harika Narumanchi. "Reinforcing Immutability of Permissioned Blockchains with Keyless Signatures' Infrastructure." Proceedings of the 18th International Conference on Distributed Computing and Networking. 2017.

[29] Salah, Khaled, et al. "Blockchain for AI: Review and open research challenges." IEEE Access 7 (2019): 10127-10149.

[30] Singh, Jatinder, and Johan David Michels. "Blockchain as a Service (BaaS): Providers and Trust." 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW). IEEE, 2018.

[31] Onik, Md Mehedi Hassan, and Mahdi H. Miraz. "Performance Analytical Comparison of Blockchain-as-a-Service (BaaS) Platforms." International Conference for Emerging Technologies in Computing. Springer, Cham, 2019.

[32] Tasca, Paolo. "Token-Based Business Models." Disrupting Finance. Palgrave Pivot, Cham, 2019. 135-148.

[33] Tapscott, Don, and Alex Tapscott. Blockchain revolution: how the technology behind bitcoin is changing money, business, and the world. Penguin, 2016.

[34] Chen, Yan. "Blockchain tokens and the potential democratization of entrepreneurship and innovation." Business Horizons 61.4 (2018): 567-575.

[35] Sallaba, M., D. Siegel, and S. Becker. "IoT powered by Blockchain– How Blockchains facilitate the application of digital twins in IoT." Deloitte Issue (2018).

[36] Morgan, Steve. "Cybersecurity Ventures predicts cybercrime damages will cost the world $6 trillion annually by 2021." Cybersecurity Ventures (2017).

[37] Ali, Muhammad Salek, Koustabh Dolui, and Fabio Antonelli. "IoT data privacy via blockchains and IPFS." Proceedings of the Seventh International Conference on the Internet of Things. 2017.

[38] Chauhan, Anamika, et al. "Blockchain and scalability." 2018 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C). IEEE, 2018.

[39] Otte, Pim, Martijn de Vos, and Johan Pouwelse. "TrustChain: A Sybil-resistant scalable blockchain." Future Generation Computer Systems (2017).

[40] Eyal, Ittay, et al. "Bitcoin-ng: A scalable blockchain protocol." 13th {USENIX} symposium on networked systems design and implementation ({NSDI} 16). 2016.

[41] Magazine, Bitcoin. "Malta Government To Enhance Public Transportation Using Blockchain Technology". Nasdaq.Com, 2018, https://www.nasdaq.com/articles/malta-government-enhance-public-transportation-using-blockchain-technology-2018-05-17.

[42] Fraile, Francisco, et al. "Trustworthy industrial IoT gateways for interoperability platforms and ecosystems." IEEE Internet of Things Journal 5.6 (2018): 4506-4514.

[43] "Chain of Things". Chain of Things, 2020, https://www.chainofthings .com/.

Automatic Detection of Plant Disease and Insect Attack using EFFTA Algorithm

Kapilya Gangadharan[1]
Research Scholar, Dept. of Computer Science and Engineering, SSE, Saveetha Institute of Medical and Technical Sciences Chennai, India

G. Rosline Nesa Kumari[2]
Professor, Dept. of Computer Science and Engineering
Shadan women's college of Engineering and Technology, Telangana, India

D. Dhanasekaran[3]
Professor, Dept. of Computer Science and Engineering
SSE, Saveetha Institute of Medical and Technical Sciences
Chennai, India

K. Malathi[4]
Associate Professor, Dept. of Computer Science and Engineering, SSE, Saveetha Institute of Medical and Technical Sciences Chennai, India

***Abstract*—The diagnosis of plant disease by computer vision using digital image processing methodology is a key for timely intervention and treatment of healthy agricultural procedure and to increase the yield by natural means. Timely addressal of these ailments can be the difference between the prevention and perishing of an ecosystem. To make the system more efficient and feasible we have proposed an algorithm called Enhanced Fusion Fractal Texture Analysis (EFFTA). The proposed method consists of Feature Fusion technique which combines SIFT- Scale Invariant Feature Transform and DWT- Discrete Wavelet Transform based SFTA- Segment Based Fractal Texture Analysis. Image as a whole can be detected by shape, texture and color. SIFT is used to detect the texture feature, it extracts the set of descriptors that is very useful in local texture recognition and it captures accurate key points for detecting the diseased area. Further extraction of texture is considered and that can be performed by WSFTA method. It adopts intra- class analysis and inter- class analysis. Extracted features trained using Back Propagation Neural Network. It improves and expands the success rate and accuracy of extraction also it provides higher precision and efficiency when compared to the other traditional methods.**

Keywords*—*Texture analysis; features; computer vision; inter-class; intra-class

I. Introduction

Digital images are typically represented in the form of texture, shape and color features. It indicates the attributes of the image. Dimensions of the raw data or information can be reduced by extracting the features. It is a process that is very much required in Machine Learning and pattern recognition. Where the required texture attributes are extracted from the specified dataset. The characteristic represents the original data property like texture, shape or color based upon the requirement. Texture is based on the feature and shape is based on the template Feature extraction mainly deals with reducing the large number of data into subset that describes the required information. The process makes the classification simpler and accurate. Machine learning methodology is mainly dependent on the appropriate and efficient feature extraction. Extraction begins with an underlying arrangement of already estimated information and determined values or features planned as instructive and non-repetitive, it accelerates the successive learning and speculation steps, also it is interpreted by human as well [16]. And it is identified with dimensionality reduction. SIFT process converts the image into smaller dimensions with the required information. It is mainly used for object matching criteria. Key point extraction through SIFT is more accurate than the other traditional methods, improvement shows 11.12 % more using SIFT [17]. SFTA is a used for image decomposition which utilizes OTSU thresholding in a conventional method. To make the decomposition much easier we follow Discrete Wavelet Transform and it is passed to STFA algorithm for further decomposition and it creates a hybrid method called WSFTA. When shift is combined with WSFTA it produces a best and accurate result and it is relatively easier for the classifier to recognize the diseased portion and the healthy portion of the plants. Our work mainly concentrates on proposing the enhanced texture feature fusion which combines SIFT and WSFTA, once after fusing the algorithms it is very important to perform selection process. Selection is done using PCA method to avoid dense feature vector creation that elaborates the vector length causing high computational cost [19]. The selection process results in the accurate and required features. The selected features are then trained and tested using Back Propagation Neural Network Classifier.

II. Related Work

The texture parameters are calculated using gray level synchronizing matrix spatial variants and to define the diseased area in the plant leaves [1]. Bark classification on texture and fractal dimension is proposed. It combines texture and the structural features to improve the accuracy [2]. Another method is proposed using the objective values, it is calculated by Kurtosis, variance, Skewness and Entropy, where homogeneity and contrast is calculated for estimating the diseased area [3]. GLCM (Gray level Covariance Matrix) is used to extract the texture features where 12 types of features are extracted which include Entropy, Skewness, Kurtosis, Smoothness, Variance, etc. [4]. A method called

HGPASO and it is combined with OCGR method is used to reduce the steps which is carried out to calculate the threshold by Multilevel OSTU method [5].

Image is processed using traditional Image processing methods and SIFT algorithm is passed through the preprocessed image to extract the color feature, where SIFT texture feature is described using Johnson SB distribution. Average precision is calculated using cross validation of 10-fold [6]. Analysis of key point extraction is done using SIFT and SURF extraction methods, it is proved that SIFT process is more accurate than SURF. The improvement is shown 11.12% more than SURF. It is classified using neural network classifier [7]. Shrinking Edge detection method is proposed and applied on recursive support vector mechanism to classify the image and to estimate the performance of the system [8]. Color histogram detects the diseased area of the image where other features are combined to generate the classification [9]. To classify and detect the diseased parts in the leaf images various methods like SIFT, HOG, Pyramid histogram word, Dense SIFT are compared [10]. SFTA extraction is composed based on the wavelet transformation, where image is decomposed into multiple sub bands using wavelet transformation and Texture feature is extracted using SFTA algorithm [11].

III. Proposed System

Feature extraction is an important tool in the Image processing field. Multiple features can be extracted from the digital image, the features referred as color, shape and texture. the plants can suffer from biotic plant diseases it could be due to pest attack or any kind of infection from fungus, bacteria or Virus. When there is a presence of disease or infection on any region of the plant say, leaf or bud or stem the diseased region will appear in different degree. Based on the difference in the characteristic of the diseased region we can estimate the change in color, shape or texture feature, also it can recognize and show the variation in each type of infection or pest attack. Before performing feature extraction method basic image processing techniques like Image acquisition, standardization, preprocessing.

Segmentation [15] should be performed so that extracting features would be much simpler and accurate. Fig. 1 describes the basic workflow of the proposed system. We have used multiple feature extraction methods to make the system more feasible. Performed Feature using SIFT and WSTFA algorithm. Feature fusion is a special technique that we have adapted in our proposed system it combines the extracted features of multiple methods using Enhanced Fusion Fractal Texture Analysis (EFFTA) algorithm.

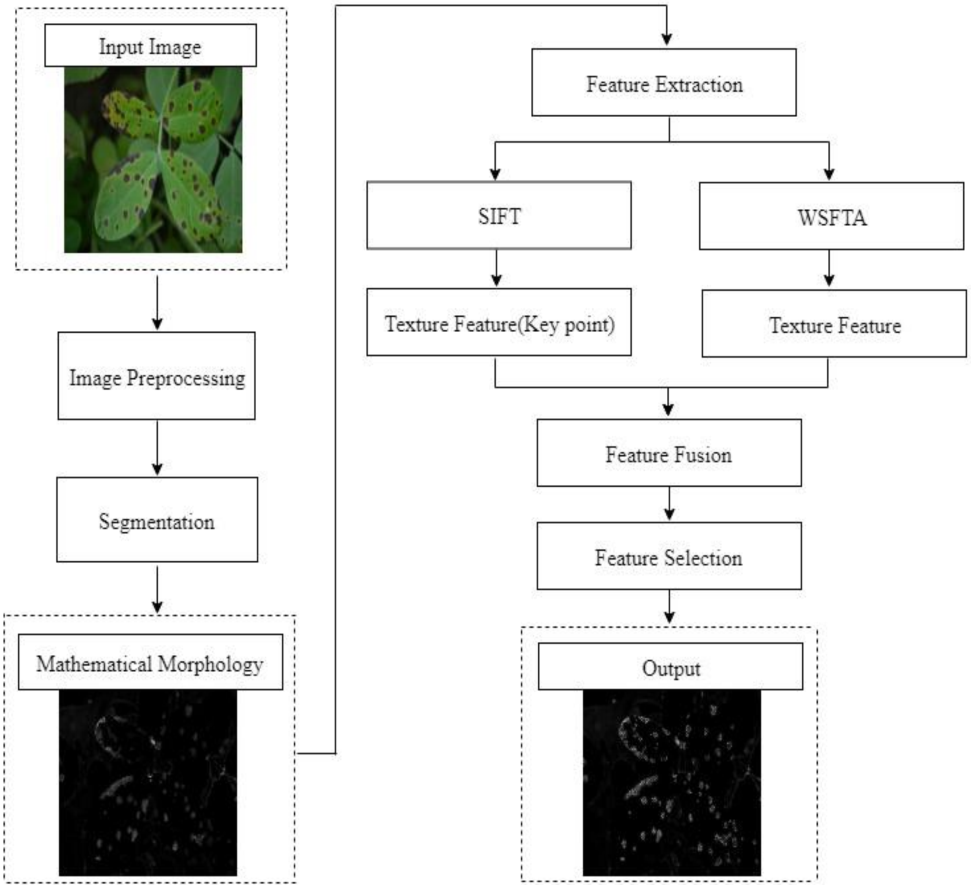


Fig. 1. Proposed System Architecture.

A. Proposed Algorithm

Step 1: Pick the image (Horizontal Gradient) extracted after Morphological Operation (Gray scaled Image)

Step 2: Apply SIFT Process:

a) Create an internal depiction of input image to check the scale variance through scale space.

b) Laplacian of Gaussian (LoG) approximation is used to find the key points. Since it is very expensive, we use DoG.

c) Calculate Difference of Gaussian (DoG), difference in Maxima and minima to estimate the key points.

d) Locate the extreme points and assign the SIFT feature.

e) Avoid or eliminate the Edges and low contrast i.e. poorly employed key points and appropriate key points are calculated.

f) Responsible key points must be assigned with orientation. Any further calculations will be done based on the orientation.

g) More depictions are generated which will be easy to identify the features.

h) Add the key points to a common list

Step 3: Mapping Discrete Wavelet Transform

a) Decompose input image into different frequency.

b) 2D-DWT is performed by applying 1D-DWT.

c) 1D-DWT is applied on the row of the Image and it is decomposed along the column.

d) If the result is 4 decomposition then the sub bands generated are LL, LH, HL and HH.

e) High pass filters produce detailed information on each level.

f) Scale factor is established by using Low pass filter.

g) Half band filter extracts signal spanning at each decomposition level. It multiplies the frequency intense into two and the unreliability in the frequency is reduced.

Step 4: Segmentation Based Fractal Texture Analysis (SFTA)

a) Low frequency sub band images are examined, which decomposes the image into binary image.

b) To decompose LL Image extracted from DWT, Two Threshold binary decomposition (TTBD) is utilized.

c) Using Fractal proportion from the examined region edges for every extracted binary image are calculated.

d) Finally, average of gray level and dimensions of region i.e., the boundaries(pixels) are calculated.

e) Add the extracted Wavelet based SFTA key points to a common list.

Step 5: Feature fusion is done for the features extracted by SIFT and WSFTA

a) Through the response value sort the common list.

b) Select the appropriate feature (diseased region) using Serial Based method.

Step 6: Perform Step 1 through Step 5 until all the features are extracted.

B. SIFT Key Point Extraction

Scale Invariant Feature Transform is an efficient algorithm that detects the local features of an image by extracting the Key points. The extracted Key points specifies the exact location and its descriptors. SIFT key points are extracted first from the set of reference images and it is saved in the database. The regions are perceived in the input image by independently comparing each component from the input image from the database. The detecting point of interest or component is the key point. A scale space is produced to represent scale variance in the input image.

Gaussian filter is applied on the Image under different scenario (scaling factor) and the difference obtained by the image which is blurred after applying Gaussian filter is taken. The Minimum and Maximum value from the difference is computed with multiple scaling factor which is also known as Difference of Gaussian (DoG).

$$D(x,y,\sigma) = L(x,y,\sigma_i) - L(x,y,\sigma_j) \quad (1)$$

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \quad (2)$$

Where, $D(x,y,\sigma)$ is the DoG Image, $L(x,y,\sigma)$ is the convolution, σ_i, σ_j is the Gaussian Blur $I(x,y)$ is the scaling factor of an image. The difference between σ_i and σ_j computed for different scales under multiple blurred condition on each pixel. The key points are differentiated with the threshold ranges 0.03, if the contrast is smaller than the threshold it is eliminated as low contrast values. DoG eliminates the edge with more accuracy, if the contrast is higher than the threshold the edge key points are rejected. So, it terminates the low contrast key point and boundary key point and it maintains only the strongest key points. Then each key point is allocated to more than one position using gradient factor extracted from the morphological operations when the positions are established. Then the magnitude and direction of each pixels are calculated with its neighboring pixel values. From which a histogram is generated, and the highest peak of the histogram is assigned to the key point, where more depictions are generated so as to make the system more feasible to identify the required features.

C. Feature Extraction using WSFTA Algorithm

Wavelet based SFTA feature extraction technique is proposed [12] to detect the infectious region from the input image. To effectively extract the texture feature we combine Discrete Wavelet Transform and SFTA as shown in the Fig. 2. The acquired feature vector will be the input for the classification algorithm to examine if the input image is identified with disease or not. Combining Wavelet Transform and SFTA algorithm it extracts 30 features from the low frequency sub bands.

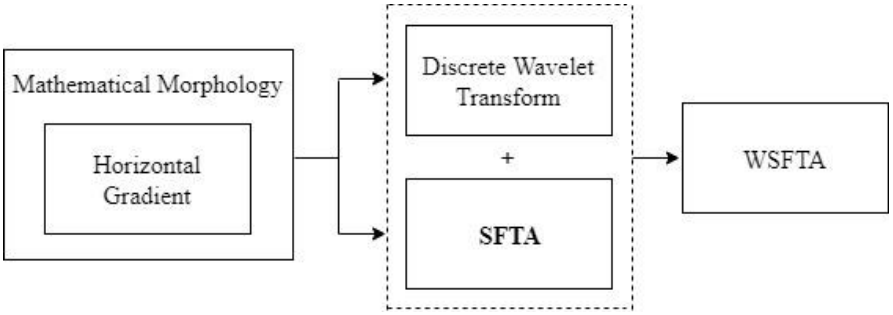


Fig. 2. Architecture of WSFTA.

a) Discrete Wavelet Transform: Feature extraction is one of the predominant aspects of classification. In this system, the extracted gradient through morphological method is sub divided into various frequency bands by applying Wavelet transform. DWT is a transformation tool that has ample of applications like feature extraction, compression, normalization, etc. It captures the frequency as well as the location information like texture, shape, etc. 2D-DWT decomposes the sample into four major sub bands, which is performed by 1D-DWT on the row of the input sample and then it is decomposed towards the column.

The result of 1D-DWT decomposition produces four types of sub band, they are Low Low (LL), Low High (LH), High Low (HL) and High High (HH). If wavelet functions are separable i.e., $f(x,y) = f_1(x)f_2(y)$, the scaling factor is low level frequency then the component of the previous scaling function is 2 dimension, which implies there is one 2D scaling factor and There are different wavelet functions $\psi^H(x,y)$, $\psi^V(x,y)$ and $\psi^D(x,y)$ for 2D scaling and filter can be applied to the wavelet function [13].

$$\phi(x,y) = \phi(x)\phi(y) \tag{3}$$

$$\psi^H(x,y) = \psi(x)\phi(y) \tag{4}$$

$$\psi^V(x,y) = \phi(x)\psi(y) \tag{5}$$

$$\psi^D(x,y) = \psi(x)\psi(y) \tag{6}$$

2D filters H_{LL}, H_{LH}, H_{HL} and H_{HH} are obtained from eq (1) to (4) and that corresponds to ϕ , ψ^H, ψ^V, ψ^D respectively. Although 2D approach involves more operation than 1D filtering this method is very efficient in implementing the feature extraction of low-level images.

$$f(x,y) = \frac{1}{\sqrt{MN}}\sum_{m}(\sum_{n}(W_\phi(j_0,m,n)\phi_{j_0,m,n}(x,y)) + \frac{1}{\sqrt{MN}}\sum_{i=H,V,D}(\sum_{j=0}^{\infty}(\sum_{m}(\sum_{n}(W_\psi^i(j,m,n)\phi_{j,m,n}^i(x,y)))) \tag{7}$$

Where

$$W_\phi(j_0,m,n) = \frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}(\sum_{y=0}^{N-1}(f(x,y)\phi_{j_0,m,n}(x,y))) \tag{8}$$

$$W_\psi^i(j,m,n) = \frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}(\sum_{y=0}^{N-1}(f(x,y)\phi_{j,m,n}^i(x,y))) \tag{9}$$

and $i = \{H,V,D\}$

b) Segmentation Based Fractal Texture Analysis (SFTA): SFTA Algorithm is mainly used to decompose the gray scaled image into multiple binary images and the output would be segmented regions which are used to define the texture patterns. So, that it will be easy for us to identify the diseased region and a normal region. The algorithm is split into two main specifications: 1. Two threshold binary decomposition in binary images. 2. SFTA calculates the fractal dimension from the boundaries of the regions. Also, pixels should be counted from all the extracted regions.

TTB Decomposition method**:** It takes gray scaled image $I(x,y)$ and split it into multiple binary images, where $I(x,y) = \{1,2,3,4,\dots.n_t\}$. Initially, Compute threshold value T, Threshold is determined by computing the equally spaced gray values in the image, Where the values range between 1 to n_t. It is done using multilevel OTSU method. The Multilevel OTSU method computes the threshold by minimizing the image intra-class variance and it is applied to each region that is split until the exact threshold is extracted.

$$I_b(x,y) = \begin{cases} 1\ if\ I(x,y) \geq T \\ 0\ , Otherwise \end{cases} \tag{11}$$

Where $I_b(x,y)$ is the binary image, T is the threshold and set the threshold obtained by equally spaced gray level value, which is called as One-Threshold Segmentation. Then the next step is to splits the gray scale image $I(x,y)$ into binary images. For which the extracted pair of thresholds Multiple T is selected and apply as Two-threshold segmentation, which computes the inter-class variance.

$$I_b(x,y) = \begin{cases} 1\ if\ T_A < I(x,y) \leq T_B \\ 0\ , Otherwise \end{cases} \tag{12}$$

Where T_A and T_B represents lower and upper threshold, after the threshold computation SFTA vector is constructed over it.

SFTA Extraction Algorithm: SFTA vector is constructed once after the threshold is computed using Two-Threshold segmentation. Where the boundaries of the fractal dimension and mean gray level is computed. The geometric phenomenon is engaged to explain the complexity and severity of the boundaries and the segmented structure. The boundary regions are $I_b(x,y)$ and the border image as $\delta(x,y)$.

$$\delta(x,y) = \begin{cases} 1\ if\ I(x,y) \in N_{30}\ (x,y) \\ I_b(x,y) = 0 \\ 0, otherwise \end{cases} \tag{13}$$

Where N_{30} is the pixel that has 30 connect to (x, y). If the pixel is in the position (x, y), then $\delta(x, y)$ picks the value 1 from the binary image $I_b(x, y)$, which is set to 1 with at least one neighboring pixel as 0, else $\delta(x, y) = 0$.

Algorithm:

```
Step1: $Threshold\ T \leftarrow Multilevel\ OTSU\ (I, n_t)$
Step2: $Threshold\ set\ 1\ from\ binary\ Image\ A \leftarrow \{\{t_i, t_{i+1}\}: t_i, t_{i+1} \in T, i \in [1,2,3, \dots [T] - 1]\}$
Step3: $Threshold\ set\ 2\ from\ binary\ Image\ B \leftarrow \left\{\{t_i, n_l\}: t_i \in T, i \in \left[1,2,3, \dots [T]\right]\right\}$
Step 4: $i \leftarrow 0$
Step5: $\left.\begin{matrix} for\ \{\ lower\ threshold, upper\ Threshold\}: \\ \{lowe\ threshold, Upper\ threshold\} \in A \cup B \end{matrix}\right\} \leftarrow$
 $do$
Step 6: $Binary\ Image\ I_b(x, y) \leftarrow$
    $Two\ threshold\ segmentation$
    $(I, lower\ threshold\ T_{A\ ,} Upper\ threshold\ T_B$
Step 7:
    $Border\ Image\ \delta(x, y) \leftarrow$
    $Find\ borders\ of\ I_b(Binary\ Image)$
Step 8:
    $SFTA\ Feature\ Vector\ [i] \leftarrow$
    $Count\ box\ of\ \delta(x, y)(Border\ Image)$
Step 9:
    $SFTA\ Feature\ Vector\ [j + 1] \leftarrow$
    $Mean\ gray\ level\ \left(I, I_b(Binary\ Image)\right)$
Step 10:
    $SFTA\ Feature\ Vector\ [j + 2] \leftarrow$
    $Pixel\ count\ of\ I_b(Binary\ Image)$
Step 11: $i \leftarrow i + 3$
Step 12: $end\ for$
Step 13: $Return\ SFTA\ Feature\ Vector$
```

D. Feature Fusion Methodology

It is an advanced method that is carried out in research to combine multiple features in one vector [12]. It upgrades the performance and precision level when compared with the individual analysis. Robust and the efficient features are combined by calculating the similarity between the extracted features on the same region. The fusion of multiple features has the objects on their own representation. It could be texture, shape color etc. Combining various features extracted by multiple methods provides better result and accuracy.

In our proposed system as show in Fig. 3 we combined the features extracted by SIFT algorithm is placed in a common list and likewise pixel count extracted by the WSFTA is placed in a common list. Fusion is done using Serial-Based Method. $f_n = \{{}^{n}/_{n} \in R^{SIFT}\}$ is the SIFT feature vector with dimension 1 X 2800 and $f_m = \{{}^{m}/_{m} \in R^{WSFTA}\}$ is the WSFTA vector with dimension 1 X 30. m, n are the feature vector. which is calculate using the sample data ϕ that perform training and testing. Let Δ= $(\phi/\phi) \in RS$ and fusion can be done as αf=nm.

We get the vector values from the common list when the features of SIFT and WSFTA is placed. f_n is the vector for SIFT and f_m is the vector for WSFTA. So, we calculate $f_v = \{f_n + f_m\}$ where the dimensions of the extracted vectors are known. Feature selection is carried out by adding the minimum distance and Pearson co-efficient. The selection method produces perfect and flawless feature from the combined vector, and it minimizes the elapsed time during execution. The selection method involves Euclidian distance (ED) calculation. Where threshold T is computed from the combined feature vector, extracted through the minimum distance feature $D(f_v)$. Mean of Minimum distance is used to calculate the skewness which requires the median value from the given set of vectors and the variance of Minimum distance.

$$E = -\sum_{v}(f_v \log_2(f_v)) \tag{14}$$

Entropy is calculated by obtaining Euclidian Distance, Minimum Distance, Skewness and Variance where PCA is applied on the extracted features for obtaining optimized features. Select the highest score from the calculated entropy to recognize the best feature of diseases part of the plant which is on observation. The Training algorithm considered in this proposed method is Back Propagation Neural network. Classifier which recognizes the infection by training and testing process.

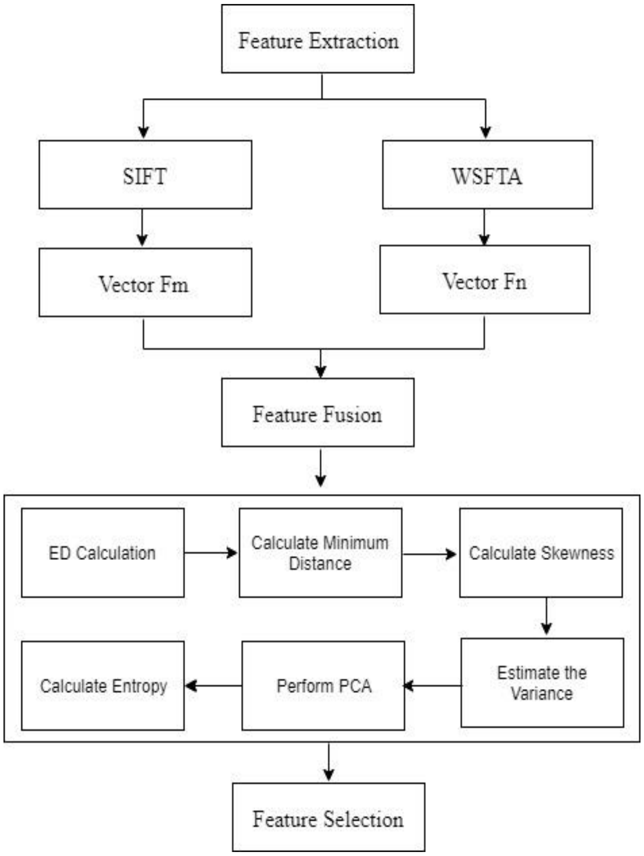


Fig. 3. Feature Fusion Method.

E. Classification Process

The selected features are trained and tested using Back Propagation Neural Network Technique. Where 80% of the image from the dataset is trained and 20% of the image from the dataset are tested using BPNN. Back propagation calculates the missed errors by performing back loops. It is based on the principle square of Euclidean distance and the calculated entropy. The internal memory is used for processing the output sequence and it showcases the dynamic temporal behavior of the system. Grading the learning system is the most important factor. For better and accurate result an optimum number should be considered for grading. Momentum is multiplied for the number of backward iterations so as to find the discrepancies in the hidden layer.

Algorithm:

```
Step 1: Allot the required input and Output
Step 2: Consider the weight from the range -1 and 1.
 Repeat Step 2 for all the available pattern in the training object.
Step 3: Initial value is predicted by performing Forward propagation.
Step 4: The discrepancies in the hidden layers are identified through backpropagation.
Step 5: Calculate local and global error.
Step 6: Repeat the steps until (maximum number of iterations < specified) and (Error function > specified).
Step 7: Increase the momentum from -0.5 to 0.5 and conduct multiple training sessions.
```

IV. Experimental Results and Discussion

The proposed method is tested and analyzed on various plant diseases and pest attacks using the dataset captured from internet [14, 18]. The experimental results include the performance of individual algorithm and the fusion and selection method. we have tested our algorithm using MATLAB to code and executed on Windows 10, 64-bit Operating System with 8 GB Ram, 2.70 GHz Intel Core i5 7th Gen Processor.

A. Performance Comparison on Scaling and other Factors

At first, we have evaluated the performance of the Fusion texture by comparing it with individual algorithm by changing the scale ranges from 0.25 to 2.0. From Fig. 4, it is observed that the proposed method Feature Fusion which is the combination of SIFT and WSFTA algorithm obtained correct match better than the other existing individual algorithms like SIFT, SFTA, BRISK, WSFTA. We have picked the normal and changed the scaling factor to test the correct match. We have tested on multiple images and taken average matching factor from total number of 300 Images with multiple diseases. The enhancement on the precision is better and accurate using proposed algorithm. Which has the robust factor for detecting the feature by extracting the maximum number of Key points.

The same way matching criteria is calculated using other factors like Rotation, Affine transform, Illumination, Gaussian Blur effect. These tests are carried out to check if we can get exact match on the proposed system, because we have fused more than one algorithm, while combining more than one texture feature method it is very important to notice if it produces good matching criteria. We adapted similarity matching factor using Serial Based method. Rotation is another major factor that is used in feature detection. We have compared the set of images with multiple degree of rotation. And it is observed that proposed method excelled in various rotation criteria starting from 30 to 60 degree as shown in Fig. 5.

Also, when the test was performed by applying affine transform, Gaussian blur and Illumination the proposed method produced better match when compared to the existing methods which is depicted in Fig. 5. When the algorithms are combined it produces better result than the individual processing.

B. Key Point Extraction

Extracting key point is one of the critical and important aspect in feature detection or extraction. In our proposed system before feature fusion, we have collected the extracted features through Key points using SIFT method. We have compared SIFT, SURF and BRISK on the number of key points extracted on a specific threshold. The observation shows that the key points obtained by applying, provides the better ways to detect and classify the diseased portion. Average key points extracted are shown in Fig. 6. Higher number of key points extracted produces better result. When we compared the three methods it is proven that SIFT produces higher key points than the other two methods.

C. Feature Analysis

The feature extraction of various images and diseases are calculated with reference to multiple texture features. We have extracted the features for multiple diseases like Cercospora leaf spot on Peanut leaves, Bacterial Blight disease on Paddy leaves and few pest attacks like, Boll Weevil on Cotton buds, European Corn Borer on Cotton leaves and Fall army Worm on the peanut leaves. We have calculated the scores of the extracted features separately by using the fusion method. Where few features are examined like ED, Variance, Skewness, Kurtosis and Entropy. After the feature extraction we have spotted the diseased area by selecting the features. Calculate entropy value for each feature and select the highest score among the list and the highest score is used for recognition. Table I shows the average calculated scores. And the highest Entropy values for individual diseases. We have used more than 300 Images for the calculation.

D. Implementation Result

The input images are collected from the datasets available in the internet [14]. In the future we are planning to use the image captured in drone with high resolution camera. We have performed morphological operations on the segmented image and selected the horizontal gradient. In this paper we have used the extracted Horizontal Gradient Image as the input image and we have performed SIFT operation to extract the key point and decomposition is done by Wavelet transform and SFTA Algorithm and extracted the texture features. And both the algorithms are combined using Fusion feature method and based on the highest Entropy value we have selected the diseased region as shown in Fig. 7 to Fig. 11.

E. Comparitive Evaluation

The proposed feature extraction technique is compared with the traditional state of art methods with slight variation in the dataset. Overall accuracy is considered for comparison since the identified disease using the other traditional methods are different. The proposed system is trained using BPNN and the other tradition methods are trained using SVM and BPNN. From the analysis it is determined that the proposed system produces better result with the accuracy of 97.69% as shown in Table II.

The performance has been evaluation of the proposed system is based on the valuation metrics. It can be classified as

True Positive (TP): Abnormal case appropriately evaluated as Abnormal.

False Negative (FN): Abnormal case imperfectly classified as normal.

False Positive (FP): Normal case imperfectly identified as abnormal.

True Negative (TN): Normal case appropriately calculated as Normal.

Sensitivity: TP/TP+FN (15)

Specificity: TP/TP+FP (16)

Accuracy: (TN+FP)/(TN+TP+FN+FP) (17)

F-Score: (2*Sensitivity*Specificity)/(sensitivity+specificity) (18)

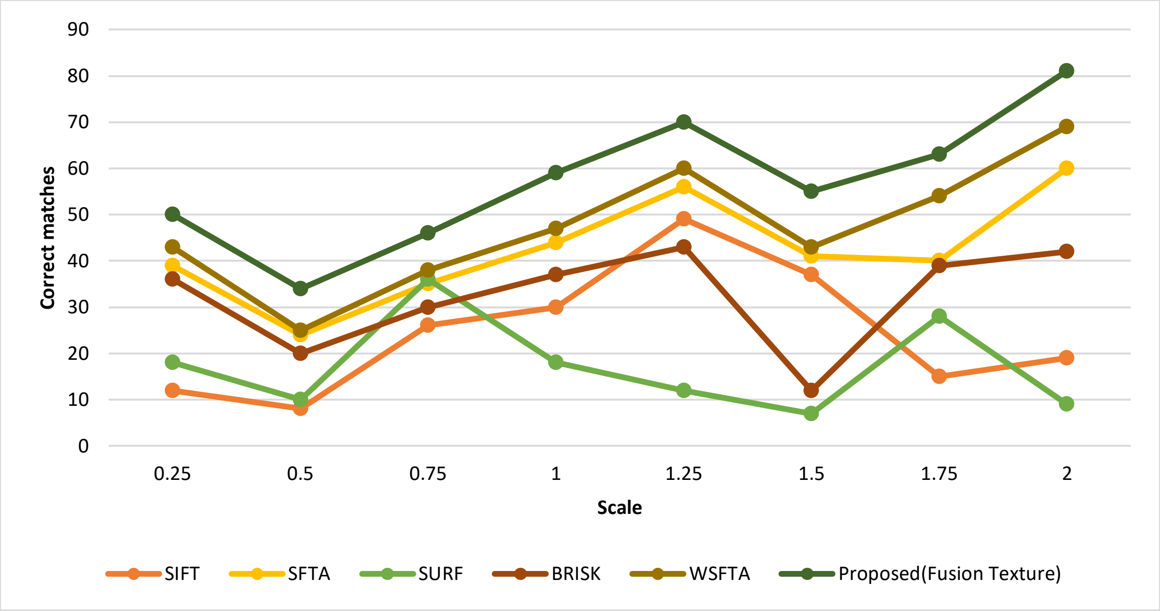


Fig. 4. Performance Comparison using Multiple Scaling Fctor.

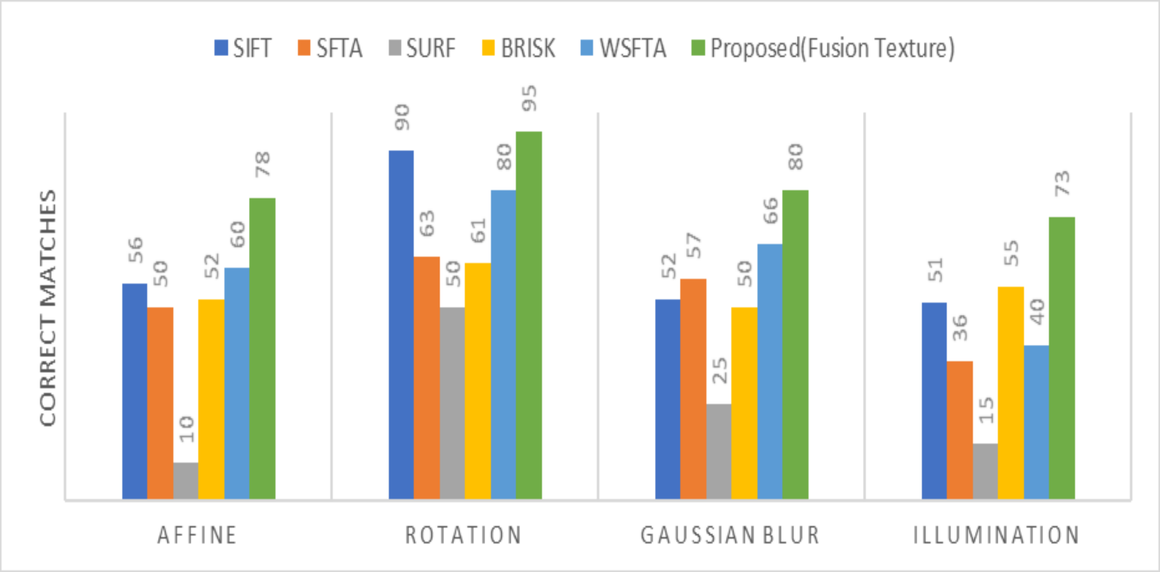
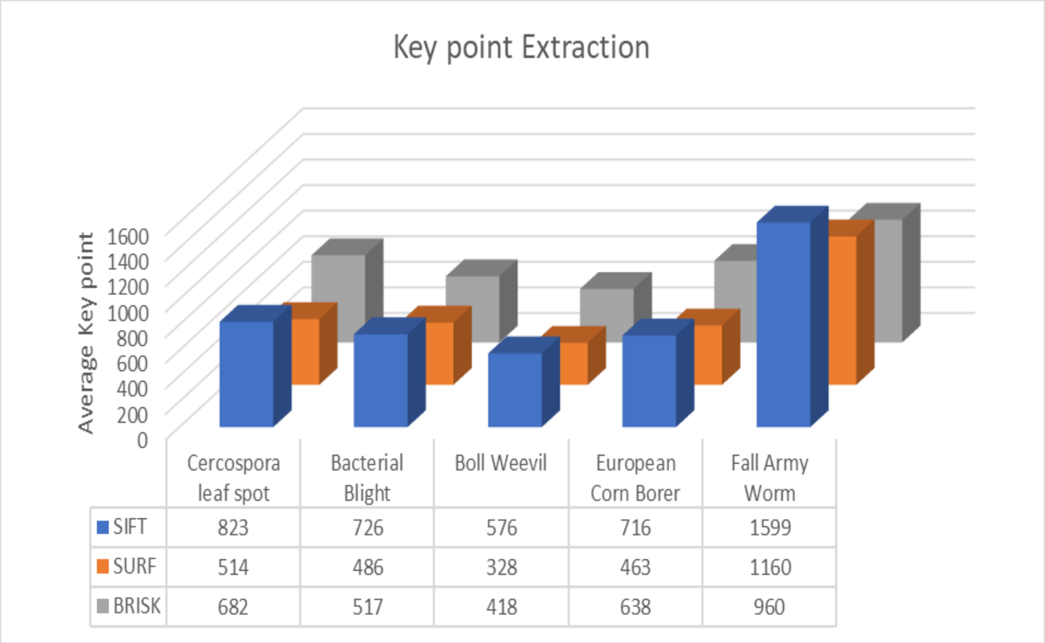


Fig. 5. Performance Evaluation using Multiple Factors.



	Cercospora leaf spot	Bacterial Blight	Boll Weevil	European Corn Borer	Fall Army Worm
🟦 SIFT	823	726	576	716	1599
🟧 SURF	514	486	328	463	1160
⬜ BRISK	682	517	418	638	960

Fig. 6. Comparison of Key Point Extraction.

TABLE. I. EXTRACTED FEATURES

Features	**Peanut leaves**	**Paddy Leaves**	**Cotton Buds**	**Cotton leaves**	**Peanut leaves**
	Cercospora leaf spot	***Bacterial Blight***	***Boll Weevil***	***European Corn Borer***	***Fall Army Worm***
Euclidian distance	48.3456	23.8312	11.5674	17.6571	33.6172
Variance	45.7434	27.8767	15.5633	19.4415	35.9787
Skewness	1.5001	2.5143	4.2657	2.9349	4.1015
Kurtosis	4.6343	8.4565	22.7562	10.2745	15.3851
Entropy	4.1004	2.5647	1.2109	1.9111	3.154

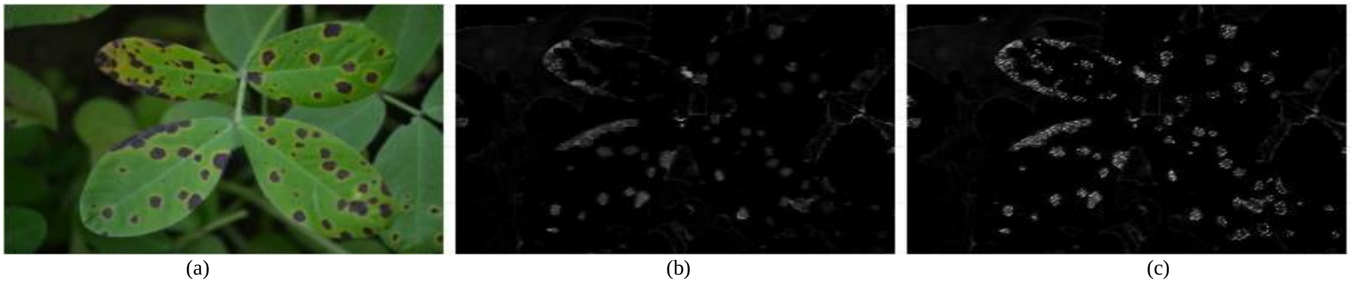

(a) (b) (c)

Fig. 7. Feature Extraction of Cercospora Leaf Spot on Peanut Leaves (a) Input Image (b) Horizontal Gradient Extracted by Mathematical Morphology (c) Extracted Features using Proposed Method.

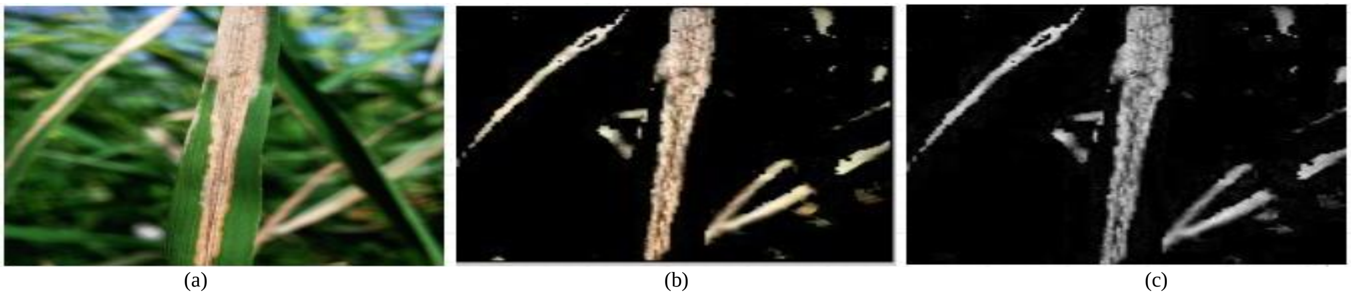

(a) (b) (c)

Fig. 8. Feature Extraction of Bacterial Blight Disease on Paddy Leaves (a) Input Image (b) Horizontal Gradient Extracted by Mathematical Morphology (c) Extracted Features using Proposed Method.

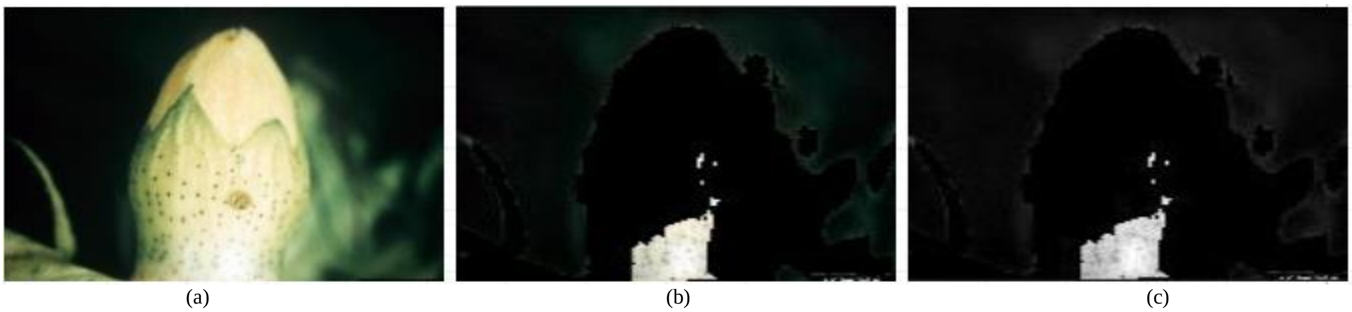


Fig. 9. Feature Extraction of Boll Weevil Attack on Cotton Buds (a) Input Image (b) Horizontal Gradient Extracted by Mathematical Morphology (c) Extracted Features using Proposed Method.

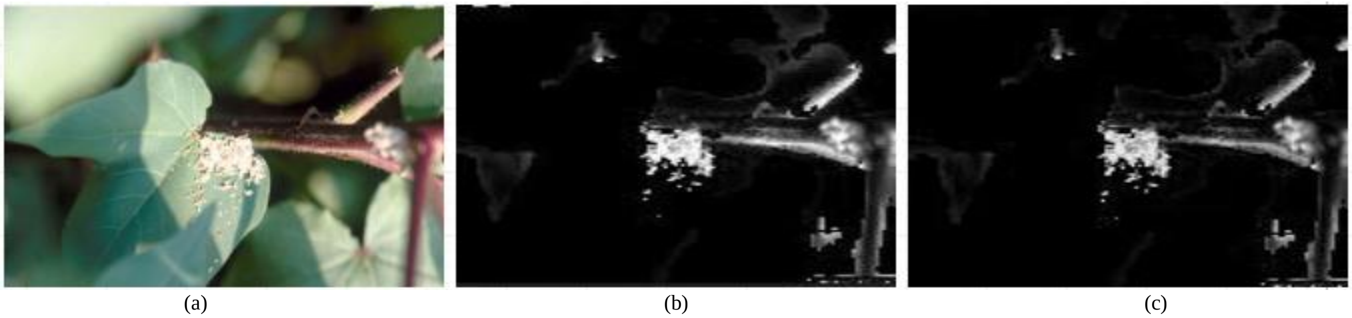


Fig. 10. Feature Extraction of European Corn Borer attack on Cotton Leaves (a) Input Image (b) Horizontal Gradient Extracted by Mathematical Morphology (c) Extracted Features using Proposed Method.

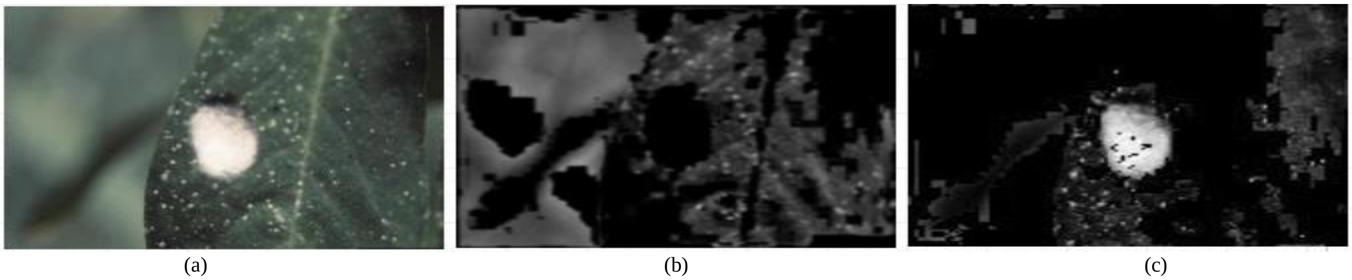


Fig. 11. Feature Extraction of Fall Army Worm attack on Peanut Leaves (a) Input Image (b) Horizontal Gradient Extracted by Mathematical Morphology (c) Extracted Features using Proposed Method.

TABLE. II. COMPARITIVE EVALUATION

Method	**Disease/Insect Attack**	**Accuracy**
Proposed Algorithm [FFTA+BPNN]	Cercospora, Bacterial Blight, Boll Weevil, fall Army Worm, European Corn Borer	97.69%
GLCM+SVM [9]	Bacterial Leaf Blight, Sheath Blight and Leaf Blast	97.02%
GLCM + BPNN [20]	Frogeye Disease, Downy Mildew and Bacterial Pustule	93.03%
GLCM+BPNN [1]	Bacterial Blight, Leaf Blast and Brown Spot	89.60%
Hu's Moment + BPNN [21]	Myrothecium, Bacterial Blight, and Alternaria	85.52%

V. CONCLUSION

Plant disease detection is an important aspect in crop management for a stable crop production. In our proposed work, we have examined Enhanced Fusion Fractal Texture Analysis Technique for feature extraction. We have driven out a texture feature and selection method which combines SIFT and WSFTA extraction techniques and selects the best feature by performing PCA algorithm for further classification. From the testes that we have performed it is proven that the proposed method produces best result of 97.69% when extracted features are classified using BPNN. Compared to the existing traditional methods the proposed system has higher accuracy. Fusing more than two existing methods produces better result. We have used serial based fusion technique which extracts the similarity criteria among the individual algorithms. The tests were performed on more than 300 images which include multiple diseases and pest attack. The proposed method excelled in different factors including change of scale, Illumination, Gaussian blur effect etc. Including all the extracted features we have spotted the diseased regions on the leaves and the buds. It gives better result in low scaled images as well.

REFERENCES

[1] Huang KY (2007) ,"Application of artificial neural network for detecting Phalaenopsis seedling diseases using color and texture features," Computers and Electronics in Agriculture 57(1):3–11.

[2] Zhi-Kai Huang, Chun-Hou Zheng1, Ji-Xiang Du1, and Yuan-yuan Wan,Bark "Classification Based on Textural Features Using Artificial Neural Networks," ISNN 2006, LNCS 3972, pp. 355 – 360, 2006. © Springer-Verlag Berlin Heidelberg 2006.

[3] Dr. Shaik Asif Hussain, Raza Hasan, Dr. Shaik Javeed Hussain, "Classification and Detection of Plant Disease using Feature Extraction Methods," International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 4219-4226 © Research India Publications. http://www.ripublication.com.

[4] Ashwini T Sapkal, Uday V Kulkarni, "Comparative study of Leaf Disease Diagnosis system usingTexture features and Deep Learning Features, International Journal of Applied Engineering Research, " ISSN 0973-4562 Volume 13, Number 19 (2018) pp. 14334–14340 © Research India Publications, http://www.ripublication.com.

[5] Md. Junayed Hasan , Jia Uddin, "A Novel Modified SFTA Approach for Feature Extraction," iCEEiCT 2016, 978-1-5090-2906-8/16/$31.00 ©2016 IEEE.

[6] Chit Su Hlaing, Sai Maung Maung Zaw, "Tomato Plant Diseases Classification Using Statistical Texture Feature and Color Feature," 978-1-5386-5892-5/18/$31.00 ©2018 IEEE, ICIS 2018, June 6-8, 2018, Singapore.

[7] K.Malathi, R.Nedunchelian, "Efficient Method To Detect And Classify Diabetic Retinopathy Using Retinal Fundus Images," International Journal of Pure and Applied Mathematics, Volume 116 No. 21 2017, 89-97 ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version).

[8] K Malathi1, R Nedunchelian, "A recursive support vector machine (RSVM) algorithm to detect and classify diabetic retinopathy in fundus retina images," Biomedical Research 2017, ISSN 0970-938X.

[9] Yao Q, Guan Z, Zhou Y, Tang J, Hu Y, Yang B (2009), "Application of support vector machine for detecting rice diseases using shape and color texture features," In: IEEE international conference on engineering computation ICEC, Hong Kong, pp 79–83.

[10] Pires RDL, Gonc¸alves DN, Orue ˆ JPM, Kanashiro WES, Rodrigues JF, Machado BB, Gonc¸alves WN (2016), "Local descriptors for soybean disease recognition", Comput Electron Agric 125:48–55.

[11] D.Saraswathi,G.Sharmila, E.Srinivasan, "An Automated Diagnosis system using Wavelet based SFTA Texture Features," ICICES2014, ISBN No.978-1-4799-3834-6,2014 IEEE.

[12] Attique Khan, Tallha Akram,Muhammad Sharif,Muhammad Younus Javed,Nazeer Muhammad, Mussarat Yasmin, "An implementation of optimized framework for action classification using multilayers neural network on selected fused features," Muhammad, Pattern Analysis and Applications, Springer-Verlag London Ltd., part of Springer Nature 2018.

[13] Chun-Lin, Liu , "A Tutorial of the Wavelet Transform," February 23, 2010, http://disp.ee.ntu.edu.tw/tutorial/WaveletTutorial.pdf.

[14] Plant Village Images, https://plantvillage.psu.edu/topics, Accessed 17 October 2019.

[15] Kapilya Gangadharan, G. Rosline Nesa Kumari, D. Dhanasekaran, "An Efficient Plant Disease Detection System Using Hybrid Watershed Segmentation with Extended K-Means Clustering Algorithm," International Journal of Advanced Science and Technology, Vol. 28, No. 11, (2019), pp.308-320).

[16] Kapilya Gangadharan, G. Rosline Nesa Kumari, D. Dhanasekaran, "Classification and Functional Analysis of Major Plant Disease using Various Classifiers in Leaf Images," International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-9 Issue-2, December 2019.

[17] Kapilya Gangadharan, G. Rosline Nesa Kumari, D. Dhanasekaran, K.Malathi, "Plant Disease Diagnosis and Classification by Computer Vision using Statistical Texture Feature Extraction Technique and K Nearest Neighbor Classification," International Journal of Engineering and Advanced Technology (IJEAT), ISSN: 2249 – 8958, Volume-9 Issue-2, December, 2019.

[18] ForestryImages, https://www.forestryimages.org/series/series.cfm, Accessed23October 2019.

[19] Prasad S, Peddoju SK, Ghosh D (2016)," Multi-resolution mobile vision system for plant leaf disease diagnosis." Signal Image Video Process 10(2):379–388.

[20] Gharge S, Singh P (2016), "Image processing for soybean disease classification and severity estimation." In: Shetty N, Prasad N, Nalini N (eds) Emerging research in computing, information, communication and applications. Springer, New Delhi, pp 493–500.

[21] Rothe PR, Kshirsagar RV (2015)," Cotton leaf disease identification using pattern recognition techniques." In: IEEE international conference on pervasive computing (ICPC), January, pp 1–6.

A Robust Deep Learning Model for Financial Distress Prediction

Magdi El-Bannany[1]
Department of Accounting
College of Business Administration
University of Sharjah 27272, UAE

Meenu Sreedharan[2], Ahmed M. Khedr[3]
Department of Computer Science
College of Computing and Informatics
University of Sharjah 27272, UAE

Abstract—**This paper investigates the ability of deep learning networks on financial distress prediction. This study uses three different deep learning models, namely, Multi-layer Perceptron (MLP), Long Short-term Memory (LSTM) and Convolutional Neural Networks (CNN). In the first phase of the study, different Optimization techniques are applied to each model creating different model structures, to generate the best model for prediction. The top results are presented and analyzed with various optimization parameters. In the second phase, MLP, the best classifier identified in the first phase is further optimized through variations in architectural configurations. This study investigates the robust deep neural network model for financial distress prediction with the best optimization parameters. The prediction performance is evaluated using different real-time datasets, one containing samples from Kuwait companies and another with samples of companies from GCC countries. We have used the technique of resampling for all experiments in this study to get the most accurate and unbiased results. The simulation results show that the proposed deep network model far exceeds classical machine learning models in terms of predictive accuracy. Based on the experiments, guidelines are provided to the practitioners to generate a robust model for financial distress prediction.**

Keywords—Financial distress prediction; multi-layer perceptron; long short-term memory; convolutional neural network; deep neural network; optimized deep learning model

I. Introduction

Financial distress is a condition where a company faces financial difficulties, which is also referred to as Business or Corporate Failure. Financial distress can induce a great impact on any company, stakeholders, and the economy of a country. The investors rely on financial statements disclosed by any company, which can be forged by the company executives, leaving them with very little chance of getting the original financial information. Hence, a reliable distress prediction model is necessary for investors to adjust their investment strategies, so as to reduce the loss of investments. Also, it will help the company managers to take corrective measures to prevent the crisis before it happens. Many researchers have used primitive statistical techniques to generate relevant models, however, machine learning algorithms were found to create a more robust model for distress prediction.

Building a robust model with acceptable prediction performance can help the managers and investors to manage risks and take actions on time, to prevent bankruptcy before it happens. Deep learning is a field of machine learning, containing multiple layers of nonlinear processing units, to learn features from real-time data. With low cost, high computational ability and availability of different optimization techniques, deep learning has been an area of interest for many types of research. So, the question is: Can we create a robust prediction model using deep learning techniques? Many research works are available on distress prediction using classical machine learning classifiers and ensemble techniques like Decision Tree [1,4], Neural Networks [5-9], Support Vector Machines [13], etc. But to the best of our knowledge, there is no existing research work analyzing different deep learning models on financial distress prediction and how to optimize these models. Hence, this paper provides an insight into the deep learning models for financial distress prediction which will be of great significance to generate a more robust model.

This paper focuses on building deep neural networks including multi-layer perceptron, LSTM, and Convolutional Neural Network and optimization of the same using different optimization techniques. We plan to train the networks using different transfer functions and training algorithms to generate a robust model for distress prediction. In the second phase of the study, we try to further optimize the models using different architectural configurations by varying the number of deep layers and neurons at each level. In this paper, we focus on two different datasets, one from the companies in Kuwait and other from companies in GCC, to analyze the performance with real-time and varying structures. The experiments in this paper mainly focus on providing proper guidelines for any researcher, to build a robust deep learning model for financial distress prediction in the future. The results of the study clearly indicate that the proposed model has significantly higher predictive accuracy compared to classical machine learning models.

This paper is organized as follows: The next section reviews the studies related to financial distress prediction. Research methodology: data collection and data modeling are described in Section 3. The prediction performance of all the selected deep models is described, optimized, analyzed and compared in Section 4. Finally, in Section 5, we infer our conclusion, with a set of guidelines to generate a robust model for financial distress prediction using deep neural networks.

II. Literature Review

The major aim of a Financial Distress Prediction Model is to determine whether a company has a chance of experiencing financial distress in the future. Bankruptcy, Insolvency, etc. are the formal signs of financial distress in a company. Discriminant analysis [3] and logit model [12, 14] are the initial and traditional statistical models used in the field of distress prediction. These traditional linear techniques are simple but unrealistic, and therefore cannot be used to generate a robust model for making real-time predictions. In 2014, a simple hazard model for the distress prediction of banks in the Gulf Cooperation Council countries was built [2]. Machine learning using data mining techniques like Logistic Regression, Support Vector Machines [11, 13] and Neural Networks [15, 17, 19, 20] were introduced as alternatives in later researches. In 2015, Ruibin Geng, Indranil Bose, and Xi Chen evaluated the performance of machine learning techniques for the distress prediction of listed Chinese companies [21]. The paper compared the three highly used data mining classifiers and evaluated the performance by combining the results using Majority Voting. Researches on datasets, collected from different countries, using data mining algorithms are also available [10]. In 2019, an analysis of a two-stage model for distress prediction is studied in [16]. It basically focuses on feature selection as a critical step, through data envelopment analysis. Distress prediction using deep learning is presented in [18], which uses unstructured textual data in statements for prediction. This problem can be solved in case the data sets are distributed among a number of sites cross the neworks [22-30].

To the best of our knowledge, no researches are available in the literature, investigating the performance of various deep learning network models on financial distress prediction and how to optimize them. In this paper, we evaluate the deep neural network models for financial distress prediction using two different datasets. We also apply the various optimization techniques on deep models to generate a robust model. As the last phase of this study, the results of the proposed model are compared with that of classical machine learning classifiers like support vector machine and decision tree. This paper helps researchers to develop a robust model for financial distress prediction using conclusions derived at the end of this paper.

III. Research Methodology

This section explains data modeling, the algorithm, how the model is evaluated and the datasets used for simulation. The data modeling section is divided into two phases. In the first phase of the study, we analyze the performance of deep neural networks on financial distress prediction. We also investigate the predictive performance of the models by optimizing the models using various optimization techniques. In this phase, we select the best performing model obtained from the combination of optimization and activation functions for further analysis. Phase 2, focuses on optimizing the outstanding model from phase 1 by restructuring the architectural configurations i.e., Network depth and Network width All the experiments in this study apply the technique of resampling using k-fold evaluation metrics, to get unbiased and most accurate results. A schematic representation of the steps involved in this study is shown in Fig. 1.

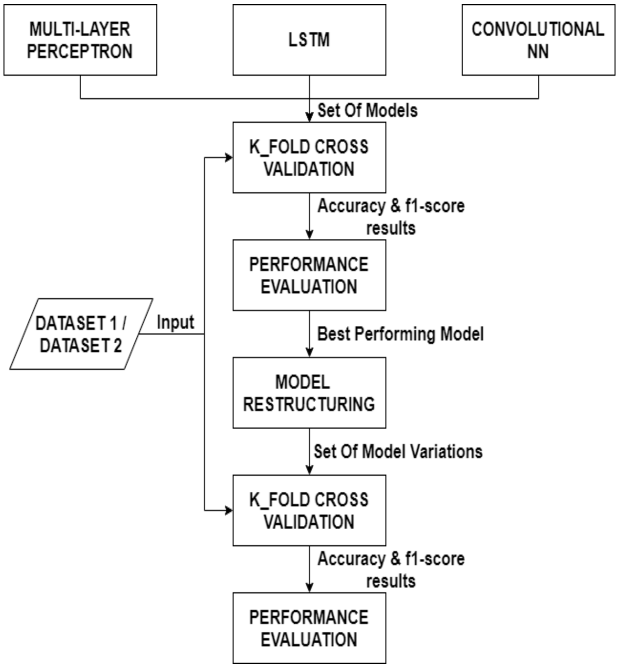


Fig. 1. Schematic Diagram Showing the Steps of Research.

A. Data Collection and Financial Indicators

1) Dataset1: The primary dataset used for prediction performance analysis contains sample data collected from the companies in Kuwait. The dataset, referred to as dataset1, contains 64 sample companies with balanced data, 32 financially healthy and 32 financially distressed companies during the period of 2010 to 2017. Dataset1 contains 24 financial indicators or attributes extracted from financial statements and balance sheets of respective companies.

2) Dataset2: A second dataset, with sample data collected across different countries, is used for modeling in order to verify whether the models can identify the common dynamics across datasets. The second dataset, referred to as dataset2, contains 120 samples from six GCC countries including UAE, Bahrain, Kuwait, Qatar, Saudi Arabia, and Oman. In order to create a balanced data sample, an equal number of financially healthy and financially distressed companies are included in dataset2 as in dataset1. Hence, dataset2 contains 60 healthy and 60 distressed companies' data collected from the balance sheets and financial statements of respective companies during the period of 2010 to 2017. Since the data contains samples collected across different countries, the number of financial indicators is only 19, less when compared to dataset1, due to missing common attributes among countries.

B. Data Interpretation and Preprocessing

The problem addressed in this paper is a binary classification problem, to determine whether a company can be labeled as financially distressed or not. The output/target attribute in the financial dataset belongs to two classes: one for financially distressed and the other for financially healthy companies. Except for this target attribute, which is binary, all the other attributes in the dataset are continuous values. The initial datasets contain incomplete samples and those with missing data and null values, which are removed during preprocessing.

C. Data Modeling

1) Artificial neural network: A neural network is a machine learning classifier based on an artificial representation of the human brain. Neural Network Architecture consists of an input layer, a layer of output nodes, and one or more intermediate layers.

The corresponding weights are multiplied with the input to calculate Y as,

$$Y_k = w_0 + \sum_{j=1}^{n} w_{jk} x_j$$

where Y_k is a weighted sum of input signals at node k; w_0 is bias value; w_{jk} is, the weight associated with the connection between node k and the input node j; x_j is a value of input node j; n is the number of input nodes. The weighted sum output is then served as input to an activation function.

$$(Y) = \frac{1}{1 + e^{-Y}}$$

The value after applying the activation function is the output value from node k, which is considered as the input to the next layer in the architecture.

Phase 1: In this phase, three highly preferred deep learning network models namely MLP, LSTM, and CNN are developed, trained and tested for financial distress prediction. All three models are trained with different activation and optimization algorithms with a fixed number of neurons at each level, to evaluate the prediction performance. We have generated different variations of each deep learning model through different combinations of transfer functions and optimization algorithms. The different transfer functions used in this study include sigmoid or logistic function, reLu or Rectified linear units and tanh or hyperbolic tangent and the optimization functions are stochastic gradient descent, Adam and Adagrad optimizers. The generated model variations were then trained and tested using two datasets. This phase aims to investigate the best deep neural network model for financial distress prediction, which is further optimized in phase 2.

Phase 2: A variation of predictive performance was observed with different layers of MLP in Phase 2. Hence the second phase of our study focuses on further evaluation of Multi-Layer Perceptron, the outstanding model from phase 1 in terms of accuracy and f1-score for different architectural configurations. There exists a myriad of hyperparameters that can be tuned to improve the predictive performance of a deep neural network. We focus on tuning the main two hyperparameters namely Network depth and Network width, which can make a difference in the algorithm exploding or converging. MLP models are designed by varying the number of hidden layers and the number of neurons at each level. The aim of this phase is to investigate the optimum parameters for Network depth and Network width, to generate a robust model for financial distress prediction. Pythons Scikit-learn and Keras packages are used for training the models and to generate the results. We have used the resampling technique called Cross-validation for performance evaluation, which is further discussed in Section 3.4.

Finally, we have also compared the performance of the proposed model with that of classic machine learning models. The simulation results indicated that the proposed model has significantly higher predictive accuracy when compared to support vector machine and decision tree classifier models.

D. Evaluation

In this paper, we have used Keras, the most powerful deep learning library in python, to build and evaluate deep learning models. The models are evaluated using k-fold cross-validation in pythons scikit-learn. Thus, we use the technique of resampling to estimate the performance of models. In this technique, the data is split into k-parts, and the model is trained using all parts except 1, which is kept aside as test data for evaluating the performance of the model. In this paper, we have chosen to repeat this process 10 times and the average value across all the built models is used as the robust prediction performance estimation. This process is stratified because it attempts to balance the number of samples belonging to each class in the k-splits.

In this study, the predictive performance of the machine learning classifiers is measured in terms of accuracy and f1-score based on the common evaluation metrics of machine learning. Training and testing accuracy measures are used for performance evaluation. Since we use a k-fold cross-validation score, the mean and standard deviation across the 10 models are calculated for the metrics training accuracy, testing accuracy, and f1-score. Training accuracy is the ratio of correct predictions on the training dataset while testing accuracy is the same calculated on the testing dataset. F1 Score is a function of precision and recall, where precision-recall and F1 score are defined as follows:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} = \frac{True\ Positive}{Total\ Predicted\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{True\ Positive}{Total\ Actual\ Positive}$$

$$\text{F1 Score} = \frac{Precision * Recall}{Precision + Recall}$$

where the true positives, true negatives, false positives, and true negatives are defined by the confusion matrix:

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

In our study negative indicates financially distressed companies and positive indicates financially healthy companies. A balance between precision and recall can be obtained if f1 score is used as a performance measure.

IV. Result Analysis

In this paper, we have performed the analysis of three types of deep neural networks namely MLP, LSTM, and CNN on two financial datasets. All these networks were trained and tested using samples from both the datasets. The results of phase1 i.e. accuracy and F1 score of deep learning models – MLP, LSTM and CNN on financial distress prediction using dataset1 and dataset2 are shown in Fig. 2, 3, 4 and 5, respectively. In phase 1 experiments, all three models were trained with different optimization technique, to evaluate the prediction performance. In the experiments, we have used combinations of different optimization functions with activation functions for optimizing each deep learning model. The above steps were repeated for dataset1 and dataset2. It was found that all the three deep learning models with sigmoid activation function and Adam optimizer outperformed any other combinations. The accuracy and f1 score of models with the best optimization techniques (Sigmoid + Adam optimizer) are shown in Fig. 2, 3, 4, and 5. The combinations of other optimization algorithms could not generate a robust predictive model (not shown).

In short, the phase 1 experiments concluded that the deep learning models designed with sigmoid activation function and Adam optimizer yielded the best predictive performance for financial distress prediction. Fig. 2 and 3 indicates the prediction performance of models for dataset1, while that of dataset2 is depicted in Fig. 4 and 5. The two graphical representations clearly show that Multi-layer Perceptron (MLP) has the highest performance in terms of accuracy, precision, and recall.

The predictive performance of dataset1 is higher compared to dataset2 because the former has more financial attributes compared to the latter, which helps to build a better classification model during training. However, the predictive performance of MLP outperforms the performance of LSTM and CNN models with dataset1 and dataset2. Hence, we can conclude that MLP is the best suited deep learning classifier for financial distress prediction. Accordingly, in the next phase of this study, we have selected Multi-Layer Perceptron model with sigmoid transfer function and Adam optimizer.

The results above represent the mean and standard deviation (in brackets) obtained using 10 times repeated random sub-sampling.

The results above represent the mean and standard deviation (in brackets) obtained using 10 times repeated random sub-sampling.

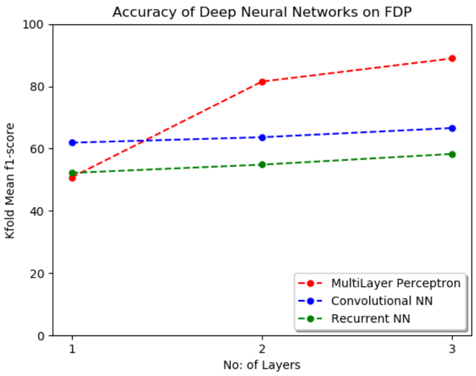


Fig. 2. Mean Accuracy (from 10 Repeated Random Sub-Sampling) of MLP, CNN and LSTM using DataSet1.

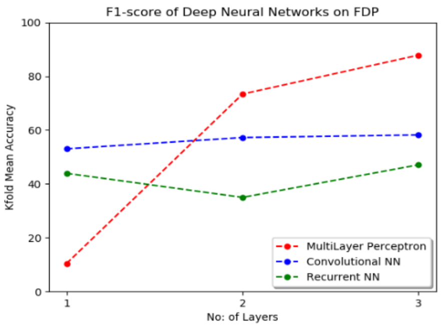


Fig. 3. Mean F1 Score (from 10 Repeated Random Sub-Sampling) of MLP, CNN and LSTM using DataSet1.

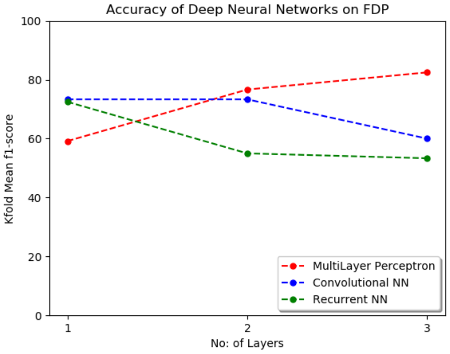


Fig. 4. Mean Accuracy (from 10 Repeated Random Sub-sampling) of MLP, CNN and LSTM using DataSet2.

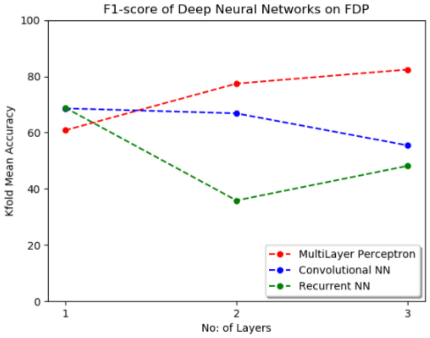


Fig. 5. Mean F1 Score (from 10 Repeated Random Sub-Sampling) of MLP, CNN and LSTM using DataSet2.

In phase 2, we further analyze the deep learning model – MLP for financial distress prediction. The results of phase 2 with different configurations of MLP are shown in Tables I and II. Based on the experiments from the preliminary study, the training method and the activation function were Adam Optimizer and sigmoid function respectively. The number of hidden layers and the number of neurons at each layer are varied in this phase, for further optimization of MLP architecture. The results of MLP on Financial distress prediction, with 16 different architectures developed, trained and tested are listed in Tables I and II. It can be noticed that any changes in the number of layers or the number of neurons in each layer affect the proficiency of the model. For example, as shown in Table I, MLP with configuration 10-10-10-10-10 had an acceptable accuracy value of 90.76% but the network with 50-50-50-50-50 configuration had a poor prediction accuracy of 70.91%. An optimized architecture for dataset1 is 10-20-10-20-10(93.79%) and that of dataset2 is 20-20-10-10-10(84.17%). The variation in the performance of models on two datasets is due to the change in the number of financial indicators in each dataset. A dataset with a higher number of attributes can be trained better and can generate a more robust model than a dataset with a smaller number of attributes.

A reduction in the accuracy (< 80% for dataset1 and < 70% for dataset2) was observed with networks containing more than 5 layers and hence were not able to generate a robust model (networks are not shown in the results table). Maximum prediction performance is obtained with a 4-layer architecture for experiments with dataset1 and dataset2. Also, the accuracy value started decreasing when the number of neurons at each level was approaching twice the number of attributes in the input dataset. A robust model was not generated after the number of nodes was set equal to and greater than 40(72.22%) and 30(68.33%) for dataset 1 and dataset 2 respectively. Hence this study indicates that higher prediction performance is obtained when the number of neurons at each level is less than twice the number of input attributes in the dataset. The prediction performance is maximum with an architecture containing a combination of 10 and 20 neuron units at hidden layers for both dataset1 and dataset2.

In the final phase of the study, we have compared our optimized MLP performance with the classic machine learning algorithms including support vector machine and decision tree. The prediction results in terms of accuracy are shown in Table III. The statistical results indicated that the prediction accuracy of the proposed optimized model was significantly higher than that of base machine learning models using both datasets.

TABLE. I. Predictive Accuracy of MLP using Dataset1

Structure	**Training Accuracy: Mean (Standard Deviation)**	**Testing Accuracy: Mean (Standard Deviation)**
5-5	72.44(21.97)	71.06(19.61)
5-5-5	95.57(2.81)	90.76(7.43)
10-10-10	87.87(6.25)	86.06(8.52)
5-5-5-5	95.89(2.75)	92.27(6.29)
10-10-10-10	97.16(3.22)	89.09(6.19)
20-20-20-20	96.49(3.22)	90.61(5.27)
20-10-20-10	96.83(2.08)	91.97(6.81)
10-20-10-20	97.15(2.38)	90.76(7.43)
5-5-5-5-5	96.20(1.91)	92.27(6.29)
10-10-10-10-10	97.46(2.60)	90.76(7.43)
20-20-20-20-20	96.50(2.55)	89.90(6.67)
30-30-30-30-30	98.10(1.09)	87.42(8.76)
10-20-10-20-10	97.77(2.04)	93.79(4.40)
20-20-10-10-10	98.72(0.90)	90.76(7.43)
50-50-50-50-50	72.77(20.43)	70.91(20.29)
100-50-100-50-100	55.33(3.12)	52.27(2.27)

TABLE. II. Predictive Accuracy of MLP using Dataset2

Structure	**Training Accuracy: Mean (Stand Deviation)**	**Testing Accuracy: Mean (Stand Deviation)**
5-5	73.33(15.82)	70.00(14. 43)
5-5-5	75.83(14.62)	73.33(13.74)
10-10-10	78.61(9.25)	76.67(10.23)
5-5-5-5	80.56(4.87)	75.83(9.82)
10-10-10-10	87.83(3.53)	82.50(3.82)
20-20-20-20	84.67(2.60)	80.83(9.75)
20-10-20-10	78.33(10.67)	78.33(5.00)
10-20-10-20	81.33(8.47)	78.83(9.43)
5-5-5-5-5	87.22(4.61)	81.67(3.73)
10-10-10-10-10	86.11(8.03)	82.50(8.62)
20-20-20-20-20	81.17(7.99)	75.17(9.75)
30-30-30-30-30	70.28(15.00)	68.33(14.04)
20-20-10-10-10	87.67(5.73)	84.17(5.34)
10-20-10-20-10	83.83(6.64)	81.67(2.89)
50-50-50-50-50	53.33(7.45)	52.50(5.59)
100-50-100-50-100	55.33(8.30)	53.33(5.53)

TABLE. III. Prediction Results (Accuracy) of MLP, SVM, and DT

	DataSet1	**DataSet2**
Deep Neural Network (MLP)	0.93	0.84
Support Vector Machine (SVM)	0.85	0.65
Decision Tree Classifier (DT)	0.80	0.67

V. Conclusion

In this paper, we have investigated the performance of deep learning neural networks namely MLP, LSTM, and CNN, on financial distress prediction. We have found that MLP networks are the best-performing distress prediction model. In the last phase of the paper, we have applied different architectural variations to MLP for further optimization. It was found that an accepted predictive performance rate can be achieved if the model is designed with 3 or 4 hidden layers with the neuron count at each level not exceeding twice the number of input attributes in the dataset. We have trained and tested the models using two different datasets with a varying number of input attributes and found that more the number of financial indicators, a better robust model can be generated. The simulation results also indicate that the proposed model has higher performance when compared to classic machine learning models like support vector machine and decision tree.

References

[1] Adrian Gepp, Kuldeep Kumar," Predicting Financial Distress: A Comparison of Survival Analysis and Decision Tree Techniques", Procedia Computer Science, Volume 54, 2015.

[2] Aktham I. Maghyereha, Basel Awartani, Bank distress prediction: Empirical evidence from the Gulf Cooperation Council countries. United Arab Emirates University, the United Arab Emirates and Plymouth University, U, 2014.

[3] Altman, E. I. , Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589–609, 1968.

[4] Chang, C. L., & Chen, C. H. (2009). Applying the decision tree and neural network to increase the quality of dermatologic diagnosis. Expert Systems with Applications, 36, 4035–4041

[5] Chia-Pang Chan, Ching-Hsue Cheng, An Attribute Selection Based Classifier to Predict Financial Distress, 2012.

[6] D. Wu, L. Liang, Z. Yang, Analyzing the financial distress of Chinese public companies using probabilistic neural networks and multivariate discriminate analysis, Socio-Econ. Plan. Sci. 42 (3) 206–220, 2008.

[7] Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. San Francisco, CA, USA: Morgan Kaufmann.

[8] Huang, M. J., Chen, M. Y., & Lee, S. C. Integrating data mining with case-based reasoning for chronic disease prognosis and diagnosis. Expert Systems with Applications, 32(3), 856–867, 2007.

[9] I. Guyon, and A. Elisseeff, "An introduction to variable and feature selection," The Journal of Machine Learning Research, vol. 3, pp. 11571182, 2003.

[10] Jae Kwon Bae, Predicting financial distress of the South Korean manufacturing industries, Expert Systems with Applications, Volume 39, Issue 10, 2012.

[11] Jie Sun, Hamido Fujita, Peng Chen, Hui Li. "Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble", Knowledge-Based Systems, 2017

[12] Jie Sun, Hui Li, Qing-Hua Huang, Kai-Yu He, Predicting financial distress and corporate failure: A review from the state-of-art definitions, modeling, sampling and featuring approaches, Knowledge-based systems, 57:41-56, 2014.

[13] J.H. Min, Y.-C. Lee, Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters, Expert Syst. Appl. 28 (2005) 128–13.

[14] J. Ohlson, Financial ratio and the probabilistic prediction of bankruptcy, J. Account. Res. 18, 109–131, 1980.

[15] L. Cleofas-Sánchez, V. García, A.I. Marqués, J.S. Sánchez, Financial distress prediction using the hybrid associative memory with translation, Applied Soft Computing, Volume 44, 2016.

[16] Mohammad Mahdi Mousavi, Jamal Ouenniche, Kaoru Tone, "A comparative analysis of two-stage distress prediction models", Expert Systems with Applications, Volume 119, 2019.

[17] M. A. Hall, and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," Knowledge and Data Engineering, IEEE Transactions on, vol. 15, no. 6, pp. 1437-1447, 2003.

[18] Rastin Matin, Casper Hansen, Christian Hansen, Pia Mlgaard,"Predicting distresses using deep learning of text segments in annual reports", Expert Systems with Applications, Volume 132, 2019.

[19] P. Ravisankar, V. Ravi a, I. Bose, Failure prediction of dotcom companies using neural network-genetic programming hybrids, Expert systems with applications, 36(3):4830-4837, 2009.

[20] P. Ravisankar, V. Ravi a, Financial distress prediction in banks using Group Method of Data Handling neural network, counter propagation neural network and fuzzy ARTMAP, Knowledge-based systems, 23(8):823-831, 2011.

[21] Ruibin Geng, Indranil Bose, Xi Chen, Prediction of financial distress: An empirical study of listed Chinese companies using data mining, 2015.

[22] Ahmed M. Khedr, and Bhatnagar R., New Algorithm for Clustering Distributed Data using k-means, Computing and Informatics, Vol. 33, pp. 1001-1022, 2014.

[23] Ahmed M. Khedr, Decomposable Naive Bayes Classifier for Partitioned Data, Computing and Informatics, Vol. 31, pp. 1511-1531, 2012.

[24] Ahmed M. Khedr, Nearest Neighbor Clustering over Partitioned Data, Computing and Informatics, Vol. 30, pp. 1001-1026, 2011.

[25] Ahmed M. Khedr and Salim A., Decomposable Algorithms for Finding the Nearest Pair, J. Parallel Distrib. Comput., Vol. 68, pp. 902-912, 2008.

[26] Ahmed M. Khedr, Learning k-Classifier from Distributed Databases, Computing and Informatics Journal, Vol. 27, pp. 355-376, 2008.

[27] Ahmed M. Khedr and Bhatnagar, R., Agents for Integrating Distributed Data for Complex Computations, Computing and Informatics Journal, vol. 26, No.2, pp. 149-170, 2007.

[28] Ahmed M. Khedr and Mahmoud, Rania. Agents for integrating distributed data for function computations. Computing and Informatics. 31. 1101-1125, 2012.

[29] Ahmed 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. 334-353, 2016.

[30] Ahmed M. Khedr, Walid Osamy, Ahmed Salim and Abdel-Aziz Salem, Privacy Preserving Data Mining Approach for IoT based WSN in Smart City, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 8, 2019.

Investigation of a 7-Level Inverter-based Electric Spring Subjected to Distribution Network Dynamics

K.K.Deepika[1]
Dept. of EEE, VIIT, Visakhapatnam, Andhra Pradesh
Scholar, KLEF, Vijayawada,A.P., India

Dr. G.Kesava Rao[2]
Professor, Department of EEE
KL EF, Vijayawada, Andhra Pradesh

Dr. J. Vijaya Kumar[3]
Professor, Dept. of EEE, ANITS
Visakhapatnam, Andhra Pradesh, INDIA

Dr. Satya Ravi Sankar Rai[4]
Dept. of EEE, Vignan's Institute of Information Technology
Visakhapatnam, Andhra Pradesh, India

Abstract—**This paper aims to provide solution to mitigate the voltage variations in critical load caused by the high penetration of DGs into distribution system using Electric Springs (ES). In this regard, there is a need for its exploration with various converter circuits. The improvised topology opens new avenues in the renewable energy powered micro grids for the implementation of ES with a Multi-Level Inverter (MLI) comprising a voltage balancing circuit providing a better quality of power system stability and voltage regulation. This paper captures the voltage dynamics of distribution system dominated by Renewable variability for varying reactive power of the DGs and constantly changing consumer demands. These are analyzed and explained using voltage profiles and power flows in Matlab/Simulink environment. It is practically shown that with the developed ES topology %THD in the system is conspicuously reduced and voltage regulation is seamlessly improved.**

Keywords—Electric spring; critical load; multilevel inverter; voltage balancing circuit; voltage regulation

I. Introduction

Increasing Distributed Generation in Smart Grid systems impacts the reactive power flow in the feeder and thereby leads to voltage variations in the distribution system [1]. Also, constantly fluctuating consumer load demands effect the load of the distribution system. It is observed that the distribution system is prone to voltage collapse under critical loading conditions. There are several strategies to maintain power quality in the distribution network [2] strengthened through proper placement of the devices [3]. Another way of providing the solution is Demand Side Management (DSM) wherein power demand follows the supply [4]. Creation of new technologies in research, stimulate to achieve the Sustainable Development Goals established by the United Nations (2015) [5]. Among the various profound DSM technologies, Electric spring has emerged to provide voltage and frequency regulation [6][7]. When embedded within a less voltage sensitive load, ES forms a smart load and enables the demand to follow the Renewable variability. The improved regulation is practically executed with power electronics.

In the emerging paradigm of power electronics, balancing the number of power switches, harmonic distortion with the preferred multilevel inverters and investigation with diverse PWM schemes [8] is important to meet the requirements of the microgrids and nanogrids [9]. Existing inverter topologies for ES focused on the %THD, number of switches and PWM techniques [10], but there is not a method that would attend to the concern of frequent non-critical load changes, in addition to the former phenomena. This paper explores the relation between dynamic loads and various modes of operation of MLI based ES. Switching operation of the loads influences the PCC voltage. This influence is efficiently handled and the same is evidently illustrated in the 4 case studies.

II. Working of MLI based Elcetric Spring

Electric Spring is a new voltage compensating device employed in smart grids using demand side management. As shown in Fig. 1, this custom power device is connected in series with a less voltage sensitive load like refrigerator, music system, etc. that have inbuilt Switched Mode Power Supply circuit to withstand voltage variations. This comprises a smart load. Smart load is connected across voltage sensitive load, termed as critical load, to maintain voltage constant. ES senses the voltage fluctuations line voltage, V_s with respect to its reference voltage, V_{s_ref} and operates analogous to mechanical spring [11], as outlined in Table I. The magnitude and phase of the voltage injected by the ES is controlled by the ES controller. Existing inverter topologies for ES [12]-[14] focused on the %THD, number of switches and PWM techniques. A major benefit of realizing 7-level output voltage with the topology under consideration [15] as illustrated in Fig. 2 is reduction in number of power switches and reduced switching losses. The proposed ES configuration consists of 8 MOSFETS, 4 diodes and 3 capacitors for input voltage division.

TABLE. I. Analogy of Mechanical and Electric Spring

Mechanical spring		**Electric spring**	
State of spring	*Mode of Operation*	*State of voltage*	*Mode of Operation*
Neutral	Neutral position	Vs = Vs_ref	Neutral
Compressed	Mechanical push (upward force)	Vs < Vs_ref	Capacitive Mode (voltage boosting)
Extended	Mechanical pull (downward force)	Vs > Vs_ref	Inductive Mode (voltage reduction)

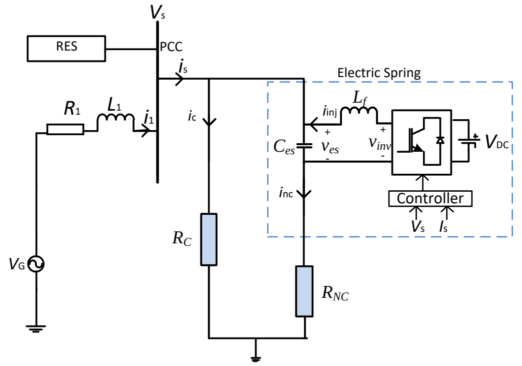


Fig. 1. Electric Spring in a Distribution System.

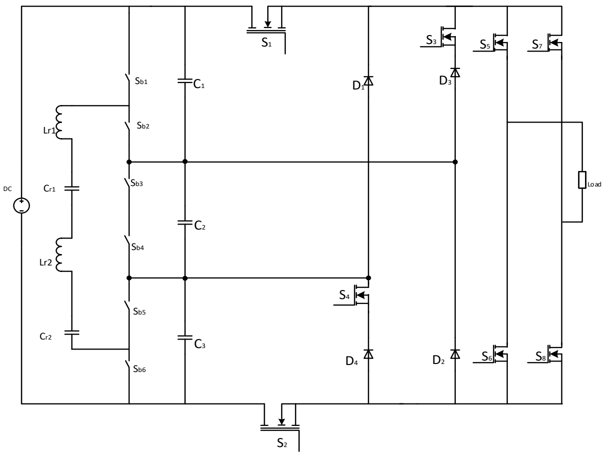


Fig. 2. A 7-Level Inverter Topology for Implementation to ES.

A. *For Voltage Level* $\frac{Vdc}{3}$

For output voltage $V_{dc}/3$, S_{b1} and S_{b3} are in ON position and the current flows through capacitor $C_{1,}$ thereby capacitor gets charged during the positive half cycle. Switches $S_{1,}$ $S_{5,}$ S_8 are turned ON and capacitor C_1 gets discharged. Voltage available at the load terminals is $V_{dc}/3$.

B. *For Voltage Level* $2\frac{Vdc}{3}$

For output voltage $2V_{dc}/3$, S_{b1} and S_{b5} are in ON position and the current flows through capacitor C_1 and $C_{2,}$ thereby capacitors C_1 and C_2 get charged during the positive half cycle. Switches $S_{1,}$ $S_{5,}$ S_4 are turned On and capacitors discharge causing voltage available at the load terminals as $2V_{dc}/3$.

C. *For Voltage Level* V_{DC}

For output voltage $V_{dc,}$ S_{b1} and S_{b6} are in ON position and capacitors $C_{1\,,}$ $C_{2,}$ C_3 are charged during the positive half cycle. Switches $S_{1,}$ $S_{5,}$ S_8 S_2 are turned ON and capacitors discharge causing voltage available at the load terminals as $V_{dc.}$

D. *For Voltage Level* - $\frac{Vdc}{3}$

For output voltage $-V_{dc}/3$, switches S_{b1} and S_{b6} of the voltage balancing circuit are in conducting mode and C_3 is charged. The capacitor gets discharged through D_1 and S_7, S_2, $S_{6.}$ The current direction through the load reverses and hence the voltage at the load terminals is$-\frac{Vdc}{3}$.

TABLE. II. SWITCHING COMBINATIONS OF MLI

Output voltage	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
$+V_{DC}/3$	1	0	0	0	1	0	0	1
$+2V_{DC}/3$	1	0	0	1	1	0	0	1
$+V_{DC}$	1	1	0	0	1	0	0	1
0	0	0	0	0	1	0	1	0
$-V_{DC}/3$	0	1	0	0	0	1	1	0
$-2V_{DC}/3$	0	1	1	0	0	1	1	0
$-V_{DC}$	1	1	0	0	0	1	1	0

E. *For Voltage Level* -$2\frac{Vdc}{3}$

For output voltage $-2V_{dc}/3$, during the positive half cycle, switches S_{b2} and S_{b6} of the voltage balancing circuit and the capacitors C_2, C_3 gets discharged. They discharge through D_3, S_3 and S_7 in the forward direction and through S_6 and S_2 in return direction.

F. *For Voltage Level –Vdc*

For output voltage $-V_{dc,}$ switches S_{b1} and S_{b6} are in ON position and capacitors $C_{1,}$ C_2, C_3 are charged. Voltage $-V_{dc}$ will now be available across the H-bridge terminals. The capacitors will discharge through S_1 and S_7 in forward path S_2 and S_6 in return path.

G. *For Voltage Level Zero*

For zero output voltage, switches S_{b2}, S_{b3}, S_{b4}, S_{b5} are short circuit and RSCC-1 and RSCC-2 are in working mode. There is no current flow through any of the capacitor hence charging or discharging. Due to the gate pulses given to switches S_5 and S_7 of MLI are in ON position and the current flows through the load. Switching combinations of the switches of MLI is outlined in Table II, to attain various output voltages.

III. RESULTS AND DISCUSSION

Dynamic behavior of the Multi-level inverter based Electric spring is subjected to an important variation of load. Performance of the Multi-level inverter based ES is commanded by the Sine-triangle Pulse Width Modulation (SPWM) to improve the quality of the voltage injected at PCC by ES, in both capacitive and inductive modes of ES.

To substantiate the dynamic voltage regulation capability of the improvised MLI based ES in all its modes, a model as shown in Fig. 3 is simulated in MATLAB/ Simulink environment. Basic parameters of distribution network and Electric Spring are mentioned in Table III. Voltage on LV side is considered as 220 V.

Random variation of 20% in output of the Distributed generation source is represented by a three phase programmable voltage source. Overall analysis of the new improvised ES is carried out in four ways as detailed in Table IV.

A. *Capacitive mode of ES with Proposed Converter Topology*

To illustrate the voltage boosting function of ES, reduction in the line voltage is simulated from 0.4 sec to 0.8 sec, as shown in Fig. 4. At 0.4 sec, ES initiates to operate in voltage support mode and injects a voltage of 40volts in series with the Non-critical load voltage, to restore the line voltage back to its nominal value of 220 Volts. In Fig. 4, waveform of critical load voltage, clearly depicts that ES succeeds to operate in voltage support mode and the load demand of non-critical load to follow the generation. This is accomplished by absorption of real power and supply of reactive power into the system by ES, as shown in Fig. 5.

TABLE. III. SIMULATION SYSTEM SPECIFICATIONS

Specifications of line and load		**Specifications of Electric Spring**		
Line Inductance (mH/km)	1.22	Low-pass filter	Inductance (mH)	5
Line Resistance (*Ω*/km)	0.1		Capacitance (μF)	13.2
Noncritical Load (*Ω*)	50.5	PI Controller	K_p	2
Critical load Z (*Ω*)	53		K_i	1.5

TABLE. IV. SYNTHESIS OF THE BEHAVIOR OF MLI BASED ES FOR VARIOUS CONDITIONS

Case Number	**Imposed condition**	**Load variation**	**Line voltage variation**	**%THD**
1.	Voltage sag	-	-20%	0.42
2.	Voltage swell	-	+20%	0.42
3.	Load switching during voltage sag	-66.67%	-20%	0.44
4.	Load switching during voltage swell	-66.67%	+20%	0.44

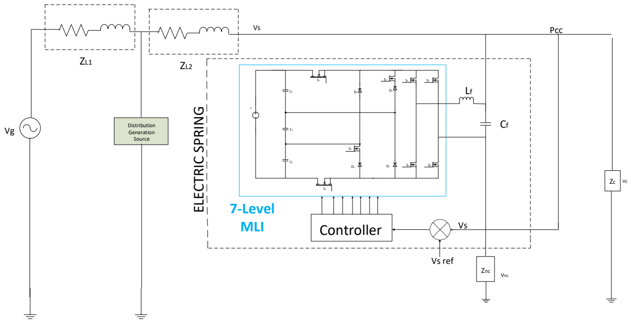


Fig. 3. ES Configuration Comprising a Converter with 8 Switches.

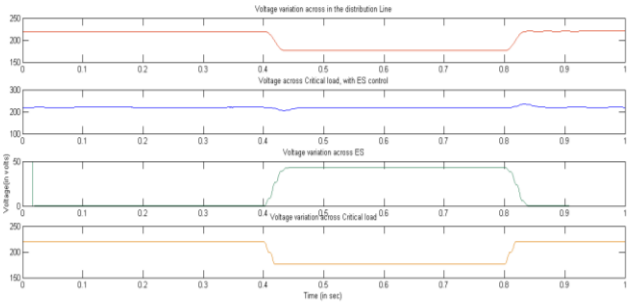


Fig. 4. Simulation Waveforms of Voltages with ES Operation in Capacitive Mode.

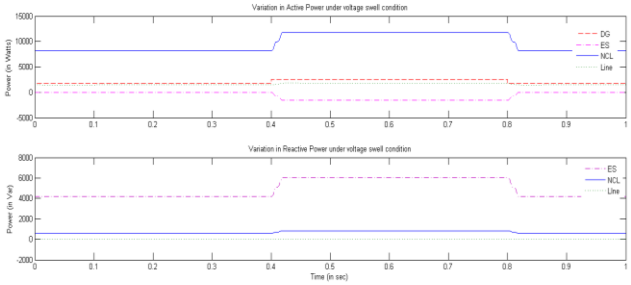


Fig. 5. Simulation Waveforms of Active and Reactive Powers of ES Operation in Capacitive Mode.

B. Inductive mode of ES with Proposed Converter Topology

To test the Inductive mode of ES, generated power by the DG source is made higher than load demand. This leads to an overvoltage condition simulated from 0.4-0.8 sec and consequent voltage profiles are shown in Fig. 6.

In this condition, ES operates in voltage reduction mode and suppresses the overvoltage. Voltage across ES decreases by 40 volts to restore the line voltage back to its nominal value of 220 Volts. ES maintains the voltage across critical load constant in voltage suppression mode. This is accomplished by injecting real power and absorbing reactive power into system as shown in Fig. 7.

C. Capacitive mode of ES with Proposed Converter Topology and Dynamic Loading Condition

To further validate the capability of ES in dynamic loading during capacitive mode of operation, voltage is reduction by 20% is simulated from 0.4 to 0.8 sec and non-critical load value is reduced at 0.6 sec. It is observed in Fig. 8 that the line voltage rises to 265 volts and is seamlessly regulated to 220 volts by ES in about 0.08 seconds. ES effectively boosts the line voltage from 0.4 sec and its regulatory performance is further increased with the dynamic loading at 0.6 sec, in addition to the voltage dip in line voltage. Fig. 8 shows that ES regulates the critical load voltage to 220V controlling the non-critical load voltage.

D. Inductive mode of ES with Proposed Converter Topology under Dynamic Loading Condition

To further validate the capability of ES in dynamic loading during inductive mode of operation, voltage rise by 20% is simulated from 0.4 to 0.8 sec and non-critical load value is reduced at 0.6 sec. It is observed in Fig. 9 that the line voltage rises to 265 volts and is seamlessly regulated to 220 volts by ES in about 0.08 seconds. ES effectively suppresses the line voltage from 0.4 sec and its regulatory performance is further increased with the dynamic loading at 0.6 sec, in addition to the overvoltage. Fig. 9 shows that ES regulates the critical load voltage to 220V controlling the non-critical load voltage.

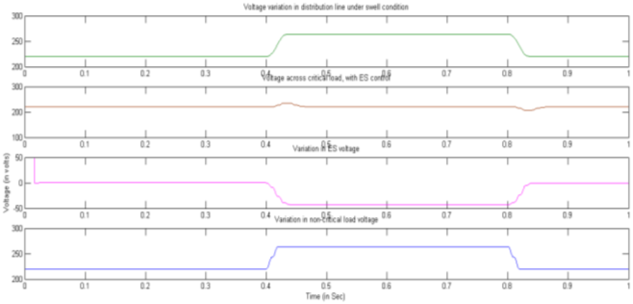


Fig. 6. Simulation Waveforms of Voltages with ES Operation in Inductive Mode

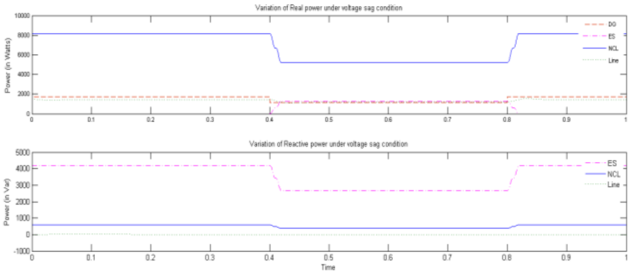


Fig. 7. Simulation Waveforms of Active and Reactive Powers with ES Operation in Inductive Mode.

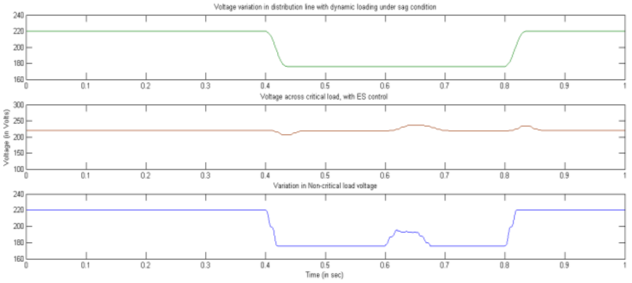


Fig. 8. Variations in Voltages for ES Operation in Capacitive mode with Dynamic Loading Condition.

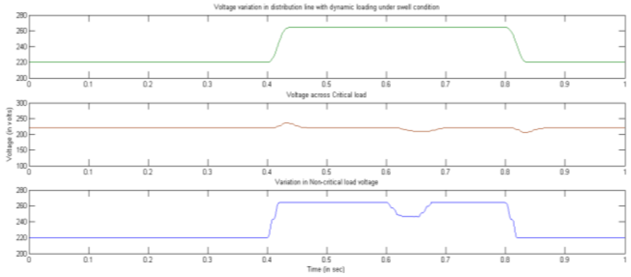


Fig. 9. Variations in Voltages for ES Operation in Inductive mode under Dynamic Loading Condition.

Gate signals given as input to the switches of MLI is illustrated in Fig. 10. These are for the operation of the switches during voltage sag, simulated in case 1. And the THD in the voltage injected by ES at PCC is 0.42% as shown in Fig. 11, which is within the standard limits. The fundamental component measures 305 Hz.

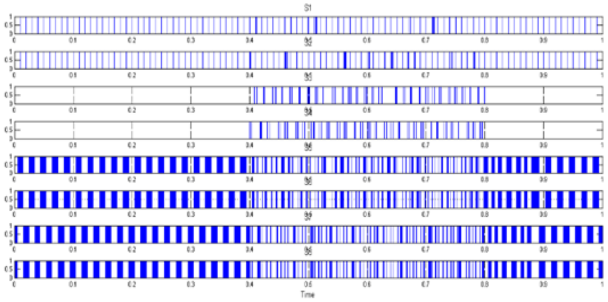


Fig. 10. Gate Signals Fed to the MLI in Capacitive mode of ES.

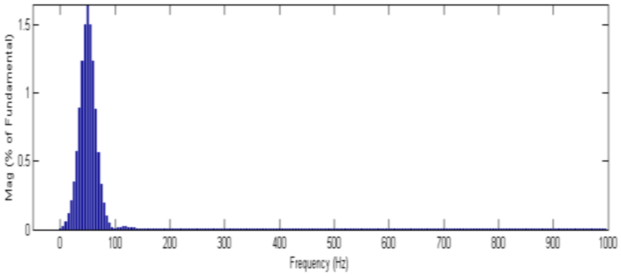


Fig. 11. %THD in the main Voltage.

IV. Conclusion

This paper presents a new MLI based ES to regulate main voltage in smart grids. The implemented ES controls the injected voltage using a 7 level inverter that has benefits of reduction in the number of switches, switching losses and also %THD in the local mains voltage. Efficacy of the ES is validated for inductive and capacitive mode of operation. Objectives are discussed with the simulation studies and the results corroborate the effectiveness of improvised topology for ES in its individual modes as well as with dynamic loading conditions.

References

[1] Nazir, Refdinal; NURDIN, Muhammad; FITRIANTO, Eka. Voltage Profile Improvement of the 20 kV Painan Distribution System with Multiple Distributed Renewable Energy Generation. International Journal of Technology, [S.l.], v. 7, n. 1, pp. 26-37, jan. 2016. ISSN 2087-2100. doi:10.14716/ijtech.v7i1.2193.

[2] S. Rahman, "An efficient load model for analyzing demand side management impacts," IEEE Transactions on Power Systems, vol. 8, no. 3, pp. 1219-1226, 1993.

[3] Arief, Ardiaty & Nappu, Muhammad Bachtiar & Antamil, Antamil. (2018). Analytical Method for Reactive Power Compensators Allocation. International Journal of Technology. 9. 602. 10.14716/ijtech.v9i3.913.

[4] Chen, Xia & Hou, Yunhe & Tan, Siew-Chong & Lee, C.K. & Hui, S.Y.. (2014). Mitigating Voltage and Frequency Fluctuation in Microgrids Using Electric Springs. IEEE Transactions on Smart Grid. 6. 10.1109/TSG.2014.2374231.

[5] Berawi, Mohammed Ali. The Role of Technology in Achieving Sustainable Development Goals. International Journal of Technology, [S.l.], v. 8, n. 3, pp. 362-365, Apr. 2017. ISSN 2087-2100. doi:10.14716/ijtech.v8i3.9296.

[6] Chen, Xia & Hou, Yunhe & Tan, Siew-Chong & Lee, C.K. & Hui, S.Y.. (2014). Mitigating Voltage and Frequency Fluctuation in Microgrids Using Electric Springs. IEEE Transactions on Smart Grid. 6. 10.1109/TSG.2014.2374231.

[7] Deepika K, Vijayakumar J, G K Rao, Chaitanya S, "Adaptive PI control of Electric Springs for Voltage Regulation under Dynamic Load Changes", International Journal of Innovative Technology and Exploring Engineering, 2019 vol: 8 (10) pp: 1051-1056.

[8] Satiawan, I Nyoman Wahyu et al. Performance Comparison of PWM Schemes of Dual-inverter FED Five-phase Motor Drives. International Journal of Technology, [S.l.], v. 5, n. 3, pp. 277-286, Nov. 2014. ISSN 2087-2100. doi:10.14716/ijtech.v5i3.609.

[9] Andreas, Jamsep & Setiawan, Eko & Halim, Suharsono & Atar, Muhammad & Nur Shabrina, Hanifati. (2018). Performance Test of 2.5 kW DC Boost Converter for Nanogrid System Applications. International Journal of Technology. 9. 1285. 10.14716/ijtech.v9i6.2429.

[10] Wang, Q.; Deng, F.; Cheng, M.; Buja, G. The State of the Art of Topologies for Electric Springs. Energies 2018, 11, 1724.

[11] S.Y.R.Hui, C.K.Lee, Fu, “Electric springs–A new smart grid technology,” IEEE Trans. Smart Grid, vol. 3, no. 3, pp. 1552–1561, September. 2012.

[12] Wamne, S. & Balpande, P. & Murme, S. & Dhakate, Parag & Bajpai, S. & Gawande, Snehal & Nagpure, R. & Waghmare, Manoj. (2018). A Novel Common Inverter Electric Spring Configuration in Smart Grid. 1-6. 10.1109/PEDES.2018.8707844.

[13] Rutuja Pawar, S. P. Gawande, S. G. Kadwane, M. A. Waghmare, R. N. Nagpure “Five-Level Diode Clamped Multilevel Inverter (DCMLI) Based Electric Spring for Smart Grid Applications”, Energy Procedia 117 (2017) 862–869.

[14] K. Gajbhiye, P. Dahiwale, S. Bharti, R. Pawar, S. P. Gawande and S. G. Kadwane, "Five-level NPC/H-bridge MLI based electric spring for harmonic reduction and voltage regulation," 2017 International Conference on Smart grids, Power and Advanced Control Engineering (ICSPACE), Bangalore, 2017, pp. 203-208.

[15] C. Hsieh, T. Liang, S. Chen and S. Tsai, "Design and Implementation of a Novel Multilevel DC–AC Inverter," in IEEE Transactions on Industry Applications, vol. 52, no. 3, pp. 2436-2443, May-June 2016. doi: 10.1109/TIA.2016.2527622.

On Exhaustive Evaluation of Eager Machine Learning Algorithms for Classification of Hindi Verses

Prafulla B. Bafna[1], Jatinderkumar R. Saini[2]
Symbiosis Institute of Computer Studies and Research
Symbiosis international (Deemed) University, Pune, India

***Abstract*—Implementing supervised machine learning on the Hindi corpus for classification and prediction of verses is an untouched and useful area. Classifying and predictions benefits many applications like organizing a large corpus, information retrieval and so on. The metalinguistic facility provided by websites makes Hindi as a major language in the digital domain of information technology today. Text classification algorithms along with Natural Language Processing (NLP) facilitates fast, cost-effective, and scalable solution. Performance evaluation of these predictors is a challenging task. To reduce manual efforts and time spent for reading the document, classification of text data is important. In this paper, 697 Hindi poems are classified based on four topics using four eager machine-learning algorithms. In the absence of any other technique, which achieves prediction on Hindi corpus, misclassification error is used and compared to prove the betterment of the technique. Support vector machine performs best amongst all.**

Keywords—Classification; eager machine learning algorithm; Hindi; prediction

I. Introduction

Most of the past and contemporary research works have targeted English corpus document classification and prediction. In online and offline systems, documents are continuously generated, stored, and accessed every day in large volumes. Classifying text according to the contents present helps to produce groups based on tokens present in the text. The maximum work is done in text classifiers focuses on English corpus, but text in Hindi on the web has come of age since the advent of Unicode standards in Indic languages. The Hindi content has been growing by leaps and bounds and is now easily accessible on the web at large. Generally, researchers have focused on Hindi text but only for Natural Language Processing (NLP) activities like word identification, stemming and summarization [1].

Classification or supervised learning groups the labeled data based on the features of data. The data is partitioned as training and testing. Classifiers are broadly divided into eager and slow learners. Eager learners require a long period for training and less time for predicting. For slow learners data gets trained early but it takes more time for a prediction. Eager classifiers give better results than lazy classifiers for text data, so these classifiers are chosen. Naive bayes, Support Vector Machines, Neural Network and Decision tree are popularly used eager classifiers. A decision tree is a classifier, which generates several rules and tables. As a result, rules are placed in the form of decision trees.

Artificial neural network (ANN) has minimum three layers , input, hidden and output. Depending upon the input given and its respective output the network consisting of nodes gets trained. All nodes and layers are interconnected with each other and pass the values generated through the functions, it means that every node present in layer n is connected to various nodes present in tier n-1, inputs connected to respective nodes and nodes present in layer n+1. Output nodes show the classes to which a particular input object belongs.

Classification and regression is carried out through “Support Vector Machine” known as a supervised machine learning algorithm. Each data item is plotted as appoint in n dimensional space. It considers features of the object which are represented by coordinates of a point. SVM differentiates points in different hyperplanes.

Naive Bayes works with text classification. Every unique term is treated as a feature while processing text. Naive Bayes is an eager learner and simple algorithm and termed as strong performer to achieve the classification of text. Naïve byes works best when features are dependent on each other [2].

To apply any classification algorithm on text data first it needs to be converted into structured form. There are several techniques like bag of words, term frequency inverse document frequency and so on, which selects important terms from the corpus based on the frequency of the terms [3].

II. Bakground

In spite of Hindi being used for communication by a large number of people in the world, lots of research work in the field of text classification [4-6] focuses on English. The reason may be processing Hindi corpus is a difficult task.

Topic models are built on Hindi corpus using algorithms, namely Latent Semantic Indexing (LSI), Non-negative Matrix Factorization (NMF), and Latent Dirichlet Allocation (LDA). Many visualizations in the form of trees were used to focus the analysis and results. The outcomes of Hindi text topic modelling gives best results as compared to some outcomes generated on English corpus [7]. To apply any classification techniques the data should be in the tabular form. Various techniques are available to store such types of data for example bag of words [8]. But it creates dimension curse, as all terms in the corpus are considered. High dimensions affect the performance of the algorithm. To reduce high dimensions, only significant words need to be considered. Classification will execute in less time if the top significant words are selected.

To improve the classification process, the text is preprocessed by removing stop words, etc. [9-10]. Generally (TF-IDF) is a popularly used technique that transforms text data into matrix form. The measure represents the significance of the token with respect to text documents considering the entire corpus. In document processing, it acts as a weighting unit. In spite of increasing word count proportional to the number of documents in which it is present, The TF-IDF ignores the most commonly occurring words, by offsetting count of the words in the entire corpus. [11]. Accuracy, and misclassification errors are used to evaluate classifiers. Hindi is a morphologically rich language. Hindi words have many morphological variants that present the same concept but differ in tense, plurality, etc. A lightweight stemmer is proposed for Hindi, which conflates terms by providing suffix list. The stemmer has been evaluated by computing under stemming and over stemming figures for a corpus of documents [12-14].

Various methods like simulated annealing, genetic algorithms and differential evolution are used which finds out the required solution. Multi-parent mutation and crossover operations are used by the differential evolution algorithm. Results of the methods are input to Naïve Bayes classifiers and its different variations. [15-18]. The proposed algorithm works well in case of text classification as compared with other existing algorithms.

ANN is used to classify the text present in the Arabic language. ANN model is generated for an Arabic corpus. Document representation using different methods along with the feature weights [26] are used and results into identifying important terms. Each Arabic document is represented by the term weighting scheme. The term weighting scheme is used to represent the document. To choose the most significant terms, SVD is used to avoid dimension curse.

Back-propagation neural network (BPNN) and modified back-propagation neural network (MBPNN) are proposed to categorize the text. To avoid dimension curse and to improve the efficiency of algorithm an efficient feature selection method is used.

Training time required for BPNN is slow thus it is modified to enhance the speed required to train. Instead of using a vector space model which is based on term frequency, latent semantic analysis is used. LSA uses only important terms and considered a semantic relationship between the terms and builds concept space. The news dataset is used to prove the efficacy of prosed technique [27].

Different Machine learning algorithms are used to classify the text present in different questions. Two approaches namely Bag-of-words and bag-of-grams are used to construct vector space. Syntactic terms present in the question are identified using a kernel function. Comparative analysis of algorithm performance is being carried out [28] Classification of Hindi text documents includes dividing the documents as training and testing corpus and applying classifiers on the labeled text. Handwritten and printed text documents are partitioned into specific classes. The algorithm is implemented on Hindi text which has Hindi printed and handwritten. The system will be useful for discrimination between handwritten and printed text [19-21].

The text is classified based on emotional features present into it. There are nine categories of emotional features. One category represents one class. Term frequency is used to handle overlapping features. Naïve byes and support vector machines are executed on a set of 55 poems having 10531 words [22-25].

This research is unique because

1) Prediction of Hindi poem using four eager classifiers is achieved.

2) Performance evaluation of the classifier is carried out.

3) Scalability is achieved by processing 697 poems.

III. Research Methodology

The proposed approach initiates with corpus removal of stop words and finds out top N frequent terms using TF-IDF weights on the corpus of poems having three groups. The N value is called a threshold, which is 50 % of maximum TF-IDF weight. Stemming and lemmatization are not used. It effectively removes all unuseful words. Different classifiers are available in the literature, the proposed approach applies all eager classifiers on the term document matrix and the model is built using each classifier. Naïve byes and random forest algorithms are applied. Their performance is evaluated using accuracy. Support vector machine performs best in comparison with remaining algorithms. Fig. 1 depicts the research methodology.

In the paper terms, dimensions, words and tokens are used as synonyms, interchangeably. The paper is organized as follows. The work done by other researchers on the topic is presented as a background in the next section. The third section presents the methodology; the fourth section depicts Results and discussions. The paper ends with a conclusion and future directions. Table I shows steps in the proposed approach.

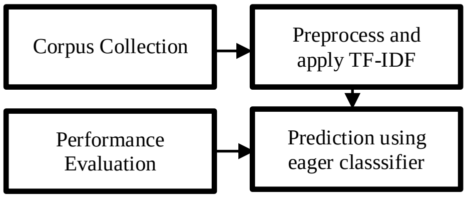

Corpus Collection
Preprocess and apply TF-IDF
Performance Evaluation
Prediction using eager classsifier

Fig. 1. Diagrammatic Representation of Research Methodology.

TABLE. I. Steps and Packages used in the Proposed Approach

Step No	Step	Library/Package/Function
1	Documents are pre-processed and stop words are removed.	library(udpipe)
2	Apply TF-IDF to calculate token weights	dtm_tfidf
3	Select terms having token weights greater than 50 % threshold	dtm <- document_term_matrix(dtm_threshold)
4	Apply and evaluate classifiers	model=naive_bayes(as.factor(type) ~., data=train),
5	Select the best classifier and Predict category of new poem	p1=predict(Naïve_byes,train)

1) Corpus collection and preparation: The proposed approach initiates with data collection and preparation. It includes the process of generating, loading and preprocessing of the corpus. Corpus containing Hindi Text is preprocessed to remove the stop word. It is then partitioned into training and validation sets. The corpus comprises of poems belonging to three categories. The classes or categories are "बाल गीत" ("Bal geet") means children's' poems, "उपदेश गीत" ("Updesh geet") means life lesson teaching poem and "भजन" ("bhajans") means devotional songs. "देश भक्ति" (Desh Bhakti) means patriotic songs. The size of the corpus is 697 and it is downloaded from different websites [29].

2) Converting unstructured data into structured data: Converting unstructured data into the structured one is the next corpus of poems is converted into a vector space model. TF-IDF is used on a set of documents, and token weight is calculated. Terms or tokens having a weight greater than or equal to the threshold are considered. The Document term matrix (DTM) is input to the classifier algorithm. This step selects important tokens present in the corpus and selected significant tokens are further used to form a vector space model.

3) Model training using different classifiers and evaluation: The labeled dataset or corpus is trained based on different values of input and its corresponding output. Eager Classifiers are applied on the DTM. Models are generated and trained using the training corpus. A confusion matrix is found out for all four algorithms and misclassification error was used to evaluate the performance of the algorithm. The best classifier is selected to predict the category of the new poem. Figure specifies the diagrammatic representation of research methodology

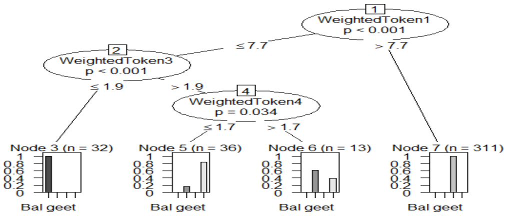
4) Prediction: The best performing classifier is used to predict the category of a poem. It was observed that the support vector machine predicts the class of a poem in a more accurate way.

IV. Results and Discussions

Fig. 2 shows a decision tree for Hindi poems' corpus along with token weights. The corpus of 697 poems is used to build the model. Each token's significance with respect to each category is generated by a decision tree. The figure depicts a particular node represented as "Bal geet" category. The rules based on the weighted tokens for each category are generated.

Fig. 3 shows Naïve bayes classification. The model is a plot for weighted token 4 on the Y axis, it represents a density of Weighted token4 for different categories of poems. The graph clearly represents four different categories of poems namely Bal geet, Bhajan Updesh geet and DeshBhakti geet. and Updesh geet are classified as Bhajans.

Fig. 4 shows the SVM plot. SVM divides the data into two significant hyperplanes. It clearly shows that the upper part of hyperplane consists of poems having category "Desh Bhakti geet". Rest of the poems are distributed in the lower part. Overlapping of poems for two classes can be observed. Confusion matrix depicts the misclassification between Updesh geet and Bhajans.



```
Fitted party:

[1] root

[2] WeightedToken1 <= 7.7

| | [3] WeightedToken3 <= 1.9: Bal geet (n = 32, err = 0.0%)

| | [4] WeightedToken3 > 1.9

| | | [5] WeightedToken4 <= 1.7: Updesh geet (n = 36, err = 16.7%)

| | | [6] WeightedToken4 > 1.7: Bhajan (n = 13, err = 38.5%)

| [7] WeightedToken1 > 7.7: DeshBhakti (n = 311, err = 0.0%)


Number of inner nodes: 3

Number of terminal nodes: 4
```

Fig. 2. Decision Tree and Classifier Rules.

```
Confusion Matrix and Statistics


 Reference

Prediction Bal geet Bhajan DeshBhakti Updesh geet

 Bal geet  11  0  0  0

 Bhajan   0  1  0  3

 DeshBhakti 0  1  85   0

 Updesh geet  0  4  0  4


Overall Statistics


 Accuracy : 0.9266

  95% CI : (0.8605, 0.9678)

 No Information Rate : 0.7798

 P-Value [Acc > NIR] : 3.524e-05


 Kappa: 0.8005
```

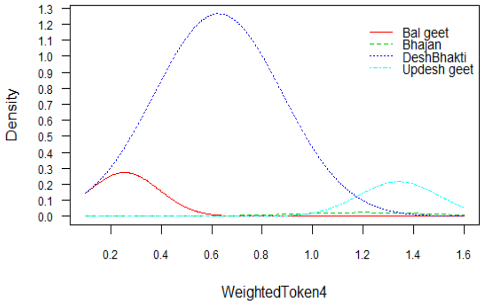


Fig. 3. Model Fitting by Naïve Bayes and Accuracy of Prediction.

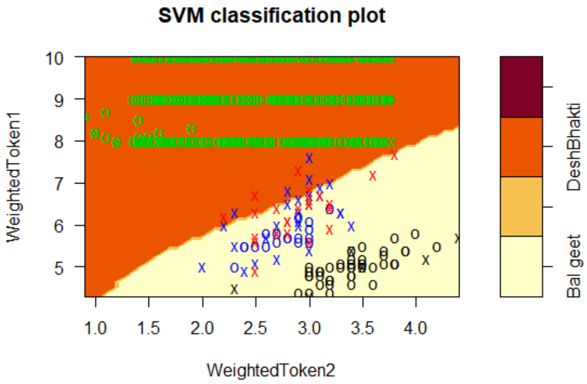


Fig. 4. Hyperplanes Produced by Support Vector Machine.

Fig. 5 shows the confusion matrix along with the Prediction of type of poem carried out using SVM. It is clear that the class accuracy is 0.96, also actual and predicted results are shown that is 11 poems actually belonging to Bhajan class are classified as Updesh geet. All categories of poem can be seen in plot represented by different colours.

Fig. 6 represents the neural network generated for all categories of the poems. Four significant tokens are acting as input to a network. Weights applied by two hidden layers are shown in the figure. The network is trained to identify the tokens most helpful in an accurate classification. These input-weight products are summed and then the sum is passed through a node's activation function. Accuracy of the prediction is calculated comes out to be 0.88 for 500 poems id depicted. Blue coloured lines represent hidden layers.

Table II shows a misclassification error produced by all four algorithms for different corpus size. The support vector machine gives best results for all samples of poems. The error produced by SVM is less for all sample sizes for all four algorithms.

```
Prediction Bal geet Bhajan DeshBhakti Updesh geet

 Bal geet  39 0  0   0

 Bhajan   0 8  0   6

 DeshBhakti  0 0  373   0

 Updesh geet  0  11  0  30


Overall Statistics


   Accuracy : 0.9636

    95% CI : (0.9424, 0.9787)

 No Information Rate : 0.7987

 P-Value [Acc > NIR] : < 2.2e-16


   Kappa: 0.8951
```

Fig. 5. Prediction of Verse type using SVM.

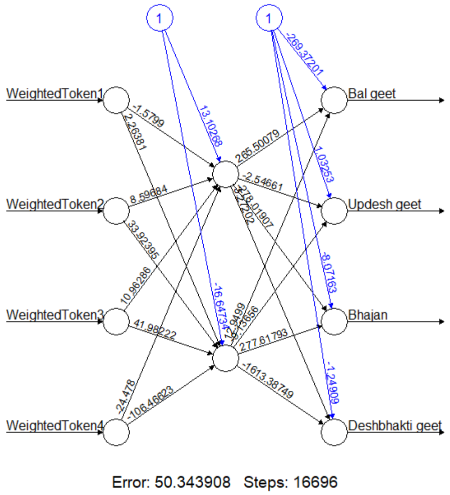


Fig. 6. Neural Network showing Input, Hidden and Output Layer.

TABLE. II. COMPARISON OF MISCLASSIFICATION ERROR PRODUCED BY CLASSIFIERS

Sr. No	Corpus Size	SVM	Decision tree	Neural network	Naïve Byes
1	100	0.45	0.61	50.34	0.51
2	250	0.46	0.63	50.45	0.53
3	400	0.47	0.64	50.82	0.54
4	697	0.47	0.66	50.85	0.55

V. CONCLUSIONS

The current study achieves the prediction of a class of Hindi poem, unlike the other published research works, which have focused on classification of English text. Additionally, the contribution of this study is the exhaustive evaluation of the eager classifiers. The formation of the classes was achieved through the TF-IDF. Government and non-government agencies can use the approach to classify reports, initiatives, different schemes, etc. Experiments are conducted on a corpus of 697 poems. The current work is the first of its kind in the world, which employs prediction and performance evaluation for Hindi corpus comprising of verses.

REFERENCES

[1] Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014, January). Word cloud explorer: Text analytics based on word clouds. In 2014 47th Hawaii International Conference on System Sciences (pp. 1833-1842). IEEE.

[2] Ray, S. K., Ahmad, A., & Kumar, C. A. (2019). Review and Implementation of Topic Modeling in Hindi. Applied Artificial Intelligence, 1-29.

[3] Ramanathan, Ananthakrishnan, and Durgesh D. Rao. "A lightweight stemmer for Hindi." In the Proceedings of EACL. 2003.

[4] Diab, D. M., & El Hindi, K. M. (2017). Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. Applied Soft Computing, 54, 183-199.]

[5] Puri, S., & Singh, S. P. (2018). Hindi Text Document Classification System Using SVM and Fuzzy: A Survey. International Journal of Rough Sets and Data Analysis (IJRSDA), 5(4), 1-31.

[6] Pal, K., & Patel, B. V. (2020). Model for Classification of Poems in the Hindi Language Based on Ras. In Smart Systems and IoT: Innovations in Computing (pp. 655-661). Springer, Singapore.

[7] Saini, J. R., & Desai, A. A. (2011). Identification of Hindi Words Used in Pornographic Unsolicited Bulk E-Mails. IUP Journal of Systems Management, 9(2).

[8] Garg, A., & Saini, J. R. A Systematic and Exhaustive Review of Automatic Abstractive Text Summarization for Hindi Language.

[9] Kaur, J., & Saini, J. R. (2017, February). Punjabi Poetry Classification: The Test of 10 Machine Learning Algorithms. In Proceedings of the 9th International Conference on Machine Learning and Computing (pp. 1-5). ACM.

[10] Kaur, J., & Saini, J. R. (2017). PuPoCl: Development of Punjabi Poetry Classifier Using Linguistic Features and Weighting. INFOCOMP, 16(1-2), 1-7.

[11] Kaur, J., & Saini, J. R. (2018). Automatic classification of Punjabi poetries using poetic features. International Journal of Computational Intelligence Studies, 7(2), 124-137.

[12] Kaur, J., & Saini, J. R. (2016). Automatic Punjabi poetry classification using machine learning algorithms with reduced feature set. International Journal of Artificial Intelligence and Soft Computing, 5(4), 311-319.

[13] Kaur, J., & Saini, J. Designing Punjabi Poetry Classifiers Using Machine Learning and Different Textual Features.

[14] Chandrakar, O., & Saini, J. R. (2016, October). Development of Indian weighted diabetic risk score (IWDRS) using machine learning techniques for type-2 diabetes. In Proceedings of the 9th Annual ACM India Conference (pp. 125-128). ACM.

[15] Audichya M.A., Saini J.R., "Computational Linguistic Prosody Rule-based Unified Technique for Automatic Metadata Generation for Marathi Poetry", proceedings of ICAIT-2019, in press, IEEE, USA.

[16] Audichya M.A. and Saini J.R., 2020, "Computational Linguistic Prosody Rule-based Unified Technique for Automatic Metadata Generation for Hindi Poetry", 1st IEEE International Conference on Advances in Information Technology, Karnatka, India, in press with IEEE.

[17] Saini J.R. and Kaur J., 2020, "Kāvi: An Annotated Corpus of Punjabi Poetry with Emotion Detection Based on 'Navrasa'", Procedia Computer Science, in press with Elsevier.

[18] Bafna P.B., Saini J.R.,2019, "Identification of Significant Challenges in the Sports Domain using Clustering and Feature Selection Techniques", 9th International Conference on Emerging Trends in Engineering and Technology on Signal and Information Processing (ICETET-SIP-19, Nagpur, India, in press with IEEE.

[19] Bafna P.B., Saini J.R.,2019, "Hindi Multi-document Word Cloud based Summarization through Unsupervised Learning", ", 9th International Conference on Emerging Trends in Engineering and Technology on Signal and Information Processing (ICETET-SIP-19), Nagpur, India, in press with IEEE.

[20] Bafna P.B., Saini J.R., 2019, "Scaled Document Clustering and Word Cloud based Summarization on Hindi Corpus, 4th International Conference on Advanced Computing and Intelligent Engineering, Bhubaneshwar, India, in press with Springer.

[21] Bafna P.B., Saini J.R., 2020, On Readability Metrics of Goal Statements of Universities and Brand-promoting Lexicons for Industries, 4th International Conference of Data Management, Analytics and Innovation (ICDMAI 2020), Delhi,India.

[22] Bafna P.B., Saini J.R., 2020, Identification of Significant Challenges Faced by Tourism and Hospitality Industry Using Association rules", 4th International Conference of Data Management, Analytics and Innovation (ICDMAI 2020), Delhi,India.

[23] Bafna P.B., Saini J.R., 2020,"Marathi Text Analysis using Unsupervised Learning and Word Cloud", International Journal of Engineering and Advanced Technology,9(3),in press.

[24] Venugopal G., Saini J.R., Dhanya P., Novel Language Resources for Hindi: An Aesthetics Text Corpus and a Comprehensive Stop Lemma List, International Journal of Advanced Computer Science and Applications, vol. 11(1), Jan. 2020, in press.

[25] Bafna, P., Pramod, D., & Vaidya, A. (2016, March). Document clustering: TF-IDF approach. In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (pp. 61-66). IEEE.

[26] Harrag, F., & El-Qawasmah, E. (2009, August). Neural Network for Arabic text classification. In 2009 Second International Conference on the Applications of Digital Information and Web Technologies (pp. 778-783). IEEE.

[27] Yu, B., Xu, Z. B., & Li, C. H. (2008). Latent semantic analysis for text categorization using neural network. Knowledge-Based Systems, 21(8), 900-904.

[28] Zhang, D., & Lee, W. S. (2003, July). Question classification using support vector machines. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 26-32).

[29] https://aajtak.intoday.in/sahitya-kavita.html.

Teachers' Experiences in the Development of Digital Storytelling for Cyber Risk Awareness

Fariza Khalid[1]
Faculty of Education, Universiti Kebangsaan Malaysia
Bangi, Selangor, Malaysia

Tewfiq El-Maliki[2]
Hepia Hes-So
Geneva Switzerland

***Abstract*—Although the Internet has positively impacted people's lives, it also has its dark side. There have been reports on the increase of cases of violence, racial abuse, cyber-bullying, online fraud, addiction to gaming and gambling, and pornography. A vital issue has emerged that Internet users still lack awareness of these online risks. In this study, our respondents were involved in the development of educational videos related to cyber risk topics using a storytelling approach. The participants in this study were 28 in-service teachers who took a master class on Resource and Information Technology. This study aims to examine the issues that participants took into consideration while planning and developing their digital stories, and their experiences developing digital stories about cyber risks. The data was collected using a written reflection. The data was then analyzed thematically using NVivo Software. The findings indicate how the respondents valued their experience in planning, developing, and evaluating their storytelling videos. The impact of learning from the videos on the students' affective domain is also discussed. We further discuss the benefits of the storytelling approach for behavior change.**

Keywords*—*Cybersecurity; awareness; education; video; digital storytelling; media; case study

I. Introduction

It is an undeniable fact that the way people learn, retrieve information, and construct knowledge has changed with the existence of the Internet and digital media [1], [2]. Although the Internet is considered the most valuable innovation ever created, it also has a dark side that may result in adverse effects on its users, including adults and children [3]-[5]. Among the potential cyber risks are cybersex [6], pornography [7], personal information exposure [6], [8], [9], cyber addiction [10], online fraud, and addiction to gaming and gambling [11].

Cybersecurity is defined as: "the protection of cyberspace itself, the electronic information, the [Information and Communication Technologies] ICTs that support cyberspace, and the users of cyberspace in their personal, societal and national capacity, including any of their interests, either tangible or intangible, that are vulnerable to attacks originating in cyberspace" [12], [13]. However, the process of cybersecurity has become a more human issue, as it demands a human-centered approach rather than merely technical controls.

According to [14], human weakness is considered the most challenging issue to deal with in relation to cybersecurity. This is because behavior, emotions, and feelings towards technology are somewhat unpredictable [15]. Therefore, to create a safe online environment for all users, everyone must be aware of cyber risks so that early prevention can be done. As the issue of awareness is closely related to the affective domain of learning, it is crucial to have a learning medium that applies an approach that can result in a rise in awareness of the importance of being safe online to avoid such risks. In this study, we launched a project in which digital stories were developed by 28 in-service teachers that covered topics related to cyber risks. This study aims to examine the issues that participants took into consideration while planning and developing their digital stories, and their experiences developing digital stories on cyber risks.

A. Digital Storytelling

Digital storytelling is an approach that has been found to be engaging for both students and teachers [16]. Storytelling is an approach for communicating information through the use of words, images, and sounds [17], [18]. Although some authors use the word 'narrative' as a synonym for 'story', narratives can be defined as predominantly factual, whereas stories are reflective, creative, and value-laden, usually revealing something important about the human condition [19]. Digital stories can be produced with a combination of visuals, photos, drawings, voice narration, and music to present a narrative [18], and with the existence of multimedia applications, digital stories can be made even more powerful with the integration of the elements such as texts, graphics, audio, video, and animation. In line with this, [18] also points out that digital stories are similar to short films, with a continuous narrative line.

Digital storytelling has potential for learning, as a combination of images, music, narrative, and voice can help promote deep dimensions and vivid colors for characters, situations, experiences, and insights [20], elevating both students' and teachers' experiences [21] and thus accelerating students' comprehension by boosting their interest in discovering new ideas [22]. The use of digital storytelling also helps teachers to build constructivist learning environments that encourage creative problem-solving based on collaboration and peer-to-peer communication [23]. Digital storytelling has been found to facilitate integrated approaches to curriculum development, and to engage learners in higher-order thinking and deep learning [24], creating more engaging and exciting learning environments [25].

The benefits of digital storytelling have not only been reported in relation to the cognitive domain, but it has also been shown to be helpful in promoting affective learning. [26], for example, highlights that the use of digital stories can develop a sense of connection among learners and mold their attitudes. This is because, through storytelling, learners will develop their listening skills and be able to identify the key messages of the stories, thus expanding their sense of respect and openness [19].

There are three major categories of digital stories: a) personal narratives–stories that contain accounts of significant incidents in one's life; b) historical documentaries–stories that examine dramatic events that help us understand the past; and c) stories designed to inform or instruct the viewer on a particular concept or practice [21]. [27] has formed a comprehensive five-part definition of digital stories: They must a) include a compelling narration of a story; b) provide a meaningful context for understanding the story being told; c) use images to capture and/or expand upon emotions found in the narrative; d) employ music and other sound effects to reinforce ideas; and e) invite thoughtful reflection from their audience(s).

On the other hand, [1] suggests seven elements of digital storytelling: a) the point of view (what is the perspective of the author?); b) a dramatic question (a question that will be answered by the end of the story); c) emotional content (serious issues that speak to us in a personal and powerful way); d) the gift of your voice – a way to personalize the story to help the audience understand the context; e) the power of the soundtrack (music or other sounds that support the storyline); f) economy (simply put, using just enough content to tell the story without overloading the viewer with too much information); and g) pacing (related to economy, but specifically dealing with how slowly or quickly the story progresses).

In this study, the process of developing digital storytelling required teachers who were involved in this project to collaboratively design an educational video using a digital storytelling approach as the end product of the learning process. In this process, they needed to consider themselves as teacher-designers. Participants had to learn specific software skills as and when required by their evolving project. The researchers anticipated that when participants were engaged in their collaborative work, they would develop skills in designing the product, such as how to operate the software needed to develop their animations, video editing, the use of timelines, and how to execute the content to be effectively delivered to end-users. The main role of the instructors was to act as facilitators and problem-solvers, rather than as content experts [28]. Learning in this context involved becoming a practitioner, not just learning about practice [29]. Most importantly, by engaging themselves in the design process, the teachers would build a better understanding of the subject matter, and in how to selecting specific instructional goals instead of general ones. Therefore, according to [28], every act of design is always a process of placing the components of technology, content, and pedagogy together.

II. Methodology

This study employed a single case study research design [30]. A case study is an approach that focuses on one or a small number of groups to investigate a contemporary phenomenon within its real context [31], for the purpose of gaining an in-depth understanding of the "events, relationships, experiences or processes occurring in that particular instance" [32, p. 52].

The total number of respondents was 28 teachers, the majority of them are female (n=24), and the rest male (n=4). In terms of their race, the majority was Malay (n=24), while two were Indian, one Chinese, and one Bidayuh, i.e. one of ethnic groups in Sarawak. The participants were masters students who enrolled in an Information and Technology Resources course. However, the subject they were teaching in school varied. As shown in Table I, the largest number were English Language teachers (n=8), followed by Information and Communication Technology (n=6) and Mathematics (n=5); three participants each taught Science and Malay Language, two participants taught Islamic Education, and one participant taught Tamil Language.

The spread of their teaching experience can be seen in Table II. Nine participants had less than three years of teaching experience. The same number had four to nine years of experience, and eight participants had ten to 15 years of experience.

Concerning their former skills in developing digital storytelling, only six of them admitted to having previous skills in this area, while the other 18 had never been involved in digital storytelling. As this project was about spreading awareness of cyber risks, we also asked about their prior involvement in any project related to spreading awareness about cyber risks. However, none of them had ever had experience in this area.

One of the learning outcomes of the course in which the participants were enrolled is that participants be able to produce animated videos using a storytelling approach. As an assessment of this learning outcome, they were challenged to become teacher-designers so as to allow for ownership of their project. The task was for them to plan, design, and develop videos about topics related to cyber risks using any animation software.

TABLE. I. Respondents' Background

Subject taught	**n**	**%**
English Language	8	28.57
Malay Language	3	10.71
Islamic Education	2	7.14
Science	4	14.28
Information and Communication Technology	6	21.42
Mathematics	4	14.28
Tamil Language	1	3.57

TABLE. II. Respondents' Background

Years of teaching experience	**n**	**%**
> 3 years	9	32.14
4-9 years	10	35.71
10-15 years	8	28.57
18-20 years	1	3.57

To complete the tasks, they were grouped into groups of four or five, and each group was given a choice of topics to be developed into digital stories. The topics were: a) cyberbullying; b) Internet addiction,; c) pornography; d) games addiction; e) oversharing of personal information; and f) grooming. They were then given the freedom to plan, arrange, and develop the content of their digital stories, starting from developing a storyboard, to developing the animation or video, and designing assessments for after the audience has watched the digital story. Examples of their storyboards are shown in Fig. 1. However, an iterative process also took place through which participants were required to present their storylines to all the members of the group, and received constructive feedback on how to improve their plots. Another presentation took place after they had developed a draft video or animation. Examples of their finalised videos are shown in Fig. 2. At this point, other group members told the developers whether the story had an impact on their feelings. This was a crucial phase, as the aim of the project was to promote awareness, so it was important that the 'audience' felt the message during the preview of the digital stories. The overall process of developing the stories took ten weeks.

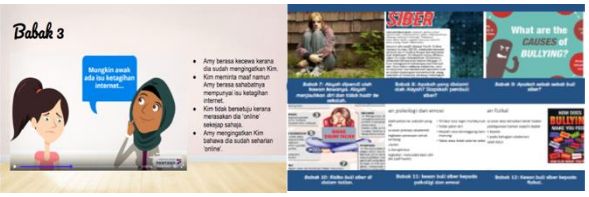

Babak 3
Mungkin awak ada isu ketagihan internet...
Amy berasa kecewa kerana dia sudah mengingatkan Kim.
Kim meminta maaf namun Amy berasa sahabatnya mempunyai isu ketagihan internet.
Kim tidak bersetuju kerana merasakan dia 'online' sekejap sahaja.
Amy mengingatkan Kim bahawa dia sudah seharian 'online'.
SIBER
What are the CAUSES of BULLYING?
BULLYING

Fig. 1. Examples of Storyboards Developed by the Participants.

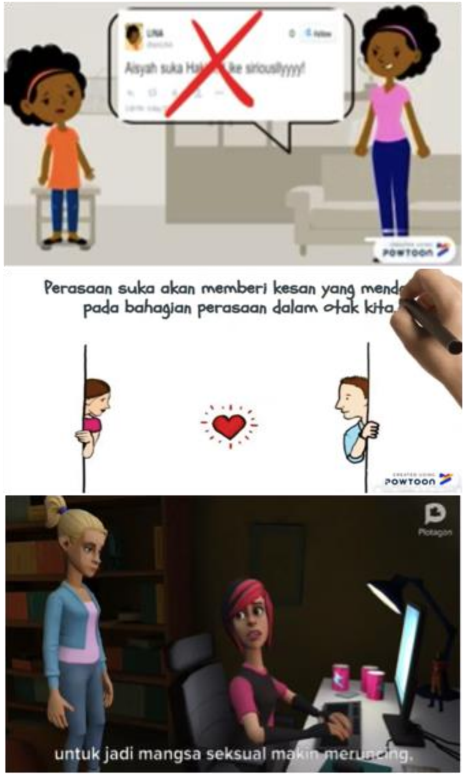

Aisyah suka Ha... ke siriouslyyyy!
POWTOON
Perasaan suka akan memberi kesan yang mend
pada bahagian perasaan dalam otak kita.
POWTOON
Plotagon
untuk jadi mangsa seksual makin meruncing.

Fig. 2. Examples of Finalised Storytelling Videos Developed by the Participants.

For data collection, we used written reflections (given to the teachers as part of their tasks, and of a minimum of four thousand words) prepared during the final week of the semester. The data was then imported to Nvivo 12 Plus software for coding purposes, and analyzed using thematic analysis [32]. This study aimed to explore the experiences of the participants throughout the whole process as teacher-designers. Our research questions were:

1) What were the considerations teachers took into account in the development of their digital stories?

2) What are teachers' views on their experiences developing digital stories about cyber risks?

III. Findings and Discussions

In this section, we present the findings of the analysis.

Teachers' considerations in the development of digital stories.

The analysis resulted in the emergence of six main themes related to teachers' considerations for the planning, design, and development of their animated videos: a) content; b) characters; c) language; d) narration; e) multimedia elements; and f) duration of the video (see Table III).

1) Content: In terms of the content, as many of the teachers were not very knowledgeable about the topics they were given, the first thing they had to do was to understand the content related to the topic. This included searching for facts available from different sources such as journal articles, books, and websites. As teachers, they were trained to ensure that the delivery of content was based on the facts, and that the arrangement of the information needed to be to suit the level of the target audience, i.e. school students aged from 10 to 14 years old.

Sample statements include: "Since our target audiences are primary and lower secondary school students, we were careful in the selection of the content and the development of the story. We also used simple language to make sure the message is delivered effectively." (Joecy).

"We took many factors into consideration. To make it impactful, we simply provided related facts. We created a storyline to illustrate how a teenager who is addicted to games started to decline in their academic performance, and the deterioration of his health as well. In the end, we showed the dos and don'ts of playing online games." (Ummu).

To enhance the impact of the videos, some of the respondents shared their strategy of developing a storyline that had a similarity to audiences' actual lives.

TABLE. III. Teachers' Considerations

Sub-themes	**Numbers of participants cited**	**Percentage**
Content	22	78.5%
Characters	18	64.2%
Language	15	53.5%
Narration	13	46.4%
Multimedia elements	24	85.7%

"In terms of content delivery, we tried to make our video easy-going for the students. We think that by providing environments similar to their actual lives, it will help to engage them more." (Reene).

For example, for the topic of games addiction, the team members agreed to use PUBG and Fortnite as examples. One of them explained.

"We chose PUBG and Fortnite games in our story. We know that our students love to play these games, so by choosing examples from their real lives may seem more realistic and impactful for them." (Ruhi).

Another respondent also presented an example of how her teammates used examples of students' daily lives in their video.

"Our topic is on grooming. It was a tough topic to be developed as we do not have much information on how it happens in real life, especially in our country. However, I talked to my little sister, and she pointed out the use of WeChat among teenagers, and many of them used the platform to make a connection with adult males. Her friend was once offered a thousand bucks by a man she knew from WeChat if she shared her half-naked picture! So we decided to use that scenario in our video." (Hannah).

The selection of issues related with audiences' actual lives may help them to relate the message to their personal lives and Internet use. Author in [17], in his seven elements of digital storytelling, stresses the importance of personalizing a digital story to help the audience understand the context better.

2) Characters: Participants also pointed out the aspects of the characters that they took into consideration while planning their digital stories. 18 participants mentioned that their group chose to create teenage characters to make their digital stories close to the audiences.

Some of the participants also mentioned they created characters from different races, for instance Malay, Chinese and Indian, as symbols of a multiracial country like Malaysia. For example:

"We wanted the video to have a simple yet powerful message. We chose characters that are similar to their real world. We include characters from different races in our story to promote unity and togetherness as Malaysians as well. We also created characters like counselor, parents, and peers to show who are the significant others for them to turn to if similar case happens to them. " (Shasha).

3) Language: Participants also shared their considerations about the language used in the video. All the groups agreed to use their first language, Malay language, as the medium of their stories. This was decided through discussions among peers in the class. Sample answers include:

"The approach used was very much based on the level of the audience. The narration text was developed according to the pupils' language level. The language chosen was our mother tongue, as it is more familiar to the pupils and thus would help in their total understanding of the story delivered." (Faiza).

The aspect of simplicity, or 'economy' is mentioned by [21] as a vital characteristic of digital storytelling, in that a digital story has to be simply put, using just enough content to tell the story without overloading the viewer with too much information. In addition to the selection of Malay as the language medium, participants also used simple language to suit the level of their target audiences, i.e. students aged from ten to 14 years old. Sample answers include.

"Lastly, we used simple and clear language which is suitable for the level of understanding of primary and secondary school students." (Zaidi)

"Since our target audiences are those in primary and lower secondary school, we were careful in the selection of the content and the development of the story. We also used simple language to make sure the message is delivered effectively. We made it a short and sweet kind of video." (Diane).

Another vital aspect that was taken into consideration was the narration. In order to ensure that the message was successfully delivered, the participants had to choose suitable narrators. Some sample responses include.

"Now that we have developed a script with simple language, there comes a problem. Who is going to be the narrator? It was a funny moment for us when all of us tried to record our voices, and we listened to them together to finally choose the 'best' voice! We are teachers, of course, we have loud voices, but we need to pick the most 'marketable' voice!" (Armie).

"As our digital story uses third-person narration, we need to choose the right voice. We want it to sound pleasant and professional. This is our first baby, so we want to produce the best." (Ila).

4) Narration: In terms of the narration, the majority of the participants used third-person narration. However, several mentioned using first-person narration. For example:

"We agreed to use third-person narration in our story. Our topic is Internet addiction. So our approach is to create a story that includes a self-check approach to help our audience to see whether they count as addicted to the Internet or not. In this case, we think it is better achieved if we use third-person narration." (Kumar)

"Our group used first-person as well as third-person narration. We divided our story into two main sections. The first section was a story of how a girl was bullied by her friends online. The second section included a scene where the girl reflects on how she felt about cyberbullying and what lessons should be learned from her story." (Daisy).

5) Elements of multimedia: As mentioned earlier, this project was about the development of animated videos about cyber risk topics. The participants were empowered to choose any software they preferred to develop their stories. Among the software packages used by the participants were Powtoon, Pixtoon, and Plotagon. Participants also considered aspect of multimedia such as audio, music, animation, and colors. The analysis shows that participants agreed on the fact that the use of multimedia elements such as audio, text, graphics, and animation helped them produce better and more impactful learning material. This finding corroborates [18], who also opines that the use of multimedia elements can enhance the power of message delivery. Sample answers include.

"Furthermore, the background music was added to make the video more interesting. We also included the element of feeling. The characters in the story portray feelings to influence the audience or help them relate to the feelings." (Hannah).

"I think what makes the video appealing is because we use animation and music. After getting feedback from peers during the tenth week of class, we got the chance to improve the quality of our music and sound." (Roy).

"The inclusion of recorded audio was also varied according to its suitability for the story in order to give emphasis and highlight important aspects of the story, as well as making the story more interesting for the audience." (Nani).

Digital storytelling has its own potential as a learning material, as a combination of images, music, narrative, and voice can help promote deep dimensions and vivid colors for characters, situations, experiences, and insights [20].

6) Participants' views on their experiences in developing digital storytelling: Based on the overall findings, participants shared their positive thoughts about spreading awareness of cyber risks, especially by using digital stories. 18 participants stated that the approach was a unique way to affect the hearts of the target audiences, compared to giving a lecture or telling them about dos and don'ts. For example, one participant said.

"I am positive that this kind of video can be an effective learning tool for our students. I have tried three videos with my students, who are in Form One [13 years old], and they love it so much. What is important is that we, as teachers, use the material in a meaningful way. Definitely, the video can be used to spur discussion among students, or we can ask them to do some reflections on what they have done in their daily internet use." (Ziana).

This finding is in line with [33], who posits that the use of digital storytelling in class activities can encourage students' insights, and may spur more in-depth discussions. In addition, participants also mentioned that the digital storytelling approach could convey a clear message without burdening learners' cognitive loads: this is because in digital storytelling, other elements of multimedia are used [17]. In this study, the use of animation and narration helped students to gather the information without having to read a dense text. Instead, the videos involved them in thinking about the message and relating it with their own life.

"My reflection on the use of storytelling is that it helps to lessen the cognitive loads of the students, and the presentation of the message can be done more easily with storytelling." (Wani).

Participants in this study also agreed that digital storytelling is an appropriate approach to develop learners' affective domain.

"It [storytelling] can be used to address the affective part of the students. The message in the video is not delivered like a class note; rather, the message is conveyed indirectly through the storyline. It makes learning more fun, too." (Roy).

"We know that from our experience, teaching the affective domain is not easy to achieve. What makes it more challenging is that the affective domain is difficult to assess. You may hear your students say: 'Yes, teacher, I know it is wrong and I can assure you that I will never do such a thing.' Nevertheless, how many of us can guarantee that it will never happen? Hopefully, using digital stories that cater to their affective domain will have an impact on the students." (Daisy).

This finding supports [26], who highlights that the use of digital stories can develop a sense of connection among learners and mold their attitudes. This is because, through storytelling, learners will develop their listening skills and be able to identify the key messages of the stories, thus expanding their sense of respect and openness, and promoting the development of wisdom [19].

"What is beautiful about storytelling is that it helps us to convey the messages more easily in a more powerful way of. Conservatively, we can produce a slide on dos and don'ts. Very dry and monotonous. In contrast, storytelling is more fun to work with. We were challenged to be more creative and more sensitive in the selection of the words and language. This is something that can trigger us, teachers in a school, to learn and do more about the power of storytelling." (Faiza).

Involvement in digital storytelling also made the participants feel engaged in their projects. This supports [16], who claim that digital storytelling is engaging for both learners and teachers.

Other findings derived from the participants' reflections relate to their experiences as teacher-designers. Some of the participants shared that their experiences had been memorable. For example, a number of participants highlighted that through the whole process of developing their digital stories, they had learned many skills, including storyboarding, handling animation software, video editing, voice recording, and decision-making. They explained that the project required higher-order thinking skills, as everything had to be decided among the members of their group. This is closely related to [28]: according to [28], when teachers are given a role as designers, they will develop skills related to what they have been doing. In such situations, teachers have to exercise problem-solving skills and become practitioners [29]. Through their engagement in the design process, teachers will build a better understanding of the subject matter, and learn how to select specific instructional goals instead of general ones [28].

IV. Conclusions

In this study, we have examined the considerations that teachers perceived as important while planning and developing their digital stories, and their experiences in developing digital stories about cyber risks. Although there are guidelines for digital storytelling, based on the findings, we believe that the considerations are most closely related to the cultural stance of the developers and the targeted audience, for instance in terms of the selection of characters, language, and storyline. This study also further supports the fact that digital storytelling as a powerful technological tool, not only to promote affective learning and positive values, but also to enhance the skills and content knowledge of developers.

ACKNOWLEDGMENTS

This study was conducted under GG-2019-066 and FRGS/1/2017/SSI09/UKM/02/3 grants.

REFERENCES

[1] F. Khalid, “Understanding University Students’ Use of Facebook for Collaborative Learning”, International Journal of Information and Education Technology, vol. 7, no. 8, pp. 595-600, 2017. Available: 10.18178/ijiet.2017.7.8.938.

[2] N. Zakaria and F. Khalid, “The Benefits and Constraints of the Use of Information and Communication Technology (ICT) in Teaching Mathematics”, Creative Education, vol. 07, no. 11, pp. 1537-1544, 2016. Available: 10.4236/ce.2016.711158.

[3] A. A. Karim, P. M. Shah, F. Khalid, M. Ahmad and R. Din, “The Role of Personal Learning Orientations and Goals in Students’ Application of Information Skills in Malaysia”, Creative Education, vol. 06, no. 18, pp. 2002-2012, 2015. Available: 10.4236/ce.2015.618205.

[4] N. Ahmad, U. A. Mokhtar and Z. Hood, “Cyber Security Situational Awareness among Parents”, in Cyber Resilience Conference, Putrajaya, Malaysia, 2019, pp. 7-8.

[5] F. Khalid, M. Y. Daud, M. J. A. Rahman and M. K. M. Nasir, “An Investigation of University Students’ Awareness on Cyber Security”, International Journal of Engineering & Technology, vol. 7, no. 421, pp. 11-14, 2018 [Accessed 1 January 2020].

[6] V. Ratten, “A cross-cultural comparison of online behavioural advertising knowledge, online privacy concerns and social networking using the technology acceptance model and social cognitive theory”, Journal of Science and Technology Policy Management, vol. 6, no. 1, pp. 25-36, 2015. Available: 10.1108/jstpm-06-2014-0029.

[7] M. D. Griffiths and D. Kuss, “Online addictions, gambling. video gaming and social networking”, in The handbook of the psychology of communication technology, S. Sundars, Ed. Chichester: John Wiley, 2015, pp. 384-406.

[8] L. Mosalanejad, A. Dehghani and K. Abdolahifard, “The Students’ Experiences Of Ethics In Online Systems: A Phenomenological Study”, Turkish Online Journal of Distance Education, vol. 15, no. 4, 2014, pp/ 205-216. Available: 10.17718/tojde.02251.

[9] D. Ktoridou, N. Eteokleous and A. Zahariadou, “Exploring parents’ and children’s awareness on internet threats in relation to internet safety”, Campus-Wide Information Systems, vol. 29, no. 3, pp. 133-143, 2012. Available: 10.1108/10650741211243157.

[10] F. Annansingh and T. Veli, “An investigation into risks awareness and e-safety needs of children on the internet”, Interactive Technology and Smart Education, vol. 13, no. 2, pp. 147-165, 2016. Available: 10.1108/itse-09-2015-0029.

[11] L. Muniandy and B. Muniandy, “State of Cyber Security and the Factors Governing its Protection in Malaysia”, International Journal of Applied Science and Technology, vol. 2, no. 4, pp. 106-112, 2012.

[12] R. Van Solms and J. van Niekerk, “From information security to cybersecurity”, Computers & Security, pp. 1-6, 2013.

[13] A. Klimburg, National cybersecurity framework manual. Talinn: NATO CCD COE Publications, 2012.

[14] S. Pfleeger and D. Caputo, “Leveraging behavioral science to mitigate cyber security risk”, Computers & Security, vol. 31, no. 4, pp. 597-611, 2012. Available: 10.1016/j.cose.2011.12.010.

[15] M. Bada, A. M. Sasse and J. R. Nurse, “Cyber Security Awareness Campaigns: Why do they fail to change behaviour?”, ArXiv, 2014.

[16] Dockter, D. Haug and C. Lewis, “Redefining Rigor: Critical Engagement, Digital Media, and the New English/Language Arts”, Journal of Adolescent & Adult Literacy, vol. 53, no. 5, pp. 418-420, 2010. Available: 10.1598/jaal.53.5.7.

[17] B. Robin, “The Power of Digital Storytelling to Support Teaching and Learning”, Digital Education Review, vol. 30, pp. 17-29, 2016.

[18] T. A. Campbell, “Digital Storytelling in an Elementary Classroom: Going Beyond Entertainment”, Procedia - Social and Behavioral Sciences, vol. 69, pp. 385-393, 2012. Available: 10.1016/j.sbspro. 2012.11.424.

[19] C. Haigh and P. Hardy, “Tell me a story – a conceptual exploration of storytelling in healthcare education”, Nurse Education Today, vol. 31, no. 4, pp. 408-411, 2011. Available: 10.1016/j.nedt.2010.08.001.

[20] M. Razmi, S. Pourali and S. Nozad, “Digital Storytelling in EFL Classroom (Oral Presentation of the Story): A Pathway to Improve Oral Production”, Procedia – Social and Behavioral Sciences, vol. 98, pp. 1541-1544, 2014. Available: 10.1016/j.sbspro.2014.03.576.

[21] B. Robin, “Digital Storytelling: A Powerful Technology Tool for the 21st Century Classroom”, Theory Into Practice, vol. 47, no. 3, pp. 220-228, 2008. Available: 10.1080/00405840802153916.

[22] L. Burmark, “Visual presentations that prompt, flash & transform”, Media and Methods, vol. 40, no. 6, pp. 4-5, 2004.

[23] E. Dakich, “From ICT competencies to the social practice of ICT-rich pedagogies: Results of a Delphi study”, in Australian Association for Research in Education (AARE) Conference 2004, Melbourne, Australia, 2004.

[24] M. Smeda, E. Dakich and N. Sharda, “Developing a framework for advancing e-learning through digital storytelling”, in IADIS International Conference e-learning, Freiburg, Germany, 2010, pp. 169-176.

[25] N. Smeda, E. Dakich and N. Sharda, “The effectiveness of digital storytelling in the classrooms: a comprehensive study”, Smart Learning Environments, vol. 1, no. 1, 2014. Available: 10.1186/s40561-014-0006-3.

[26] T. A. Abma, “Learning by Telling”, Management Learning, vol. 34, no. 2, pp. 221-240, 2003. Available: 10.1177/1350507603034002004.

[27] B. Alexander, The new digital storytelling. Santa Barbara, Calif: Praeger, 2017.

[28] M. Koehler and P. Mishra, “Teachers Learning Technology by Design”, Journal of Computing in Teacher Education, vol. 21, no. 3, pp. 94-102, 2005.

[29] J. S. Brown and P. Duguid, “Organizational Learning and Communities-of-Practice: Toward a Unified View of Working, Learning, and Innovation”, Organization Science, vol. 2, no. 1, pp. 40-57, 1991. Available: 10.1287/orsc.2.1.40.

[30] M. Denscombe, The good research guide, 4th ed. Maidenhead: Open University Press, 2014.

[31] R. K. Yin, Applications of case study research, 4th ed. Thousand Oaks, CA: Sage, 2009.

[32] V. Braun and V. Clarke, “Using thematic analysis in psychology”, Qualitative Research in Psychology, vol. 3, no. 2, pp. 77-101, 2006. Available: 10.1191/1478088706qp063oa.

[33] R. A. Berk, “Teaching strategies for the net generation”, Transformative Dialogues: Teaching & Learning Journal, vol. 3, no. 2, pp. 1-23, 2009.

e-Participation Model for Kuwait e-Government

Zainab M. Aljazzaf[1], Sharifa Ayad Al-Ali[2], Muhammad Sarfraz[3]
Department of Information Science, Kuwait University
Shdadiya, Kuwait

***Abstract*—Internet has an influence on every aspect of modern life. The increasing interest in e-government has led to increase in public expenditure on communication technologies. The technology provides and facilitates opportunities for citizens to interact with e-government, so called e-participation. In fact, it makes the citizens involvement higher in the delivery of services, administration, and decision making. People need to engage themselves and participate in e-government to achieve its objectives. e-Government literature explored the factors that influence people to participate in e-government. However, the study of e-participation is new in Kuwait. Therefore, this paper aims to find out the critical factors affecting e-participation in Kuwait. To attain the purpose of the research study, a conceptual model has been developed, keeping the context of Kuwait society in view. Then, a questionnaire has been designed and used to test the conceptual model. The results indicate that technical factors, social influence, political factors, perceived usefulness, and perceived ease-of-use are the significant factors that influence the citizen's intention to participate in Kuwait e-government. Consequently, the results of this study need to be adopted by the government to enhance e-participation in Kuwait e-government.**

Keywords—e-Government; e-participation; e-participation factors; e-participation model; e-information; e-consultation

I. Introduction

In the recent years, societies have reached to very high levels of complexities, as compared to its past, in its day to day life. It has been highly surrounded by technology such as internet, computers, and mobile phones. Consequently, the high amount of work cannot be successfully achieved without the effective communication systems. It has been vital to make use of various technologies including Web and mobile networks. There is a growing trend among the citizens and public administrations to get along with the electronic life to communicate with each other conveniently.

e-Government is related to Information and Communication Technologies (ICTs) and it aims to develop the efficiency and quality of public administration [1]. e-Government initiatives can be traced back to the 1960s, but in the late 1990s, the term e-government began to take form [2]. According to West [3], e-government is the use of internet or other digital means to deliver the government information and services online. Moreover, the United Nations described e-government as "Utilizing the Internet and the world-wide-web for delivering government information and services to citizens" [4].

In the literature, e-government has been used occasionally to refer to e-participation [5]. Participation is related to community, public, and close in meaning to engagement, empowerment, and involvement [6]. The United Nations [7] expressed the e-participation as: provide citizens with greater e-information for decision making, promote e-consultation for the processes of participation and deliberation, and strengthen e-decision making through improving citizen input in decision making.

Moreover, the Organization for Economic Co-operation and Development (OECD) [8] defined e-participation as using ICT to support information delivery to citizens where this information is related to public policies and government activities. In general, participation is supported by ICT in government and governance for administration, policy making, decision making, and service delivery; where the users are citizens and customers [9].

e-Participation aims to attain many goals, such as: minimizing coordination and transaction costs in political and society relationships, better deliberativeness, promoting the ability of information-processing in information technology, raising e-information, enhancing e-consultation, and supporting e-decision making [6][10]. E-information, e-consultation, and e-decision making are the three dimensions of e-participation framework [11].

Moreover, e-government and e-participation research assisted governments to refocus on the citizens, businesses, technologies, and tools, which lead to effective and efficient public administration systems [5]. e-Participation focuses on the development of ICTs which support participation in government processes [12]. For example, ICT tools that have been implemented in e-participation initiatives include e-mails, online discussion forums, online chat, online surveys, and group support systems [13].

United Nation Survey [14] measures e-participation index based on the e-participation framework dimensions: e-information, e-consultation, and e-decision-making, as shown in Table I. The table shows the leading countries in e-participation, such as Denmark, Finland, and Republic of Korea. However, Kuwait has an e-participation index between 0.50 and 0.75, which needs to be improved.

The main objective in e-participation is to achieve involvement and engagement of practitioners in decision making. This means that it is important to have high e-participation from citizens in the government's portals where the online services and information are available.

e-Government literature explores the factors that affect e-participation. However, the research work regarding e-participation in Kuwait is relatively new. Therefore, this research is dedicated towards e-participation in Kuwait e-government. Specifically, it identifies the reason behind people participation in e-government and seeks to find out factors which can influence e-participation in e-government in Kuwait. Consequently, it examines the following points:

1) What are the factors affecting e-participation in Kuwait?

2) Does Kuwait have a low level of e-participation in e-government portal and why?

3) How to attract citizens and residents to participate in e-government portal?

The rest of the paper is organized, as follows: The related work is presented in Section II. Section III discusses the research methodology. Data analysis and discussion is presented in Section IV and Section V, respectively. Section VI concludes the paper.

TABLE. I. E-PARTICIPATION INDEX TOP 5 COUNTRIES IN 2018 [14]

Rank	Country	Name
1	Denmark	1
1	Finland	1
1	Republic of Korea	1
4	Netherlands	0.9888
5	Australia	0.9831
5	Japan	0.9831
5	New Zealand	0.9831
5	Spain	0.9831
5	United Kingdom of Great Britain and Northern Ireland	0.9831
5	United States of America	0.9831

II. RELATED WORK

Allowing citizens to interact with government through e-participation faces many critical issues. Several studies were done to examine the factors affecting e-participation [15]-[22]. In fact, many studies have found the factors that influence citizens' engagement in e-government services and e-participation models.

Colesca and Dobrica, [23] explored the factors that affect citizens' adoption of e-government services in Romania. They have used Technology Acceptance Model (TAM) and realize that citizen's higher perception of usefulness, ease of use, quality, and trust of e-government services directly enhanced their satisfaction and implicitly the level of adoption of e-government.

Reddick [15] examined citizen interaction with e-government using three e-participation models. They used quantitative method (Survey) and found out that citizens have mostly used e-participation for management activities and less for consultative and participatory activities. Moreover, the factors that affect e-participation level are: demand by citizens for e-government, the digital divide, and political factors.

Macintosh [11] presented three factors that affect e-participation. These are ICT infrastructure, human capital, and governance. The ICTs infrastructure is measured by the indicators: PCs, Internet users, telephone lines, online population, mobile phones, and TVs. The indicated capital is measured by education, income, productivity, skills and knowledge.

Moreover, a study in Zambia [18] stated that there are many factors that affected adoption of e-government where ICT has been employed to sustain e-government initiatives. The paper assessed the issues, opportunities, and challenges at the same time with the e-government adoption criteria. The findings reported that the factors that led to delay in the adoption of e-government in Zambia are due to lack of adequate ICT infrastructure and political will, lack of appropriate change management procedures, provision of content in English other than local languages, and non-contextualization of e-government practices.

Millard, Nielsen, Warren, Smith, and Macintosh [13] identified factors affected e-participation in Singapore and determined many factors which are divided in two groups. The first group is called access factors, it includes infrastructure, platforms, website accessibility, financial assistance, and access to e-service. The second group is called knowledge and involves international collaboration, knowledge training, and content availability. Furthermore, a significant study in [17] maps the factors that shape the development of e-participation. It developed an analytical framework to recognize the key variable, internal factors, and external factors. The internal factors are top-level impacts, middle level outcomes, base level operational outputs and raw material. The external factors are political culture, public service culture, legal environment, policy environment, autonomy, technology, and socio-economic environment.

Stoiciu [20] stated that according to various studies and surveys by different organizations in many countries, there are problems in implementing e-participation. However, strengthening e-participation faces four types of barriers, which are: Political Barriers, social barriers, technology barriers, and human/emotional barriers.

The study concluded with four solutions to engage citizens, benefits of citizen inclusion, and better e-participation tools, as follows: (a) Involvement of the civil society in decision making power to develop the value of associative life and democratic systems, (b) Better use of resources and the appropriate development arise as the effectiveness and the quality of the governance increase, (c) Empowerment of citizens should be supported by public authorities and non-government organizations community who organize interventions, (d) Better motivations where adequate and long-term participation needs that regional and local governments and authorities involve in a transparent and open process.

Ahmad, Markkula, and Oivo [24] explored the factors that influenced end-user adoption of e-government services in Pakistan. The research work is based on Unified Theory of Acceptance and Use of Technology (UTAUT) model. It finds that performance of expectancy; effort expectancy, facilitating conditions, and social influence are the factors that affect citizen's adoption of e-government services in Pakistan. As a result, they realize that it is important to understand citizens' needs, run advertising campaigns to increase citizens' awareness, present the role of citizens, and raise the users' confidence in the system.

Ali and Ali [21] investigated the factors that affect citizen's acceptance and readiness to use e-participation tools in Kingdom of Bahrain. These factors are optimum, innovation, insecurity, and discomfort. They used Technology Readiness Acceptance (TRA) model, which combined TAM and Technology Readiness Index (TRI) to find out the positive and negative aspects regarding the technology beliefs. As a result, optimism and innovation affect usefulness and ease of use factors while insecurity and discomfort did not affect usefulness factor. However, insecurity does not affect ease of use factor.

AlAwadhi and Morris [25] studied the factors that influence the acceptance of Kuwait e-government services, making e-government initiatives success depends on two points; government support and the adoption of e-government services by citizens. The authors used the UTAUT model and found the factors that influence the acceptance of e-government services. The factors are linked to technology issues, lack of awareness, usefulness, ease of use, cultural and social influences, and reforming bureaucracy.

Aljazzaf [26] studied the factors influencing people in Kuwait to trust e-government. The author developed a model and tested it through a survey. The result showed that factors such as perceived usefulness, security, perceived ease of use, and Website quality affect people in Kuwait to trust and use e-government.

The presented researches express various factors that affect e-participation. The work, in this paper, aims to find out about factors that affect e-participation in e-government in Kuwait. This study proposes the factors influencing e-participation in Kuwait in concern to many previous studies. Consequently, a model is built to identify the critical factors to help increase participation in e-government and improve citizen's satisfaction in Kuwait e-government.

III. Research Methodology

This section presents the data methodology and includes the proposed research model and questionnaire.

A. *The Proposed Research Model*

This section presents the research proposed model, as shown in Fig. 1. The figure displays all the factors and relationships among them that represent the hypotheses. The model places the constructs used by TAM model, extracted from the literature, and other factors that are mostly related to Kuwait culture. The factors are the technical, demographic, social influence, political, perceived quality, perceived usefulness, perceived ease of use, and intention to participate.

The following presents the discussion of the factors in the research proposed model and the hypotheses:

1) Technical factors: Technical factors refer to the website design and content, channel of communication, and infrastructure. In fact, having good technical factors lead to better e-government services, lower cost, and reduce wastage. First, Website design and content impact the users experience and how they interact with the website. The more clear, easy, and simple website design the more users enter and use the services.

Second, using channels of communication which are classic such as telephone, email, Fax, and SMS; and the other communication channels used by most young users like; social media, television, radio, and mobile apps. Knowing which communication channel(s) the users preferred is important to easily connect with the government.

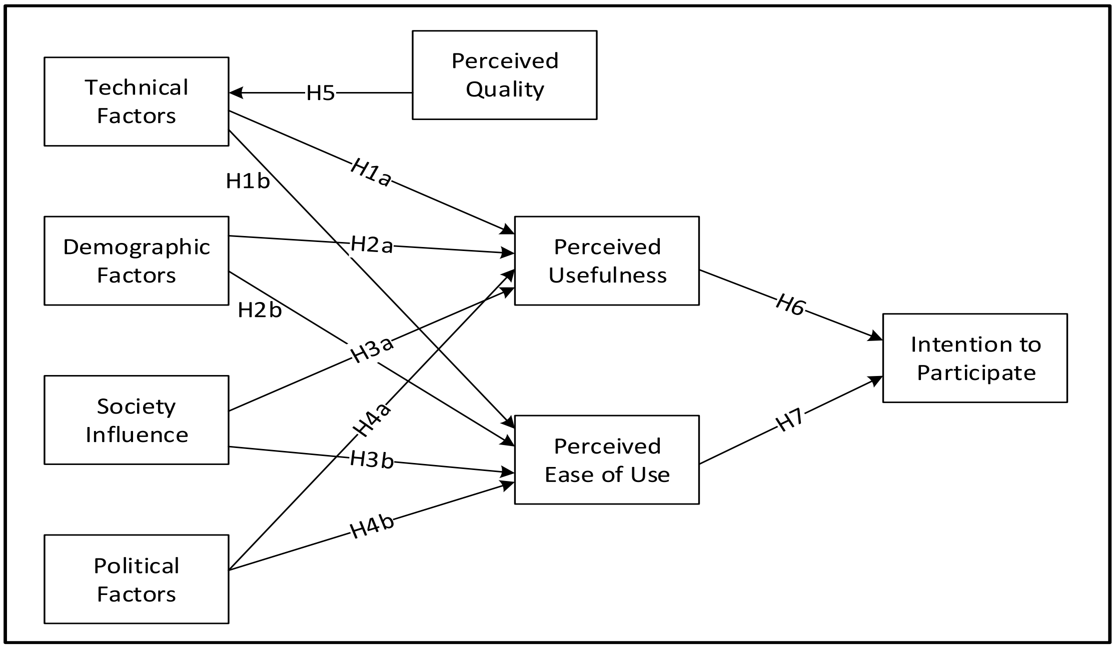


Fig. 1. The Research Model.

Third, the infrastructure refers to the hardware, software, network resources, and servers. It allows the government to deliver the services to users in the best possible way. Technology infrastructure is important in the success of the e-government. However, good infrastructure saves users and government time, money, and effort. The following are the hypotheses which assigned to the technical factors.

- H1a: Technical factors are positively related to perceived usefulness.
- H1b: Technical factors are positively related to perceived ease of use.

2) Demographic factors: Demographic factors are the characteristics of the population expressed statistically such as; the gender, education level, internet experience, and occupation. Demographic is used by government and other institutions to help understand people's characteristics more clearly. Therefore, the following are the hypotheses which presented the demographic factors.

- H2a: Demographic factors have influence on perceived usefulness.
- H2b: Demographic factors have influence on perceived ease of use.

3) Social influence: Social influence is an important factor especially in Kuwait culture. It is the persuasive influence we have on one another. Many users change their view about a new system because of the effect that a person does on others, such as their peers and respected superiors. The following are the social influence hypotheses.

- H3a: Social influence is positively related to perceived usefulness.
- H3b: Social influence is positively related to perceived ease of use.

4) Political factor: The political factor is the level of trust in government, government commitment, and the political party affairs, which have effect on the user's intention to e-participate. This leads to the following hypotheses:

- H4a: Political factors are positively related to perceived usefulness.
- H4b: Political factors are positively related to perceived ease of use.

5) Perceived quality: Using a high quality system increases the technical factors quality. Examples of quality measures include Website usefulness and accuracy. Therefore, this leads to the following hypothesis:

- H5: Perceived quality is positively related to Technical factors.

6) Perceived usefulness: The more the system enhances citizen`s job performance the more citizens will go for electronic participation, Therefore, this leads to the following hypothesis:

- H6: Perceived of usefulness is positively related to e-participation level.

7) Perceived ease of use: More the system saves effort; more the citizens will go for electronic participation. Therefore, this leads to the following hypothesis:

- H7: Perceived ease of use is positively related to e-participation level.

Table II summarizes the hypotheses and their description.

B. *The Research Questionnaire*

After developing the research model, a questionnaire has been designed and the data collected. The first task done here was designing the questionnaire to match the questions to the factors.

The questionnaire was developed to experimentally verify the proposed model and gather the necessary information. The questionnaire is divided into sections. Each section has a set of questions that refers to a factor in the model.

This study surveyed various groups of citizens and residents to get their point of view about Kuwait e-government portal, their experience, and their intension to participate on e-government portal.

TABLE. II. THE RESEARCH HYPOTHESES

N	Hypothesis	Description
H1a	Technical factors is positively related to perceived usefulness	Having better technical factors lead to better job performance.
H1b	Technical factors is positively related to perceived ease of use	Having excellent technical factors save effort.
H2a	Demographic factors have influence on perceived usefulness	Gender, education level, internet experience, and occupation influence the job performance.
H2b	Demographic factors have influence on perceived ease of use	Gender, education level, internet experience, and occupation influence the effort spent in using the system.
H3a	Social influence is positively related to perceived usefulness	The more peers and respected superiors who have positive experience using a system the more new users will exist.
H3b	Social influence is positively related to perceived ease of us	The more peers and respected superiors who have positive experience using a system the more users will save effort.
H4a	Political factors is positively related to perceived usefulness	The more a citizen trust in government the more the user will believe that the system enhance his/her job.
H4b	Political factors is positively related to perceived ease of use	The more a citizen trust in government the more the user will believe that using the system will save effort.
H5	Perceived quality is positively related to technical factors	Using a high quality system increase the technical factors quality.
H6	Perceived usefulness is positively related to e-participation level	The more the system enhance citizen`s job performance the more citizens will go for electronic participation.
H7	Perceived ease of use is positively related to e-participation level	The more the system saves effort the more the citizens will go for electronic participation.

The questionnaire consists of 38 questions, easy and clear for participants to answer, as shown in Appendix 1. The first part includes questions regarding demographic factors such as age, gender, and familiarity with internet. Other questions are regarding the proposed e-participation factors. The questions are measured on a five-point scale of “Strongly disagree” to “Strongly agree”. For example, Part C of the questionnaire presents the technical factors. It includes six questions about the web portal usage, contents, and internet, as shown in Table III.

The questionnaire was conducted using a professional survey website SurveyMonkey. The questionnaire was distributed online only via Twitter, WhatsApp and email. The target population of the study was chosen to be the citizens and residents of Kuwait. These were selected as the questionnaire population because they are the main users of Kuwait e-government portal. Also, knowing their point of view about the Kuwait e-government will help improving it. Initially, the questionnaire was pretested to an appropriate sample of 15 people varying in gender, age, occupation, education level, and internet usage to make sure that it is applicable for distribution. More importantly, it was translated into Arabic and opened for a month from April 13 to May 12, 2015. During this period of time, as many as 508 responses were received.

TABLE. III. THE TECHNICAL FACTOR IN THE QUESTIONNAIRE

C.	**Technical Factor**
1	Registration in the portal is easy
2	Using the e-government portal is easy
3	The e-government portal involves many important services
4	The services provided online make the procedures easier and simple
5	The government services provided online are of high quality
6	Weak internet in Kuwait prevents citizens from the portal usage

IV. DATA ANALYSIS AND FINDING

This part presents the analysis of the data collected through the online survey. As many as 508 persons have responded to the questionnaire and only 188 completed it. Table III presents the respondents’ profile that represents the respondents’ gender, age, level of education, occupation, internet experience, and familiarity with e-government portal. For example, Table IV shows that 51.77% of respondents are females, while 48.23% are males. Hence, Moreover, number of females responded was more than males. Moreover, 51.18% of respondents are between 18-29 years old.

In addition, 65.75% of respondents have bachelor degrees and most of them 59.25% are employees. Nearly half of the respondents 42.72% are very familiar with internet, and only 1.97% are not familiar with e-government.

To validate the hypotheses, different tests are conducted. These tests are: T-Test, One-way ANOVA, and Correlation. For example, to test the hypothesis H1a, correlation was used with Technical factors and perceived usefulness. The results of the Correlations are displayed in Table V. Correlation is significant at the 0.01 level (2-tailed). The value of Sig is .000 < 0.01. Therefore, based on the Correlation test, there is a strong positive relationship (0.761) between technical factors and usefulness, hence H1a is accepted.

To test the gender, T-Test is required. Table VI presents T-test for Gender. Based on the means results, there are no differences in means. Therefore, Gender has no influence on perceived usefulness.

To test the Age, a one-way ANOVA was used with Age and usefulness. The results of the ANOVA are displayed in Table VII and Table VIII. Sig value is .637 > .05 and there are no differences in means, therefore age has no influence on perceived usefulness.

Similar to Tables V to VIII, other tables are not shown here due to brevity. However, overall findings are summarized in Table IX. The table shows the hypotheses and the relationship strength between the factors. In addition, it shows whether the hypotheses are accepted or rejected.

TABLE. IV. RESPONDENTS’ PROFILE

Variable	**Percentage**
Gender	
Male	48.23%
Female	51.77%
Age	
18-29	51.18%
30-39	27.76%
40-49	13.78%
50-Above	7.28%
Education	
Secondary	10.04%
Diploma	16.93%
Bachelor	65.75%
Master and above	7.28%
Occupation	
Student	24.41%
Employee	59.25%
Business owner	3.35%
Retired	7.28%
Not Working	5.71%
Familiarity with internet	
Very Familiar	42.72%
Fairly Familiar	38.39%
Familiar	16.93%
Not Familiar	1.97%
Familiarity with e-government	
Very Familiar	10.04%
Fairly Familiar	27.56%
Familiar	37.60%
Not Familiar	24.80%
Use e-government portal	
Yes	53.15%
No	46.85%

TABLE. V. CORRELATIONS BETWEEN TECHNICAL FACTORS AND PERCEIVED USEFULNESS

Correlations			
		Technical Factors	Perceived Usefulness
Technical Factors	Pearson Correlation	1	.761**
	Sig. (2-tailed)		.000
	N	188	188
Perceived Usefulness	Pearson Correlation	.761**	1
	Sig. (2-tailed)	.000	
	N	188	188

** Correlation is significant at the 0.01 level (2-tailed).

TABLE. VI. T-TEST FOR GENDER

Demographic Factors	**N**	**Mean**	**T**	**df**	**P.value (sig2tailed)**
Gender					
Male	87	18.09	-.556	186	.579
Female	101	18.41	-.545	158.34	.587

The results show that all the hypotheses are accepted except for two, which are H2a and H2b. It represents that there is no influence of demographic factors on perceived usefulness and ease of use. Other hypotheses are accepted, which represent the influence between the factors within the hypotheses with different influence rates. For example, based on the Correlation test, there is a strong positive relationship (0.713) between technical factors and ease of use, therefore, H1b is accepted. Furthermore, based on the Correlation test, there is a strong relationship (.702) between perceived usefulness and intention to participate, thus, H6 is accepted. Although H4b is accepted, but there is a very weak relationship between political factors and ease of use.

TABLE. VII. ONE-WAY ANOVA FOR AGE AND PERCEIVED USEFULNESS

Perceived Usefulness	**Sum of Squares**	**Df**	**Mean Square**	**F**	**Sig**
Between Groups	27.129	3	9.043	.568	.637
Within Groups	2929.57	184	15.922	-	-
Total	2956.70	187	-	-	-

TABLE. VIII. DESCRIPTIVE TABLE FOR AGE

Perceived Usefulness	**N**	**Mean**	**Std. Dev.**	**95% Confidence Interval for Mean**	
				Lower Bound	Upper Bound
18-29	85	17.964	4.004	17.10	18.82
30-39	60	18.250	3.689	17.29	19.20
40-49	36	19.000	4.021	17.63	20.36
50 & above	7	18.285	5.964	12.76	23.80
Total	188	18.266	3.976	17.69	18.83

TABLE. IX. SUMMARY OF HYPOTHESES RESULTS

N	**Hypothesis**	**Results (Accept/Reject)**	**Relationship Strength**
H1a	Technical factors are positively related to perceived usefulness	Accept	Strong 0.761
H1b	Technical factors are positively related to perceived ease of use	Accept	Strong 0.713
H2a	Demographic factors have influence on perceived usefulness	Reject	No influence
H2b	Demographic factors have influence on perceived ease of use	Reject	No influence
H3a	Social influence is positively related to perceived usefulness	Accept	On average 0.491
H3b	Social influence is positively related to perceived ease of use	Accept	Midum 0.423
H4a	Political factors are positively related to perceived usefulness	Accept	Weak 0.279
H4b	Political factors are positively related to perceived ease of use	Accept	Weak 0.315
H5	Perceived quality is positively related to technical factors	Accept	Strong 0.731
H6	Perceived usefulness is positively related to e-participation level	Accept	Strong 0.702
H7	Perceived ease of use is positively related to e-participation level	Accept	On average 0.650

V. DISCUSSION

The research questions are highlighted together with their responses in the light of conducted study, as follows:

1) What are the factors affecting e-participation in Kuwait?

The research result shows that the following are the factors that influence the citizens and residents' e-participation in Kuwait e-government:

a) Technical Factors refer to the website design and content, channels of communication, and infrastructure. The technical factors have a strong positive relationship with perceived usefulness (PU) and perceived ease of use (PEU).

b) Social Influence is the persuasive influence people have on one another. It has an average relationship on PU and PEU.

c) Political Factors are related to the level of trust in government, government commitment and the political party affairs. The research found that political factors have a weak relationship on PU and PEU.

2) Does Kuwait have a low level of e-participation in e-government portal and Why?

As stated in the questionnaire, Kuwait has a low level of e-participation. Only 53.15% of the respondents used e-government portal, but the sample size should have to be bigger to get better results.

3) How to attract citizens to participate in e-government portal?

Attracting citizens to use the e-government portal is essential. In the questionnaire, citizens mentioned some of the reasons why they did not use the portal. Most of the reasons mentioned are about the website contents, privacy, availability of services, and information the portal provide. This means that the government must communicate with citizens, know their requirements, and reflect accordingly the changes and improvements in the portal.

To make people use e-government portal, the services must be sincerely useful to the targeted users. They must be efficient and meet citizen's specific requirements. For an effective participation in e-government portal, attractive awareness campaigns must be launched directing potential users appropriately to notify them about the real benefits they would gain out of participating in e-government.

VI. Conclusion

Most governments provide online information and services to their citizens and residents which is very common in the world nowadays. This research study aimed to identify the critical factors that led to participate in e-government in Kuwait. The result showed that technical factors, political factors, social influence, perceived quality, perceived ease of use, and perceived usefulness influence citizens' intention to participate in e-government portal.

The results from the statistical analysis concluded that a large portion of people 46.85% do not use the e-government portal. It is important to mention that the results, in this study, are helpful for decision makers to understand the citizen`s needs and requirements. The proposed research model has proved to be a useful guideline that would support e-government strategy in Kuwait.

Although the developed model and its implementation are effective, yet there is a room to extend the study further. As a future work, one can consider the followings: Using a larger population sample; extend the number of factors in the model to include, for example, security and trust; and study the factors influencing each level of e-participation framework, which are e-information, e-consultation, and e-decision-making.

References

[1] Digitales.oesterreich.gv.at., "What is e-Government?", 2015. Retrieved Oct 2017, from Digital Austria: REF: http://www.digitales.oesterreich. gv.at/ site/6506/default.aspx.

[2] Russell, T. Electronic Government Barriers and Benefits as Perceived by Citizens Who Use Public. Walden University, 2013.

[3] West, D. M. "E-government and the Transformation of Service Delivery and Citizen Attitudes", Public Administration Review, Vol. 64, No. 1, pp. 15-27, 2004.

[4] Palvia, S. C. J. and Sharma, S. S., "E-government and E-governance: Definitions/Domain Framework and Status around the World". Foundations of E-government, International Conference on E-governance, USA: Computer Society of India, 2007.

[5] Peristeras, V., Mentzas, G., Tarabanis, K. A., and Abecker, A. "Transforming E-government and E-participation through IT", Guest Editors' Introduction, IEEE Intelligent Systems, September/October, 2009, pp. 14-19.

[6] Smith S. and Dalakiouridou, E. "Contextualising Public (e)Participation in the Governance of the European Union", European Journal of ePractice, No. 7, pp. 1-9, 2009.

[7] UN E-government Survey in the News, (n.d.). Retrieved from The United Nations: http://unpan3.un.org/egovkb/egovernment_overview/ eparticipation.htm, 2018.

[8] Loulis, E., Macintosh, A., and Charalabidis, Y., "E-participation in Southern Europe and the Balkans: Issues of Democracy and Participation Via Electronic Media", Routledge, 2013

[9] Delakorda, S., "E-participation". e-Democracy Conference. Institute for Electronic Participation, 2011.

[10] Ekelin, A., The Work to Make eParticipation Work. Doctoral Dissertation Thesis, Blekinge Institute of Technology, SWEDEN, 2007.

[11] Macintosh, A., "Characterizing E-participation in Policy-Making". Proceedings of the 37th Hawaii International Conference on System Sciences, pp. 1-10, 2004.

[12] Lamrabat, A. and Jiang, N., "E-participation: Empowering People through Information Communication Technologies (ICTs)", Aide-Mémoire, Expert Group Meeting, United Nations Department of Economic and Social Affairs Division for Social Policy and Development, 2013.

[13] Millard, J., Nielsen, M. M., Waren, R., Smith, S., Macintosh, A., "European eParticipation Summary Report". European Commission, pp. 7-12, 2009.

[14] UN "E-government Survey, 2018" E-government for the Future We Want, Department of Economic and Social Affairs, United Nations, New York, 2018.

[15] Reddick, C. G., "Comparing citizens' use of e-government to alternative service channels", International Journal of Electronic Government Research (IJEGR), Vol. 6, No. 2, pp. 54-67, 2009.

[16] Daniele M. Nascimento, "Information flows in e-participation applications implications in government service-delivery in Brazil", International Conference on Information Society (i-Society), Pages: 51 – 52, 2016.

[17] UN, "Singapore's Experience: Initiatives to promote e-participation", INFOCOMM Development Authority of Singapore, E-participation: Empowering People through Information Communication Technologies (ICTs), 24-25 July 2013.

[18] Bwalya, K. J. (2009) "Factors affecting adoption of e-government in Zambia," The Electronic Journal of Information Systems in Developing Countries, Vol. 38, pp. 1-13.

[19] Khoirunnida; A. Nizar Hidayanto; Betty Purwandari; Riski Yuliansyah; Meidi Kosandi, "Factors influencing citizen's intention to participate in e-participation: Integrating Technology Readiness on Social Cognitive Theory", ICIC Second International Conference on Informatics and Computing, Pages: 1 – 7, 2017.

[20] Stoiciu, A., "Strengthening e-participation: Overcoming Obstacles for Better Consultation Mechanisms". E-participation Empowering People Through ICTs. Geneva : IMDD, 2013.

[21] Ali, H. and Ali, T., "E-participation: an investigation of Government Readiness in the Kingdom of Bahrain", Journal of e-government Studies and Best Practices, IBIMA Publishing, (2015), pp. 1-13, 2015.

[22] Bernd W. Wirtz, Peter Daiser & Boris Binkowska, E-participation: A Strategic Framework, International Journal of Public Administration, 41:1, 1-12, 2018. DOI: 10.1080/01900692.2016.1242620.

[23] Colesca, S. E., and Dobrica, L., "Adoption and Use of e-government Services: The Case of Romania", Journal of Applied Research and Technology. Vol. 6. pp.204-217, 2008.

[24] Ahmad, M. O., Markkula, J., Oivo, M. (2013) “Factors affecting e-government adoption in Pakistan: a citizen's perspective”, Transforming Government: People, Process and Policy, Vol. 7, No. 2, pp. 225-239, 2013.

[25] AlAwadhi S. and Morris, A., “Citizen Awareness to e-government Services for Information Personalization”, Journal of Software, Vol 4, No 6, pp. 584-590, 2009.

[26] Aljazzaf, Zainab M.. “Evaluating Trust in E-government: The case of Kuwait.”5th International Conference on Computer and Technology Applications (ICCTA) 2019 . DOI:10.1145/3323933.3324073.

APPENDIX 1

Table X presents the research questionnaire. The questionnaire presents the factors and their criteria.

TABLE. X. RESEARCH QUESTIONNAIRE

1. Perceived Quality
The information in the portal is useful
The information in the portal is up to date
The information in the portal is easy
The information in the portal is accurate
The portal provides information in an adequate level of detail
The interaction with the portal is clear and comprehensive
The design is appropriate for the type of the website for e-government
The portal provides a positive experience
It is safer to complete the transactions on the portal
I feel that my personal information is secure on the portal
2. Political Factors
I use services in the e-government portal because I trust the government
I use e-government portal because my political party allows me to do
Lack of implementation of policies leads to low level of e-participation
Changes in the local political events changed my view about e-participation
3. Technical Factors
Registration in the portal is easy
Using the e-government portal is easy
The e-government portal involves many important services
The services provided online make the procedures easier and simple
The government services provided online are of high quality
Weak internet in Kuwait prevents citizens from the portal usage
4. Social influence
I suggest using the e-government portal to others
I use the e-government portal because my family uses it
I use the e-government portal because my friends use it
I use the e-government portal because important people such as celebrities use it
I use the e-government portal because my work colleagues use it
5. Perceived Usefulness
The portal enables me to accomplish tasks more quickly
The portal enhances my effectiveness on doing my job
I can easily search and navigate in the e-government portal
I find the suitable help I expect
Access to the services is provided for citizen, resident, & person with disabilities
6. Perceived Ease of Use
I rarely make errors when using the portal
I rarely become confused when using the portal
I found the e-government portal flexible to interact with
Overall, I found e-government portal ease to use
My interface with e-government portal was clear and understandable
7. Intention to participate
I intend to use e-government portal because I found it useful
I intend to use e-government portal because I found it easy
I intend to use e-government portal in the future

Dynamic Changes of Multiple Sclerosis Lesions on T2-FLAIR MRI using Digital Image Processing

Marwan A. A. Hamid[1*], Walid Alhaidari[2], Waseemullah[3], Najeed Ahmed Khan[4], Bilal Ahmed Usmani[5], Syed. M. Wasim Raza[6]
Department of Biomedical Engineering[1, 2, 5,6]
Computer Science. and Information[3 ,4]
NED University of Engineering andTechnology, Karachi, Pakistan[1, 3, 4, 5,6]
Vladimir State University, Vladimir, Russia[2]

***Abstract*—Multiple Sclerosis (MS) is a complex autoimmune neurological disease affecting the myelin sheath of the nerve system. In the world, there are about 2.5 million patients with MS, in South and East Asia the ratio of MS is high. This disease affects young and middle-aged people. The MS is a fatal disease, and the numbers and volumes of MS lesions can be used to determine the degree of disease severity and track its progression. The detection of multiple sclerosis is a critical problem in MRI images because MS is described as frequently involves lesions, it can be appeared on a scan at one time-point and not appeared in subsequent time points. Also, MS on the T2 FLAIR MRI image is more often manifested by the presence of focal changes in the substance of the brain and spinal cord, which complicate their dynamic control according to MRI data. The detection and extraction of the MS lesions features are not just a tedious and time-consuming process, but also required experts and trained physicians, so the computer-aided tools become very important to overcome these obstacles. In this paper, we present a novel computer-aided approach based on digital image processing methods for enhancing the structures, removing undesired signals, segmenting the MS lesions from the background, and finally measuring the size of MS lesions to provide information about the current status of MS, which represent MS lesions that are either new, increasing or shrinking. The accuracy of the proposed methodology was 96%, according to the results presented in data. The lack of accuracy is related to some errors in segmentation.**

Keywords—Multiple sclerosis; T2-FLAIR; magnetic resonance imaging; digital image processing; image segmentation

I. Introduction

Multiple Sclerosis (MS) is a chronic autoimmune disease in which the myelin sheath in the fibers of the cerebral nerve and spinal cord is affected by demyelination. The demyelination is a formation of foci MS which destroys the myelin of the white matter of the brain and spinal cord. MS disease targets young and middle-aged people. The absence of timely diagnosis and treatment leads to disability and loss of working capacity due to the simultaneous damages of various parts of the nerve system. Patients with MS have multiples symptoms such as weakness [1], fatigue, blurred vision, and bladder symptoms. The exact cause of the disease (MS) is still unknown; some interested researchers connect the disease with genetic and environmental causes[2].

In the world, there are about 2.5 million patients with MS [3]. In some regions of Europe, the incidence of multiple sclerosis is quite high and is in the range of 8-7 cases per million [4]. In large industrial areas and cities, it is higher. According to epidemiological studies and research in the field of diagnosis and treatment of demyelinating [5] diseases, South and East Asia are in a zone of a high probability of demyelinating diseases.

The diagnostic by a neurologist allows evaluating the clinical manifestations of the disease. Magnetic resonance imaging (MRI) is used to visualize the location of MS and assess the morphological features of the affected areas [6]. The characteristics of metabolic processes of foci multiple sclerosis are investigated using positron emission tomography (PET)[7]. It is important to note that neurological sometimes cannot diagnosis the MS; this condition is called a radiologically isolated syndrome. In this case, the effectiveness of treatment and diagnosis is determined only by MRI.

Generally, MRI is a non-invasive medical imaging technique that comprises of a strong magnetic field, radiofrequency wave, and computer. The output of this technique is high-quality images of anatomical structures of the brain. The most common MRI modes are T1-weighted, T2-weighted, and T2-Flair images[8]. Fig. 1 shows the brain tumor on the T2-FLAIR and T2- weighted modes.

In this study, the T2-FLAIR image is used because MS lesion has vague boundaries and low contrast in T1-weighted and T-2 weighted [9]. Also, T2-FLAIR shows MS lesion brighter than other modes[10]. MRI images are characterized by the presence of random noises and fuzzy boundaries in the process of their formation. Moreover, T2-FLAIR is a very complicated biomedical object for analysis because it is achieved with the help of special procedures and equipment to visualize real biological objects that have certain properties that make their analysis difficult.

Multiple sclerosis on T2-FLAIR [11] is more often manifested by the presence of multiple changes in the substance of the brain and spinal cord, which complicates their dynamic control according to MRI data. The absence of timely diagnosis and treatment leads to disability and loss of working capacity due to the simultaneous damages of various parts of the nerve system. The location and volume of MS lesions are important in determining the degree of treatment progress [12].

*Corresponding Author.

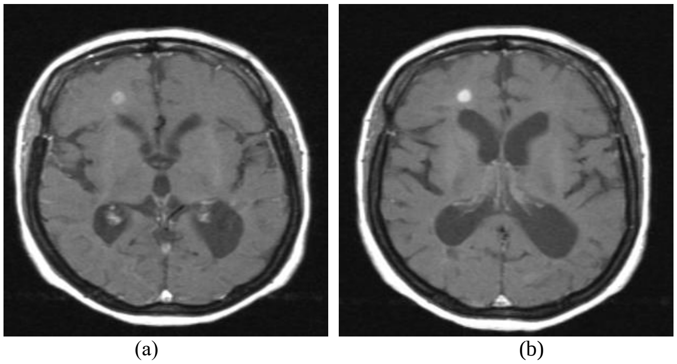


Fig. 1. Brian Tumor on different MRI modes a) T2-Weighted, b) T2-FLAIR.

The Radiologists and clinical experts manually measure the parameters of the MS lesions, and the accuracy of their results depend on their training and experience. The detection and extraction of the foci MS features are not just a tedious and time-consuming process, but also required experts and trained physicians, so the computer-aided tools become very important to overcome these obstacles. All published MS computer-aided approaches were applied on different data, mostly not calibrated, and their outputs were usually not directly comparable, making difficult the choice of the most effective method adopted to a clinical application. The MS is a fatal disease, and the information about MS lesions can be used to determine the degree of disease severity and track its progression. In this paper, we proposed a computer-aided tool that provides information about the current status of MS lesions (such as shrinking, growing, or presenting of a new one) using digital image processing techniques.

The research paper is organized as follows: Section 2 presents the critical analysis and comparison of recent related literature with our research work, Section 3 describes the proposed methodology and the stepwise illustrations, Section 4 outlines the result and discussion, and at last Section 5 concludes the paper with brief recommendations for future work.

II. Literature Review

Medical imaging is a non-intrusive technique applied in the medical field where the internal organs can be viewed without opening up of human body surgically. An MRI is a technique of medical imaging that produces images by an absorbed and emitted radio frequency signal from the human body [13]. In clinical practice, T1-weighted and T2-weighted are mostly used, but there are also other modes such as T2 FLAIR, which used to suppress the signal of the cerebrospinal fluid to get a better visualization of various multiple sclerosis [14]. Despite that, MRI has superior quality as a method for the clinical diagnosis of MS lesions, as well as understanding the size and location of the diseases.

The process of MRI image formation and storing interferes with random signals and noises that disrupt the intensity distribution of the image creating fuzzy boundaries. Different methods are being used for pre-processing MRI images. Generally, there are two techniques for noise reduction in medical images, the first is by increasing the acquisition (computational load and cost of the biomedical equipment) and the second one is by applying some processing techniques to remove or reduce the noises, which usually requires less acquisition time and can reduce the computational load. Gupta et al.[15] reviewed several linear and non-linear filtering algorithms for denoising digital images. In their study, the main goal was smoothing and enhancing the visual quality of the images. In their approach, the median filter was adopted due to flexibility and multiple-uses. It preserves most details and can remove most kinds of noises, such as impulse and Rician noises. Also, Kaur et al. [16] reviewed the noise characteristics in MRI images and applied different nonlinear filters to reduce Rician noise. As known, Noises are undesirable information [16] or random signals that cause damages by producing unreal boundaries, objects, and indistinct backgrounds[17], so the reduction of noises are mandatory in the medical image processing.

The uniformity of intensity distribution creates fuzzy boundaries, which lead to the need for filtration to remove the noises and segmentation to separate the lesions from the background of the images. Tanya et al. [18] improved a 3D segmentation approach using a convolutional neural network (CNN) to process four voxel-based uncertainties. Their results showed that filtering based on uncertainty greatly enhanced the accuracy of small MS lesions detection (around 40% of the dataset). Moreover, their result of segmentation provides clinicians or radiologists with information permitting them to assess whether to accept or reject lesions of high uncertainty quickly.

The location and volume of MS lesions are important in determining the degree of treatment progress. Ghribi et al.[19] suggested some recent segmentation (semi-automatic and automatic) methods. They gave a brief review of some directed methods to MS segmentation. However, no one of them can be regarded as a model approach. Many MS segmentation methods have been proposed in the last few years [20] to develop new techniques that give hopeful results.

In the MRI brain white mater images, segmenting of the MS is an essential process before therapy or surgical planning. Ameli et al. [21] presented a study of multiple sclerosis segmentation algorithms. In their study, they explored thirteen segmentation methods on the 53 MS cases. They gathered a database of 53 MS patients. The results of their study still trailing human expertise on both detection and delineating criteria.

Many multiple sclerosis segmentation methods based on intensity distributions such as thresholding method [22] and Mixtures of Gaussians can be implemented with the distribution of the intensity to differentiate between normal and infected tissues [23]. Recently, Schmid et al. [24] introduced a pipeline segmentation of FLAIR hyperintense white matter lesion changes between two points. Their method segmented significant changes in white matter lesions. The result of using their approach leads to more coherent results. The limitation of their work that they cautioned that their algorithm was validated using high-quality MRI data in the early stage of MS. Hence it may not work well with other situations.

Roy et al.[25] presented details of the longitudinal white matter lesion segmentation of MS challenge that was mentioned during the 2015 international symposium on biomedical imaging. Different lesion segmentation algorithms evaluated their submitted results. Their experiment shared a rich data set, collaborated and comprised of various avenues of research, reviewed the refinement of the evaluation metrics.

Over the past few years, the convolution neural network (CNN) gained a lot of interest in classifying MS lesions after the segmentation process. Roy et al. [26] proposed a CNN approach to separate the white matter lesions from multi-contrast MRI. Their approach divided into two paths: the first contains multiple parallel convolutional filters, and the second produces a thresholding function for binarization of images. Their approach scored a 90.48 in the International Symposium on Biomedical Imaging challenge.

Aslani et al. [27] introduced an automated method for segmenting MS lesions from multi-modal brain magnetic resonance images. Their approach based on a deep-end to end 2D CNN. They included a down-sampling path that can encode information from multiple modalities. The volume [28] of MS lesions can be used to determine the degree of disease severity and track its progression. Therefore, the segmentation of MS in white matter plays an essential role in understanding the nature of the dynamic behaviors of MS and helps to investigate the progression of the disease. The MS lesions are small in terms of size and indistinct borders. In T2 FLAIR may possess low resolution and often has imaging artifacts. Multiple sclerosis observed with high inter-rater variability according to Carass et al. [29].

The mentioned studies above have been organized a competition in MS lesion filtering, segmenting, and measuring. All multiple sclerosis approaches are evaluated on different datasets, mostly not determining the size and location of the lesions. Also, their results are making difficult the choice of the most relevant method adapted to clinical use.

Multiple sclerosis is described as frequently involves lesions. It can appear on a scan at one time-point and not appeared in subsequent time points. Determining the MS at one scan without reference to other scans may cause errors in the estimation of damaged tissue. The damaged tissues (MS lesions) have an indistinct correlation, and they change their location during treatments. Multiple sclerosis on MRI is characterized by the presence of multiple changes (shrinking or growing) in the substance of the brain and spinal cord, which complicates their dynamic control according to MRI data. Thus, there is an apparent need to investigate the dynamic changes of multiple sclerosis lesions on T2-FLAIR MRI using digital image processing. In this paper, we proposed a comparable methodology for overcoming the problem of dynamic changes in MS lesions by selecting the most relevant digital image processing methods to provide information about the current status of MS, which represent MS lesions that are either new, increasing or shrinking.

III. Proposed Methodology

In this study, around 38 scans before and after a month of treatment for each one of 55 MS cases from different real clinical cases were gathered and presented in DICOM format. For retrieving the images, neurologists, and neurosurgeons were consulted to differentiate between diseased and healthy (normal) images. Various processes were performed on the selected images based on a novel digital image processing methods for enhancing the structures, removing undesired signals, segmenting the MS from the background, and finally measuring the size of MS lesions to provide information about the current status of MS, which represent MS lesions that are either new, increasing or shrinking. Fig. 2 illustrates the stages of different processes of the proposed methodology

A. *Documentation*

In this paper, T2 FLAIR mode is used. The registration of images was performed on a Siemens magnetic resonance imaging machine with a field strength of 1.5 T.

B. *Pre-Processing Stage*

During formation and transformation in various devices, medical images can be distorted and degraded by different random signals or errors. The first step in digital image processing is the enhancement by eliminating noisy information.

- Image Enhancement

This stage is essential to distinguish the lesions and enhance the quality of images. In this section, spatial filters were used because we are dealing with additive noises, and we need to have direct results. Linear and non-linear filters such as median, Gaussian, and Laplacian smooth the images by averaging the value of pixels. For example, median filter denoise images from the impulse noise by determining the position of the impulse and replace it with median value while keeping other pixels of the image intact[30].

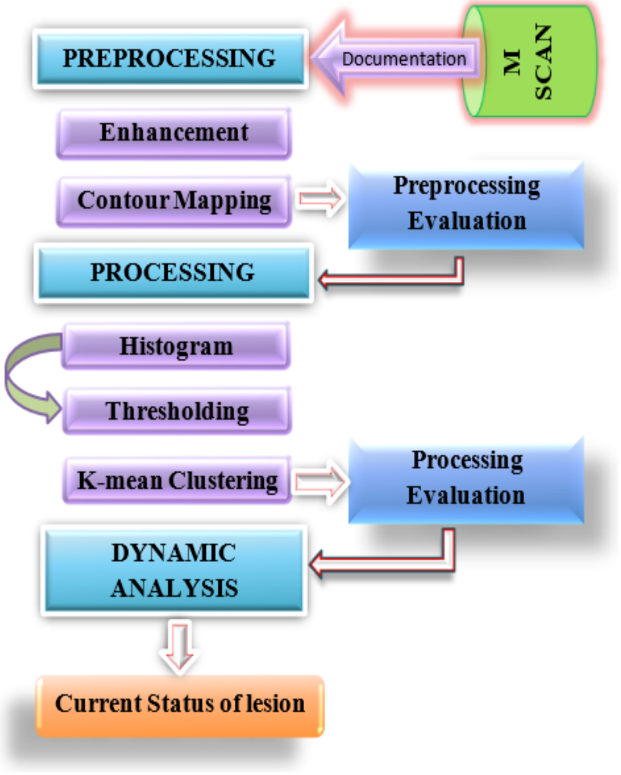


Fig. 2. Schematic Diagram of the Proposed Methodology.

- Contour Mapping

Contour is a graph that shows the 3D image in the 2D plane. It plots two variables, variable on the Y-axis and a response variable Z as contours [31]. In this section, we analyzed the results of the pre-processing stage by using a contour mapping technique to draw the intensity structure of MS lesions before and after the pre-processing stage. The goal of this step is to evaluate the enhancement of MS structures on the T2-FLAIR images.

C. Processing (Segmentation) Stage

Segmentation plays an important role in many medical diagnostics, and the result of segmentation influences the entire analysis. The following points contain detailed information related to the proposed segmentation methods.

- Histogram

In this section, the histogram was proposed to study the distribution of the image ingredients and also to select the thresholding value to separate the lesions from the normal tissues [32].

- Threshold Method

The meaning of segmentation in MRI medical images is dividing an image into a set of meaningful, homogenous, non-overlapping regions. One of the effective ways for separating the interest is the Threshold method [33].

$$I(x,y) = \begin{Bmatrix} 1, if\ f(x,y) > T \\ 0, if\ f(x,y) \leq T \end{Bmatrix} \quad (1)$$

The output of image I(x,y) depends on the T. The main logic in this method is the selection of a threshold value (T). The selection of threshold value (T) in this method was made using the histogram of the intensity distribution.

- K-means Clustering Method

Clustering is a technique for splitting the image into clusters of pixels. The pixels in each cluster must have similar attributes and vary from other clusters. In this method, there are not common pixels between clusters. In other words, one pixel can only belong to one cluster only[34]. K-means clustering algorithm (formula 2) has two phases: determining the number of clusters (k=5) is first, and the second one is taking each point of the cluster, which has the nearest centroid from the respective data point.

$$C_k = \frac{1}{k} \sum_{y \in C_k} \sum_{x \in C_k} p(x,y) \quad (2)$$

Where k is the number of selected clusters. Although k-means has a great advantage of being easy to implement, it has some drawbacks. The quality of the final clustering results depends on the arbitrary selection of initial centroid.

In this section, we examined the threshold and k-means clustering methods by comparing the size of MS lesions after segmentation with the results that done automatically (manually) by radiologists without segmentation. The goal of this section is to select the most optimal method to separate MS lesions.

D. Dynamic Analysis

In this section, a quantitative assessment of pathological changes was carried out by calculating [35] the size of the MS in each slice after and before the treatment to investigate and evaluate the dynamic changes (shrinkage or growth) of MS lesions. The expression that was used to calculate the size of the lesions after segmentation is:

$$S_o = \sum_{i=1}^{n} (N_p * h_p * w_p) \quad (3)$$

Where So is the calculated area of MS, Np is the number of segmented pixels on i slice; hp and wp are the height and width (pixel dimensions), respectively. We should fix the filtration and segmentation methods (do not change from image to image), to get accurate and useful results.

IV. Result and Discussion

A. Pre-Processing Results and Discussion

The basic ideas of the most enhancing filters (such as Laplacian or median) are based on the fact that dark and light pixels are found to be adjacent (next to each other) on the borders. The spatial filters were implemented to denoise and improve the quality of the images. Fig. 3 demonstrates that Laplacian (masks center -8) and Sobel filters highlight the edges of the objects, reinforce the boundaries between the light and dark pixels of the images. The Gaussian and Median remove unnecessary noises and smooth the borders on the images, which will reduce the lack of accuracy in the segmentation, shown in Fig. 4. Generally, the essential task of filters is smoothing, denoising, sharpening and emphasizing borders.

The output of this stage is the finer details of the images. Besides, the noises were reduced to the minimum degree from these images. Also, the MS lesions zones are clear and made it detectable as compared to raw images.

B. Contour Mapping

The contour method allows for mapping the intensity distribution. The numerical values represent the intensity values of images as groups of pixels corresponding to the normal and pathological brain tissues.

Fig. 5 presents the results of using the contour mapping method on the T2-FLAIR MS image. The results show that normal brain tissue has an intensity of pixel ranging from 250 to 300, and pathological has intensity higher than 350.

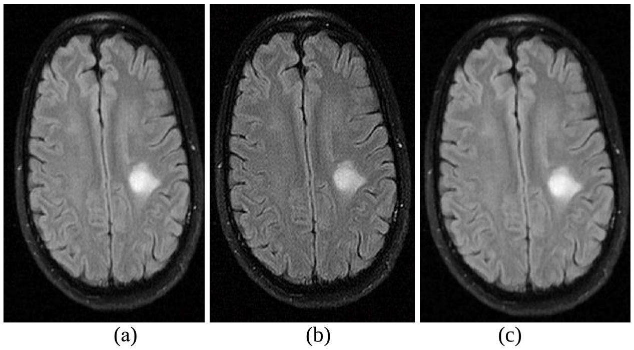


Fig. 3. Spatial Filters with T2 FLAIR MRI Images: (a) Original Image, (b) Image with Gaussian Filter, (c) Image with Median Filter.

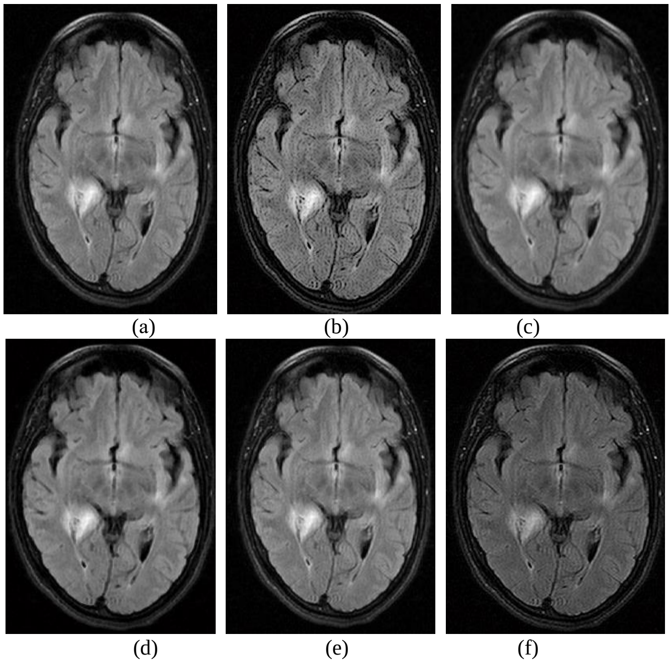


Fig. 4. Spatial Filters with T2 FLAIR MRI Images: (a) Original Image, (b) Image with Laplace Filter (-4 mask), (c) Image with Gaussian filter, (d) Image with Median Filter, e) Image with Laplace Filter (-8 Mask), (f) Image with Averaging Filter.

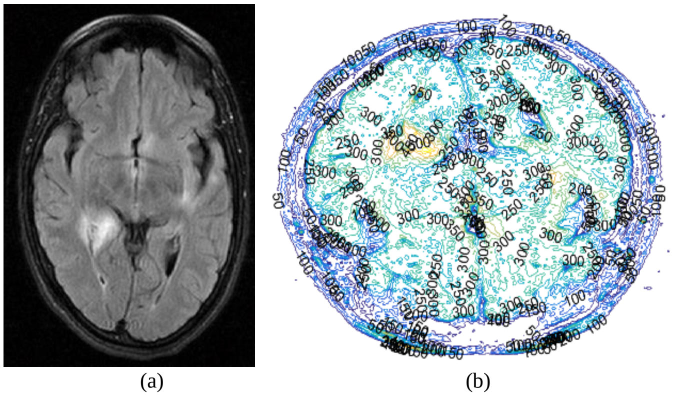


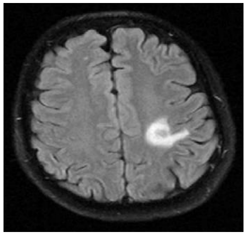
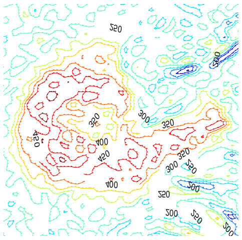
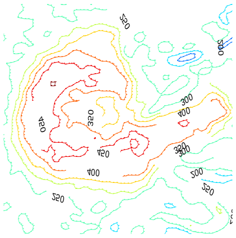
Fig. 5. Result of using Contour Mapping Method: (a) Original Image, (b) Contour of Image.

C. Results and Discussion of Pre-Processing Evaluation

In this section, we suggest the contour method to study the results of filtration on T2 FLAIR MS images and also to map the lesions. The purpose of this section is to evaluate the results of the pre-processing stage on MS lesions structures and compare them with the structures of the same lesions before preprocessing.

Table I demonstrates that the structures of multiple sclerosis lesions after the pre-processing stage have noticeable values. Also, it can be judged that the MS lesions before and after having varied constructions. The MS lesion has after pre-processing observable constructions and clear intensity distributions. The use of filtering in this paper (with T2-FLAIR) was mandatory to give accurate and optimal results for the following steps.

TABLE. I. EVALUATION OF MS LESIONS BEFORE AND AFTER PRE-PROCESSING

Original image of a patient with Multiple sclerosis	
The changes in the structure of the MS lesion before and after pre-processing	
Before pre-processing	
After pre-processing	

D. Processing (Segmentation) Results and Discussion

The meaning of segmentation in medical imaging is dividing an image into a set of meaningful, homogenous, non-overlapping clusters (regions or classes). Each region has similar attributes (such as intensity, depth of pixels, or textures) and should be different from other regions.

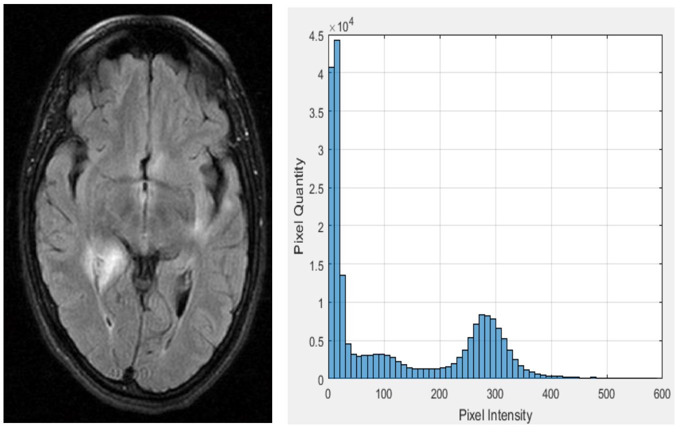
- Histogram

Histogram of an image provides a visual interpretation of the intensity distribution over an image. It can be used in many digital image processing applications, such as study the result of filtering techniques by tracking the intensity distribution over the image. Another use for the histogram is a binarizing image (segmentation) and determine the borders between different image components. In this section, we used the histogram method for the logical selection of thresholding segmentation values. Fig. 6 shows the histogram representation of normal and abnormal tissues.

The result of the histogram shows that the normal tissues have larger peaks than multiple sclerosis. The distribution of the MS tissues locates above 350. The histogram helped us easily to set the value of thresholding segmentation method.

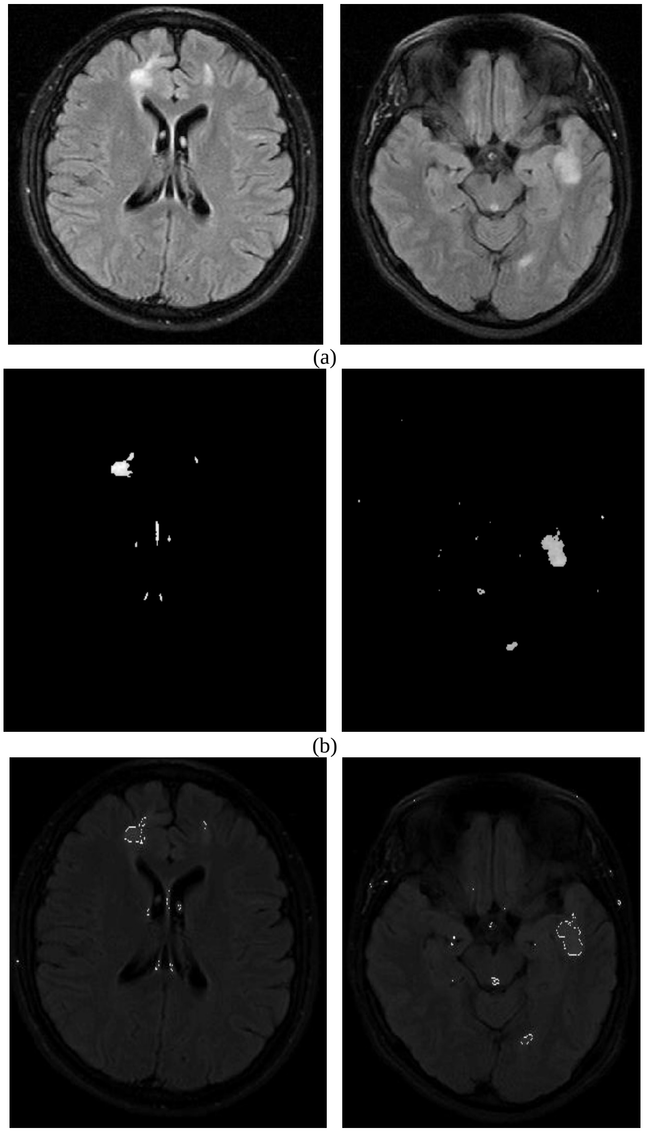
- Thresholding method

The histogram helped us easily to set the value of the thresholding segmentation method. The optimal thresholding segmentation value is T=350. Fig. 7 shows the result of segmentation using the thresholding method.



(a) (b)

Fig. 6. Histogram of the MRI Image (a) Original Image, (b) Histogram Representation of Normal and Abnormal Tissues,



(a)

(b)

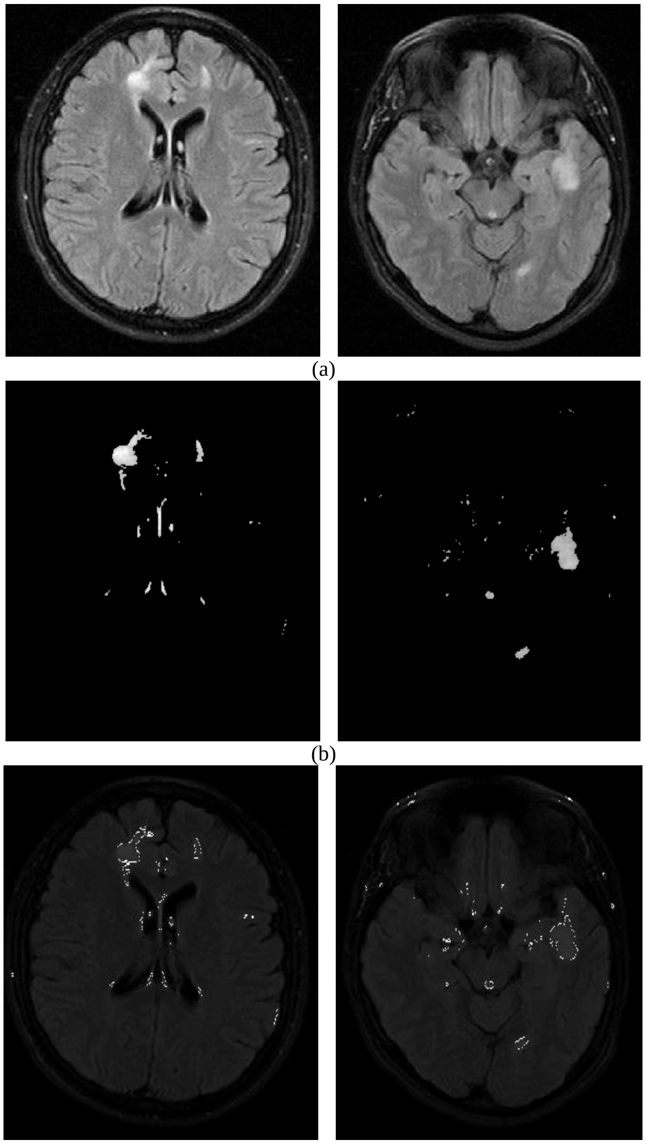
(c)

Fig. 7. Result of Thresholding Segmentation Method: (a) Original Image, (b) Segmented Image, (c) Mapping the Segmented Image with the Original Image.

- K-means clustering method

Clustering is a technique for splitting the image into clusters of pixels. The pixels in each cluster must have similar attributes and vary from other clusters. In this method, there are not common pixels between clusters, and we set the number of clusters equal to five (k=5). Fig. 8 shows the result of segmentation using the K-means clustering method.

All segmentation methods convert a given image into a grayscale image, then separate the infected tissues from healthy tissues by binarizing the image. The selection of the thresholding and k-means clustering methods was based on flexibility and popularity. Both methods successfully separated the MS lesions from the background.



(a)

(b)

(c)

Fig. 8. Result of K-means Clustering Segmentation Method: (a) Original Image, (b) Segmented Image, (c) Mapping the Segmented Image with the Original Image.

E. Results and Discussion of Processing Evaluation

The segmentation methods play a major role in determining the size and real boundaries of multiple sclerosis lesions. It can give accurate or sometimes misguided results. In this section, we evaluate the proposed segmenting methods to select a suitable method for the following steps. The accuracy and reliability of the results were confirmed by the data presented in the physician's (automatic evaluation by specialists) calculation.

TABLE. II. EVALUATION OF MS LESIONS SIZE WITH THRESHOLDING AND K-MEAN CLUSTERING SEGMENTATION METHODS

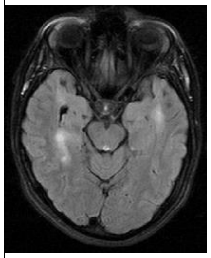
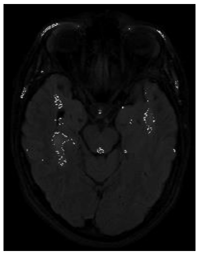
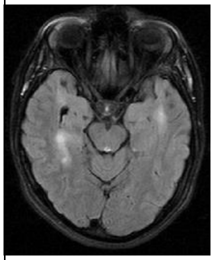
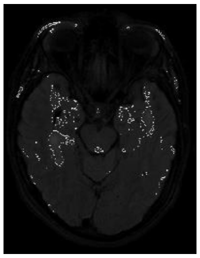
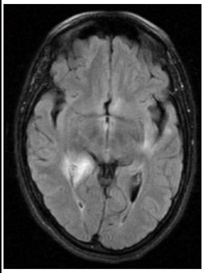
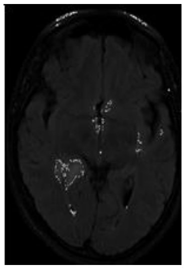
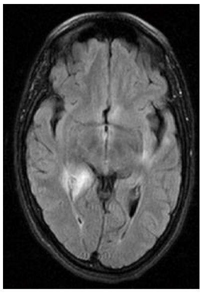
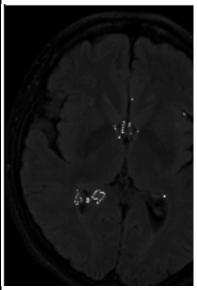
Method	**Original Image**	**Segmented Image**	**Size of MS,m m**
Patient no 1. is a woman, 43 years old, estimating MS lesion size was (2557,85) mm^2			
With threshold method			2214.72
With k-means clustering method			2135.52
Patient no 2. is a man, 27 years old, estimating MS lesion size was (1800,52) mm^2			
With thresholdmethod			1375.2
With k-means clustering method			1140.4

Table II presents the changes in MS lesion size with different segmentation methods and illustrates that the thresholding method approximately determined the boundaries of the MS area and gave a satisfactory result, whereas the k-means clustering method provided critical results. The separation of multiple sclerosis lesions from healthy tissues using segmentation methods is a real challenging task. The results of using thresholding and k-means clustering were optimistic, according to the presented data. In this paper, we selected thresholding as a method to segment the lesions because of the flexibility, accuracy, and capability of handling big dimensionality.

F. Results and Discussion of Dynamic Analysis of MS Lesions

Multiple sclerosis (MS) is described as frequently involves lesions. It can be appeared on a scan at one time-point and not appeared in subsequent time points, as shown in Fig. 9. Determining the MS at one scan without reference to other scans may cause errors in the estimation of damaged tissue. The damaged tissues (MS Lesions) have an indistinct correlation, and they change their location during treatments.

In this section, we investigate the accuracy of the implemented methodology using 38 scans divided equally as 19 scans before and 19 scans after a month of treatment for each one of 55 MS cases. Also, we fixed the filtering (Laplacian and Median) and segmenting (thresholding) methods for all scans. Then the obtained results were calibrated with the results of the automatic evaluation (done by specialists) of MS lesions that are either new, increasing, or shrinking. In the following, we present the evaluation results of two real clinical patients.

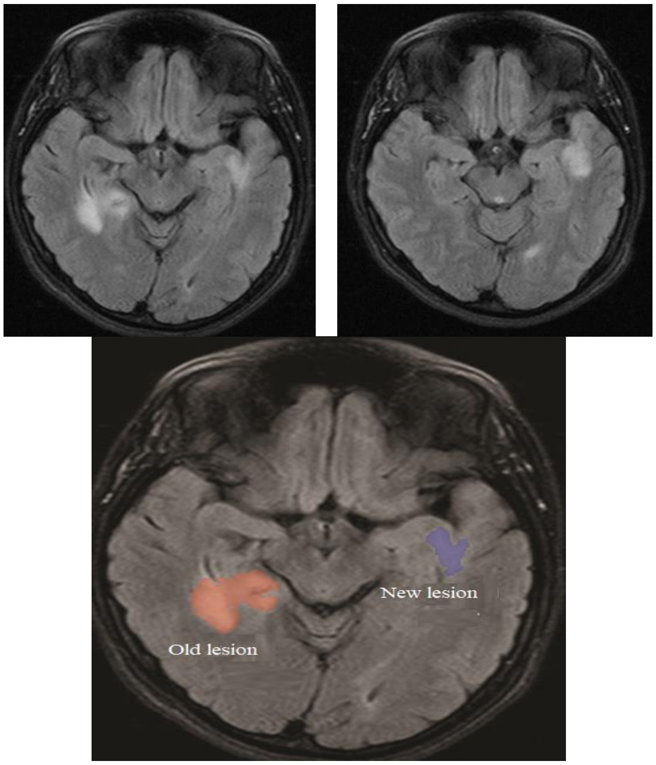


Fig. 9. Dynamic Changes of MS Lesions During Treatment.

- First Patient

The first case is a woman, 45 years old, the result of MRI scans show that she has most likely pseudotumor of the demyelinating disease in the left parietal lobe, as shown in Fig. 10.

The results of automatic evaluation by specialists before and after one month of treatment for the MS lesion size are presented in Table III, and the results of MS dynamic changes after and before treatment for the MS lesion size using our proposed methodology are shown in Table IV.

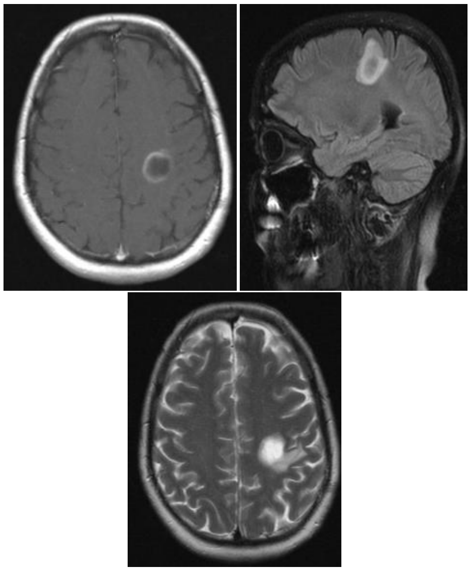


Fig. 10. MS Lesion in different Slices: Patient no.1.

TABLE. III. RESULTS OF THE AUTOMATIC EVALUATION OF DYNAMIC CHANGES OF MS LESIONS: PATIENT NO 1

Slice No.	**Size of MS lesion (mm^2)**	
	Before Treatment	*After One Month of Treatment*
1	0	0
2	0,10	0
3	0	0
4	0	0
5	1360,9	1209,3
6	0	0
7	10	0
8	0	0
9	1442,4	0
10	190,77	198,48
11	0	0
12	0	0
13	0,15	0
14	0	0
15	611,32	830,8
16	0	0
17	0	0
18	8	0
19	1510,8	1050,8
Total area, S_o	5116,4	3289,4

TABLE. IV. RESULTS OF DYNAMIC MS LESIONS CHANGES USING THE PROPOSED METHODOLOGY: PATIENT NO.1

Slice No.	***Size of MS lesion (mm^2)***	
	Before Treatment	*After One Month of Treatment*
1	0	0
2	0	0
3	0	0
4	0	0
5	1460,9	1169,3
6	0	0
7	0	0
8	0	0
9	1472,6	0
10	190,08	204,8
11	0	0
12	0	0
13	0	0
14	0	0
15	598,32	838,8
16	0	0
17	0	0
18	0	0
19	1491,8	1000,8
Total area, S_o	5213,52	3213,26

- Second Patient

The second case is a woman, 32 years old, the results of her MRI scans illustrate that she has an acute and chronic demyelinating disease, as shown in Fig. 11.

The results of automatic evaluation by specialists before and after one month of treatment for the MS lesion size are presented in Table V, and the results of MS dynamic changes after and before treatment for the MS lesion size using our proposed methodology are shown in Table VI.

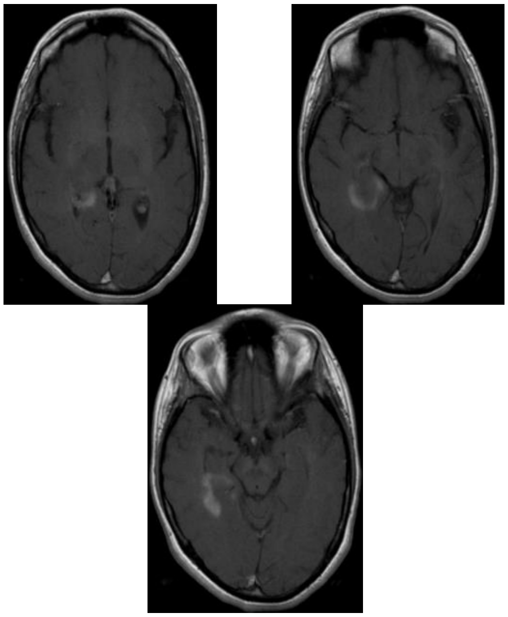


Fig. 11. MS Lesion in different Slices: Patient no.2.

TABLE. V. RESULTS OF THE AUTOMATIC EVALUATION OF DYNAMIC CHANGES OF MS LESIONS: PATIENT NO 2

Slice No.	**Size of MS (mm^2)**	
	Before Treatment	*After One Month of Treatment*
1	0	0
2	838,8	221,04
3	854,64	1854
4	0	129,6
5	0	0
6	0	0
7	743,04	1386
8	800,64	322,56
9	185,04	52,56
10	190,77	198,48
11	0	0
12	300,24	749,52
13	1748,2	676,8
14	517,68	175,68
15	0	0
16	0	0
17	1218,2	1864,1
18	414,72	713,52
19	152,64	5,06
Total area, S_o	7773,84	8150,4

TABLE. VI. RESULTS OF DYNAMIC MS LESIONS CHANGES USING THE PROPOSED METHODOLOGY: PATIENT NO 2

Slice No.	**Size of MS lesions (mm^2)**	
	Before Treatment	*After One Month of Treatment*
1	0	0
2	738,6	221,04
3	554,54	1454
4	0	100,6
5	0	0
6	0	0
7	641,03	1289
8	602,55	212,65
9	135,4	46,51
10	0	0
11	0	0
12	199,2	678,47
13	1600,34	800,89
14	420,47	578,6
15	0	0
16	0	0
17	1122,2	1746,02
18	512,71	811,24
19	255,24	6,02
Total area, S_o	7080.0	8237.0

All tables show appearing of new and disappearing of old multiple sclerosis lesions after or before the treatment. For example, slices number 2 (patient no. 1) and slice number 4 (patient no. 2). Atrophy and growth of the multiple sclerosis lesions were presented in the results of the proposed methodology as the slices number 5, 7, 9, 10, 13, 15, and 19 in Table IV. Also, the slices number 2,3,7,8,9 ,12,13,14,17,18 and 19 in Table VI. The results of the proposed methodology (Table IV and Table VI) were calibrated with the results of the automatic evaluation (Table III and Table V) for the same slices in both cases. The result of calibration shows that the errors between automatic evaluation and proposed methodology for patient no. 1 before and after treatment are nearly 1.89% and 2.31%, respectively. The patient no. 2 had 8.9% before treatment and 1.06% after one month of treatment.

In this paper, the proposed methodology used T2-FLAIR MRI scans to measure and investigate MS lesion shrinking, growing, and appearing of a new one in 55 MS cases. The result shows that shrinkage of lesions presented in 29 cases, whereas 19 cases showed more severe growth in old lesions. Also, new lesions presented in 7 cases. The accuracy of the proposed methodology was 96%, according to the results presented in data. The lack of accuracy is related to the errors of filtration and segmentation.

V. CONCLUSION

The selected MRI images were T2-FLAIR because these types of images show distinct boundaries of multiple sclerosis and have acceptable contrast. In this paper, the use of spatial filters was mandatory to detect and enhance the visuality of MS lesions. The results of spatial filters were tested using the contour method. The results of the contour method showed that MS lesions after filtration have observable construction and clear intensity distribution. The segmentation methods play a major role in determining the size and real boundaries of multiple sclerosis lesions. It can give accurate or sometimes misguided results. The thresholding method (with T=350) approximately determined the boundaries of the MS area and gave a satisfactory result. Whereas the k-means clustering method with (k=5) provided critical results. In this paper, we selected thresholding as a method to segment the lesions because of the flexibility, accuracy, and capability of handling big dimensionality. The size of MS lesions was measured using the quantitative method after fixing the filtration and segmentation methods. The evaluation of the proposed methodology was done using T2 FLAIR MRI scans to measure and investigate MS lesion shrinking, growing, and appearing of a new one in 55 MS cases. The result shows that shrinkage of lesions presented in 29 cases, whereas 19 cases showed more severe growth in old lesions. Also, new lesions presented in 7 cases. The accuracy of the proposed methodology was 96 %, according to the results presented in data. The lack of accuracy is related to some errors in segmentation. The proposed method can be used for filtering, segmenting, and tracking the information about the current status of MS, which represent MS lesions that are either new, increasing or shrinking. The improvement of accuracy will be better with more data. This approach can be used in diagnostics room for automatic decisions in critical multiple sclerosis cases.

VI. CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

References

[1] J. Snoek, “Fatigue , sleep disturbances and circadian rhythm in multiple sclerosis,” June 2014, 1993.

[2] J. Imitola, “New age for progressive multiple sclerosis,” vol. 116,. 18, pp. 8646–8648, 2019.

[3] D. A. Rudko, “Magnetic Resonance Imaging of Multiple Sclerosis -,”. February, 2018.

[4] N. Raza, M. Waheed, and R. Mehboob, “Multiple sclerosis lesions on magnetic resonance imaging , their characterization and pathological correlation with musculoskeletal disability in Pakistanis,” vol. 10, no. 6, pp. 567–574, 2016.

[5] S. Eskandarieh, P. Heydarpour, A. Minagar, and M. A. Sahraian, “Multiple Sclerosis Epidemiology in East Asia , South East Asia and South Asia : A Systematic Review,”. September 2017, 2016.

[6] S. K. Sah et al., “SM Gr up Multiple Sclerosis : Conventional and Advanced,” 2016.

[7] M. Petracca, M. Margoni, and G. Bommarito, “Monitoring Progressive Multiple Sclerosis with Novel Imaging Techniques,” Neurol. Ther., vol. 7, no. 2, pp. 265–285, 2018.

[8] D. A. Pollacco, “Magnetic Resonance Imaging,”. April, 2017.

[9] F. Martins, L. M. De Le, M. C. Pinho, N. M. Rofsky, and A. D. Sherry, “Basic MR Relaxation Mechanisms and Contrast Agent Design,” pp. 545–565, 2015.

[10] B. Jeevanandham, “To compare post contrast 3D T2 FLAIR , T1-SPACE and MP-RAGE sequences to select the ideal sequence for leptomeningeal abnormalities at 3 T MRI,” no. September, 2017.

[11] S. B. Vos et al., “Evaluation of prospective motion correction of high-resolution 3D-T2-FLAIR acquisitions in epilepsy patients ☆,” J. Neuroradiol., vol. 45, no. 6, pp. 368–373, 2018.

[12] M. I. Vargas, V. Garibotto, Þ. M. Viallon, and R. Guignard, “Clinical Applications of Hybrid PET / MRI in Neuroimaging,”. December 2017, pp. 0–3, 2013.

[13] L. Wald, “MR Image Encoding,” 2006.

[14] I. Blystad et al., “Synthetic MRI of the brain in a clinical setting Synthetic Mri of the Brain in a Clinical Setting,” vol. 53,. 10, pp. 1158–1163, 2012.

[15] G. Gupta, “Image Filtering Algorithms and Techniques : A Review International Journal of Advanced Research in Image Filtering Algorithms and Techniques : A Review,”. October 2013, pp. 1–6, 2018.

[16] M. Kaur and R. Kaur, “A Comparative Analysis of Noise Reduction Filters in MRI Images,” pp. 238–242, 2016.

[17] T. Julliand, V. Nozick, and H. Talbot, “Image Noise and Digital Image Forensics,”. April 2018, 2016.

[18] D. L. A. A. Tanya Nair, Doina Precup, “Exploring Uncertainty Measures in Deep and Segmentation,”

[19] O. Ghribi, I. Njeh, W. Zouch, and C. Mhiri, “Brief review of multiple sclerosis lesions segmentation methods on conventional magnetic resonance imaging,”. March, 2014.

[20] D. García-lorenzo et al., “multiple sclerosis white matter lesions on Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging,”. October, 2015.

[21] R. Ameli, “Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure,”. July, 2018.

[22] S. Jain, “Automatic segmentation and volumetry of multiple sclerosis brain lesions from,”. May, 2015.

[23] P. G. L. Freire and R. J. Ferrari, “Author ’ s Accepted Manuscript,” Comput. Biol. Med.,. April, 2016.

[24] P. Schmidt et al., “NeuroImage : Clinical Automated segmentation of changes in FLAIR-hyperintense white matter lesions in multiple sclerosis on serial magnetic resonance imaging,” NeuroImage Clin., vol. 23,. March, p. 101849, 2019.

[25] S. Roy et al., “Longitudinal multiple sclerosis lesion segmentation : Resource and challenge NeuroImage Longitudinal multiple sclerosis lesion segmentation : Resource and challenge,”. January, 2017.

[26] S. Roy, H. M. J. Foundation, and J. A. Butman, “Multiple Sclerosis Lesion Segmentation from Brain MRI via Fully Multiple Sclerosis Lesion Segmentation from Brain MRI via Fully Convolutional,”. July 2019, 2018.

[27] S. Aslani et al., “Multi-branch Convolutional Neural Network for Multiple Sclerosis Lesion Segmentation,” 2019.

[28] P. Sati et al., “CONSENSUS,” Nat. Publ. Gr.,. November, 2016.

[29] A. Carass et al., “Data in Brief Longitudinal multiple sclerosis lesion segmentation data resource,” Data Br., vol. 12, pp. 346–350, 2017.

[30] M. D. Sontakke and M. S. Kulkarni, “Different Types of Noises in Images and Noise Removing Technique,” no. 03, pp. 102–115, 2015.

[31] A. Burykin, M. Costa, B. Israel, D. Medical, L. Citi, and A. Goldberger, “Dynamical density delay maps : simple , new method for visualising the behaviour of complex systems,” no. January, 2014.

[32] N. Senthilkumaran and J. Thimmiaraja, “A Study on Histogram Equalization for MRI Brain Image Enhancement,” 2014.

[33] K. Bhargavi, “A Survey on Threshold Based Segmentation Technique in Image Processing,” vol. 3, no. 12, pp. 234–239, 2014.

[34] N. Dhanachandra, K. Manglem, and Y. J. Chanu, “Image Segmentation Using K -means Clustering Algorithm and Subtractive Image Segmentation using K -means Clustering Algorithm and Subtractive Clustering Algorithm,” Procedia - Procedia Comput. Sci., vol. 54,. December, pp. 764–771, 2016.

[35] I. I. T. Bombay and F. Iete, “Detection of Shapes of Objects Using Sophisticated Image Processing Techniques,” vol. 1, no. 4, pp. 32–37, 2010.

Scientific VS Non-Scientific Citation Annotational Complexity Analysis using Machine Learning Classifiers

Hassan Raza[1], M. Faizan[2], Naeem Akhtar[3], Ayesha Abbas[4], Naveed-Ul-Hassan[5]
School of Computer Sciences
National College of Business Administration and Economics
Lahore, Pakistan

***Abstract*—This paper evaluates the citation sentences' annotation complexity of both scientific as well as non-scientific text related articles to find out major complexity reasons by performing sentiment analysis of scientific and non-scientific domain articles using our own developed corpora of these domains separately. For this research, we selected different data sources to prepare our corpora in order to perform sentimental analysis. After that, we have performed a manual annotation procedure to assign polarities using our defined annotation guidelines. We developed a classification system to check the quality of annotation work for both domains. From results, we have found that the scientific domain gave us more accurate results than the non-scientific domain. We have also explored the reasons for less accurate results and concluded that non-scientific text especially linguistics is of complex nature that leads to poor understanding and incorrect annotation.**

***Keywords*—Classification; machine learning; sentimental analysis; scientific citations; non- scientific citation**

I. Introduction

The popular research area in this era is sentiment analysis [14]. Researchers widely used different types of textual data to perform sentiment analysis. Every business and organization need their clients to review for the betterment of their products and services. To analyze the opinion, perception, mindset, and experience of the user is known as sentiment analysis. Judging the sentiments of citing paper' writer about cited paper is termed as sentiment analysis [17]. From the literature work, it has been identified that no work has been done on the problem of evaluating the annotation complexity of both scientific as well as non-scientific text related articles. To perform this work we are needed to prepare experimental data sets of both domains. To prepare scientific corpus we selected Elsevier Computer & Operations Research Journal and prepared a corpus consisted of 5161 citation sentences extracted from 262 research papers published in 2015-2019. On the other hand, we selected SJR Applied Linguistics Journal to prepare a nonscientific corpus consisted of 4989 citation sentences extracted from 250 research papers in 2015-2019. Different machine learning classification algorithms e.g. Naïve-Bayes (NB), Neural Network (NN), Support Vector Machine (SVM), Logistic Regression (LR), Gradient Boosting (GB), Decision Tree (DT), K-Nearest Neighbor (KNN), and Random Forest (RF) are implemented. Using evaluation metrics e.g. f-score, and accuracy score, the system' accuracy is evaluated and improved using different data processing features selection techniques e.g. Lemmatization, NGrams, Tokenization, Case Normalization, and Stop Words Removal.

II. Literature Review

The current state of the domain is analyzed by conducting a literature review in this research work. With the passage of time, researchers' interest has been aroused towards sentiment analysis. The major attention of this domain is towards the construction of framework, extraction of features, and determination of polarities.

Mainly supervised and unsupervised learning approaches are used for sentiment analysis [10]. In a supervised learning mechanism, classifiers' training needs annotated data. To prepare annotated data we need some annotation guidelines. Labeled data is beneficial for a supervised learning approach. Classifiers are trained by this labeled data and also testing of classifier's accuracy is performed. Another approach is unsupervised learning, in this approach data doesn't need to be labeled while there is a need for sentiment lexicons and considered as difficult as it needs various types of lexicons for various genres.

Sentiments are often not well expressed in scientific citation [3]. This may be due to the overall strategy of avoiding critique because of the citation's sociological aspect [12]. [25] mentioned that many works of "politeness, nationalism, or piety" are cited. Negative feelings, still available as well as observable to humans, are articulated in intricate positions and maybe suppressed, particularly when they cannot be explained quantitatively [9]. In scientific literature, citation sentences are often neutral in terms of opinion, either because they critically define an algorithm, strategy or technique, or because they favor a fact or argument [3]. [13] have worked on Sentiment Analysis of Roman Urdu. Most of the research works have been done on different subjects like "English", and "Chinese" etc. No work has been done on sentiment analysis of non-scientific literature because non-scientific literature is totally different from other literatures. Non-scientific citations are very difficult to understand for a non-linguistic person because most of the unfamiliar words are used. So we decided to go for the evaluation of both scientific as well as non-scientific articles' citation sentences' annotation complexity.

III. Methodology

Fig. 1 highlights the purposed methodology adopted in this research work. First of all, in order to analyze the citation sentences' annotation complexity of both scientific as well as non-scientific text related articles we prepared our own data sets of separate domains mentioned in section IV. As we are following the supervised learning approach so there is a need for labeled data sets. For labeling the data we developed some annotation guidelines mentioned in section 4(B) and performed the annotation procedure with the help of human annotators. The annotators classified the citation sentences into 3-classes positive, negative, and neutral. After data is completely labeled, we developed a classification system using python based library named "Sickit-Learn". Test Train Split method is used to divide the data randomly through 60 percent of training data and 40 percent of test data. Experiments are conducted in two phases. In the first phase, we just applied uni-gram, bi-gram, and tri-gram features on data and computed F-scores and Accuracy Scores. Additionally, to boost the quality of the evaluations, we applied different features selection techniques (punctuations and stop words removal, lemmatization, case normalization, etc.) along with n-grams and then computed the above-mentioned metrics again. The later approach helped out in minimizing noise and data complexity. In order to calculate average results, thirty iterations of each experiment were carried out and a total of six experiments were carried out. Finally, we have explored the reasons for less accurate results for non-scientific data classification and concluded that non-scientific text especially linguistics is of complex nature that leads to poor understanding and incorrect annotation process.

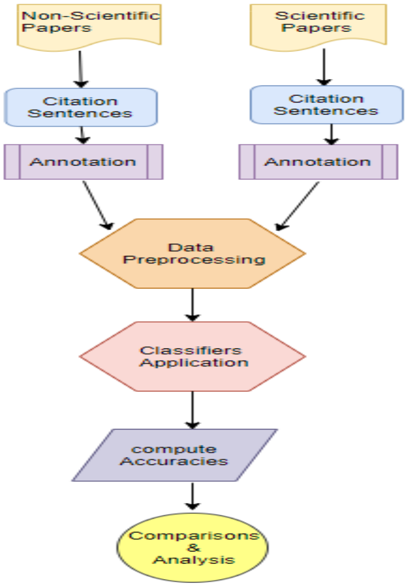


Fig. 1. Step by Step Process Working Stream.

IV. Corpus Construction

To evaluate the citation sentences' annotation complexity of both scientific as well as non-scientific text related articles we need corpora of these domains. We developed two different corpora. To prepare the scientific citations' corpus we choose a science-related journal named "Elsevier Computers and Operations Research Journal" and developed a data set consisted of 5161 citing sentences retrieved from 262 research articles published in 2015 – 2019. On the other hand to prepare non-scientific citations' corpus we specifically choose a non-scientific domain-related journal named "SJR Applied Linguistics" and extracted 4989 citation sentences from 250 linguistics research papers published in 2015-2019.

A. Citation Sentiment Annotations

After preparing the data set the next step was to label the data using a data annotation procedure. We executed this process by applying our own defined guidelines. Citation sentences are categorized by three separate positive, negative, and neutral classes. Annotation guidelines used are as follows:

B. Annotation Guidelines

We have developed some annotation rules according to different scenarios and categorize them as follows:

a) Positive: All those citation sentences which based on words that express *attitude of writers* contains the feeling of "compatibility", "appreciation", "positivity", "excellence", "interest", "admiration", "proposed", "introduced", "analysis", "refers", "thankful" regarding cited paper will be annotated as ''positive". Citations that contain *positive terms* like "outperformed", "accurate", "better", "fast", "favorable", "high quality", and "excellent" etc. Citation sentences that just contain positive terms except the negation terms that reverse the meaning of a sentence like "no", "not", "never", "neither", "nor", and "none" etc.

b) Negative: All those citations sentences based on words that express the *attitude of writers* contain the feeling of "negativity", "doubt", "ambiguity", "criticism", "un-clarity", "degrade" regarding cited paper will be annotated as ''negative". Citation sentences based on *negative terms* like "burden", "complicated", "inability", "lack", "poor", "unclear", and "unexplored" etc. Citation sentences just contain negative terms except for negation terms that reverse the meaning of a sentence like "no", "not", "never", "neither", "nor", and "none" etc.

c) Neutral: All sentences that not contain any positive word and negative words considered as neutrals like "This work was done and evaluated".

C. Statistics of Annotated Corpus

Scientific citation' corpus consists of 5161 and non-scientific citation consists of 4989 sentences. These data sets were annotated using the own defined categories mentioned in Section 4(B). Here are the statistics of the annotated scientific and non-scientific citation sentences' corpus in Table I and Table II.

TABLE. I. Scientific Citations' Statistics

Polarities	**Notations**	**Total Count**	**Percentage**
Positive	P	2014	39.02%
Negative	N	272	5.27%
Neutral	O	2875	55.71%
Total		**5161**	**100**

TABLE. II. NON-SCIENTIFIC CITATIONS' STATISTICS

Polarities	**Notations**	**Total Count**	**Percentage**
Positive	P	2616	52.4
Negative	N	201	4.0
Neutral	O	2172	43.6
Total		**4989**	**100**

V. CLASSIFICATION PROCESSING

This section briefly explains the process of classification applied in this research work. This process consists of various sub-processes including pre-processing data, features' application, classifiers' application, and evaluation metrics.

A. Pre-Processing Data

Data preprocessing is a technique of data mining involving the transformation of raw data into a concise format. Real-world data is often incomplete, contradictory, and lacking in certain habits or patterns, and is likely to contain several mistakes. Preprocessing data is a proven way to solve these problems. Preprocessing the data allows raw data to be processed further. Citations sentences are annotated using 3-classes (Target attributes). Whole data was split into training and testing data using 60:40 ratio randomly.

B. Features' Application

We implemented various features for data classification including N-Grams [16][17], Stop Words Removal [17], Lemmatization [17], Tokenization [17], and Case Normalization to clean down the data.

C. Classifiers' Application

In order to perform the classification procedure we have used different classification algorithms including NB[15][17],SVM[8][17][18][21],DT[2][17],RF[7][8][11][17][18], KNN[17][20][22], LR[5][17][19][24], GB[4][6][23], and NN[1].

D. Evaluation Metrics

To determine the accuracy of a classification we have preferred to use Accuracy score [17], and F-Score [17] evaluation metrics.

VI. RESULTS

Table III represents the evaluation scores of both scientific and non-scientific data sets' classification using the F-Score and Accuracy score. In the case of scientific citation' data set SVM using Uni-gram achieved highest F-score of 70.6% and Accuracy score of 70.6% as well. While in the case of non-Scientific citation' data set LR using tri-gram feature achieved the highest F-score of 65.3% and Accuracy score of 65.3% as well. The reasons for low evaluation scores in case of non-scientific data set is its complex annotation procedure. As human annotators faced much difficulty and complexity while annotating the non-scientific citation sentences due to its complex nature that leads to poor understanding and incorrect annotation. The major reasons we have found because of achieving low accuracy scores in case of non-scientific data set are language differences as linguistic research papers are related to different languages e.g; English, Dutch, French, and Chinese that leads to difficult understanding. Appearing complex terms inside citation sentences is another reason that is responsible for the poor annotation process. Most of the terms that authors found during the annotation procedure were unfamiliar, having different meanings as considered normally. Most of the citation sentences in which the writer's view was difficult to judge that leads to neutral sentiment. Lengthy citation sentences with complex orientation of terms also lead to difficult understanding and annotation process. These are the reasons that leads to complex annotation process and less accuracy scores of non-scientific citation' as compared to scientific citation'.

TABLE. III. HIGHEST SCORES AFTER THIRTY ITERATIONS

Data Set	**N-Gram**	**Classifier**	**F-Score**	**Accuracy Score**
Scientific	Uni-Gram	SVM	70.6%	70.6%
Non-Scientific	Tri-Gram	LR	65.3%	65.3%

VII. CONCLUSION

In this section, we conclude our work done. We have evaluated the citation sentences' annotation complexity of both scientific as well as non-scientific text related articles to find out major complexity reasons by performing sentiment analysis of scientific and non-scientific domain articles by using our own developed corpora of these domains separately. We prepared a science-related data set consisted of 5,161 citation sentences, we also prepared a non-scientific dataset consist of 4,989 citation sentences and applied polarities using our some rules that are mentioned above. We classified these data sets using different classifiers by applying different features. With the evaluation results, we reached a conclusion that in case of scientific data highest f-score of 70.6% and accuracy score of 70.6% using uni-gram feature is achieved while in case of non-scientific data set highest f-score of 65.3% and accuracy score of 65.3% using the tri-gram feature is achieved. We have concluded major reasons of low accuracy scores in case of non-scientific data set are linguistic differences, Complex words, unfamiliar terms, the neutrality of author's sentiment, and lengthy citations sentences with complex orientation of terms. These are the reasons that lead to difficult and complex annotation process leads to less accuracy scores as compared to scientific citation'.

ACKNOWLEDGMENT

The writers would like to begin by thanking Almighty Allah for His grace and blessings. The writers would also like to thank their parents, friends, and colleagues for their encouragement and support.

REFERENCES

[1] Acharya, U. R., Bhat, P. S., Iyengar, S. S., Rao, A., & Dua, S. (2003). Classification of heart rate data using artificial neural network and fuzzy equivalence relation. Pattern recognition, 36(1), 61-68.

[2] Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: a case study. International Journal of Information and Education Technology, 6(7), 528.

[3] Athar, A. (2014). Sentiment analysis of scientific citations (No. UCAM-CL-TR-856). University of Cambridge, Computer Laboratory.

[4] Babajide Mustapha, I., & Saeed, F. (2016). Bioactive molecule prediction using extreme gradient boosting. Molecules, 21(8), 983.

[5] Bai, S. B., Wang, J., Lü, G. N., Zhou, P. G., Hou, S. S., & Xu, S. N. (2010). GIS-based logistic regression for landslide susceptibility mapping of the Zhongxian segment in the Three Gorges area, China. Geomorphology, 115(1-2), 23-31.

[6] Ganjisaffar, Y., Caruana, R., & Lopes, C. V. (2011, July). Bagging gradient-boosted trees for high precision, low variance ranking models. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 85-94). ACM.

[7] Gao, D., Zhang, Y. X., & Zhao, Y. H. (2009). Random forest algorithm for classification of multiwavelength data. Research in Astronomy and Astrophysics, 9(2), 220.

[8] Hasan, M. A. M., Nasser, M., Pal, B., & Ahmad, S. (2014). Support vector machine and random forest modeling for intrusion detection system (IDS). Journal of Intelligent Learning Systems and Applications, 6(01), 45.

[9] Hyland, K. (1995). The Author in the Text: Hedging Scientific Writing. Hong Kong papers in linguistics and language teaching, 18, 33-42.

[10] In European conference on machine learning (pp. 4-15). Springer, Berlin, Heidelberg.

[11] Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. Ieee Access, 5, 16568-16575.

[12] MacRoberts, M. H., & MacRoberts, B. R. (1984). The negational reference: Or the art of dissembling. Social Studies of Science, 14(1), 91-94.

[13] Mehmood, K., Essam, D., & Shafi, K. (2018, July). Sentiment Analysis System for Roman Urdu. In Science and Information Conference (pp. 29-42). Springer, Cham.

[14] Moravcsik, M. J., & Murugesan, P. (1988). Some Results on the Function and Quality of Citations: Social Studies of Science. 研究 技術 計画, 3(4), 538.

[15] Mukherjee, S., & Sharma, N. (2012). Intrusion detectionusing naive Bayes classifier with feature reduction. Procedia Technology, 4, 119-128.

[16] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up? sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10(pp. 79-86). Association for Computational Linguistics.

[17] Raza, H., Faizan, M., Hamza, A., Mushtaq, A., & Akhtar, N. (2019). Scientific Text Sentiment Analysis using Machine Learning Techniques: International Journal of Advanced Computer Science and Applications(IJACSA), 10(12), 2019.

[18] Selvaraj, H., Selvi, S. T., Selvathi, D., & Gewali, L. (2007). Brain MRI slices classification using least squares support vector machine. International Journal of Intelligent Computing in Medical Sciences & Image Processing, 1(1), 21-33.

[19] Tsangaratos, P., & Ilia, I. (2016). Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. Catena, 145, 164-179.

[20] Wang, J. S., Lin, C. W., & Yang, Y. T. C. (2013). A k-nearest-neighbor classifier with heart rate variability feature-based transformation algorithm for driving stress recognition. Neurocomputing, 116, 136-143.

[21] Widodo, A., & Yang, B. S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. Mechanical systems and signal processing, 21(6), 2560-2574.

[22] Yu, X. G., & Yu, X. P. (2006, August). The Research on an adaptive k-nearest neighbors classifier. In 2006 International Conference on Machine Learning and Cybernetics (pp. 1241-1246). IEEE.

[23] Zhang, F., Du, B., & Zhang, L. (2015). Scene classification via a gradient boosting random convolutional network framework. IEEE Transactions on Geoscience and Remote Sensing, 54(3), 1793-1802.

[24] Zhu, J., & Hastie, T. (2002). Kernel logistic regression and the import vector machine. In Advances in neural information processing systems (pp. 1081-1088).

[25] Ziman, J. M. (1968). Public knowledge: An essay concerning the social dimension of science (Vol. 519). CUP Archive.

Arabic Morphological Analysis Techniques

A Survey and Classification

Ameerah Alothman[1], AbdulMalik Alsalman[2]
Department of Computer Science, College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia

***Abstract*—Recently, activity surrounding Arabic natural language processing has increased significantly. Morphological analysis is the basis of most tasks related to Arabic natural language processing. There are many scientific studies on Arabic morphological analysis, yet most of them lack an accurate classification of Arabic morphology and fail to cover both recent and traditional techniques. This paper aims to survey Arabic morphological analysis techniques from 2005 to 2019 and to organize them into a reasonable and expandable classification system. To facilitate and support new research, this paper compares the currently available Arabic morphological analyzers, reaches certain conclusions, and proposes some promising directions for future research in Arabic morphological analysis.**

***Keywords*—Arabic analyzer; Arabic lexicon; classification morphology; morphological analysis; natural language processing**

I. Introduction

Since the advent of the computing era, researchers have been trying to develop systems which can interact with humans; these systems play an essential role in facilitating human life by saving time and improving the quality of work. Morphological analyzers are one such system and constitute an important component of many applications dealing with natural language processing (NLP), machine translation, information search and retrieval, and more.

Morphology is a challenge in Arabic natural language processing (ANLP), and a somewhat complex task. This is because the most important characteristic of Semitic languages is their nonconcatenative nature. Arabic words are composed of roots, derived from certain patterns extracted from stems and their affixes. One root and a small number of patterns with several affixes can form many stems (word formations).

Accordingly, it is necessary to study and classify the techniques of Arabic morphological analyses, because doing so may contribute to greater understanding and improved construction of morphological methodologies, and will pave the way for future researchers in the field of ANLP.

The main purpose of this article is to survey Arabic morphological analysis techniques and bridge the gap in scientific survey studies from 2005 to 2019. This paper is organized as follows: In the second section, we provide basic definitions for this article's most frequently used terms. In the third section, we propose a classification of Arabic morphological analysis techniques and describe some of the shortcomings of earlier classifications. In the fourth section, we present a survey of Arabic morphological analysis techniques. The fifth section presents a discussion of the comparative study undertaken. Finally, we conclude and summarize some important future directions for Arabic morphological analysis techniques. We adopt Buckwalter [1] for the transliteration of Arabic characters, providing transliterations in brackets where relevant.

II. Basic Definitions

There are many terms related to Arabic morphological analyses, and many papers have made great efforts towards the Arabization and standardization of these terms. The book Introduction to Arabic Language Processing, [2] as well as its translation into the Arabic language [3], is one of the most important references in this field of study. Table I presents the meanings and translations of the most frequently used terms in this research.

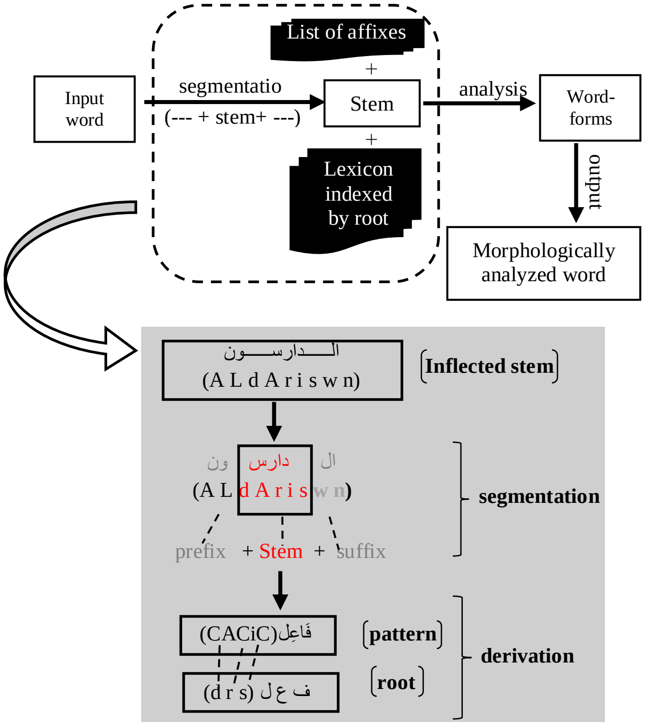


Fig. 1. Root-Pattern Morphology Process.

TABLE. I. BASIC TERMS USED FREQUENTLY IN ARABIC MORPHOLOGICAL ANALYSIS TECHNIQUES (ARRANGED ALPHABETICALLY)

Term	**Translation**	**Meaning**	**Example**
Affix	*اللواصق*	Three types that attach to the root: prefixes, suffixes, and infixes.	ال - فـ - ة
Basic Arabic letter used in patterns	*الحروف العربية الأساسية في الأوزان*	The three basic letters used to construct a pattern in Arabic, which are: [l ل- E ع- f ف].	فَعِلَ(faEila) يَفْعَلُ(yafEalu)
Inflected stem	*الجذع الإعرابي*	A stem that may have a prefix and/or suffix to provide meaningful context, also known as a surface word.	فسيكتبونها (fsyktbwnhA)
Lexeme	*المُعَيْجِمَة*	The smallest part of the lexicon that has meaning.	بيت - حقل
Long vowel	*حروف المد*	Also called the “weak letters set” (*أحرف العلة*); consists of three letters of the alphabet (“a”*الألف*, “w”*الواو*, and “y”*الياء*).	The long vowel in قال (qAl) is ا (A)
Morpheme	الوحدة الصرفية	The smallest unit of the language that has meaning.	ال, إلى أكل ,
Morphological analysis techniques	*تقنيات التحليل الصرفي*	The process used to determine all possible morphological analyses of a word.	See shaded part of Fig. 1
Pattern	*الصيغة أو الوزن الصرفي*	Abstract CV-template (C: Consonant, V: Vowel) representation of how to order the root and short vowels (and some affixes) to generate the stem. It conveys a grammatical meaning, such as part of speech (POS) and tense.	(fEl) فعل (fAEl)فاعل
Root	*الجذر*	A sequence of three (most commonly), four (less commonly), or five (rarely) consonants. It can be derived based on various patterns. It identifies the general meaning of a word.	(zrE)زرع
Short vowel or Diacritic	*الحركات*	Includes diacritics, which are marks usually written above or below a letter. Diacritics include: 1) three short vowels (“a”,*الفتحة* “u” *الضمة* and “i” *الكسرة*), and the absence of any vowel (*"ـْ" السكون*); 2) three nunations (*التنوين*) occurring in the final positions of a word in nominals only; and 3) Shadda (“ـّ” *الشدة*).	ـَ (a) ـِ (i) ـُ (u)
Stem	*الجذع*	The core of concatenative morphology, it is a surface word generated by inserting the radicals of roots and short vowels into the pattern template slots (e.g., the interdigitating of roots with the patterns).	Stem in (fsyktbwnhA) (فسيكتبونها) is (yktbwn) (يكتبون)

III. CLASSIFICATION OF ARABIC MORPHOLOGICAL ANALYSIS TECHNIQUES

Many scientific papers have tackled the classification of Arabic morphology, and several reviews exist of the most cited Arabic morphological analyzers [4-9]. These studies have many shortcomings, including the following: 1) They are very general in their classification process, and most existing analyzers are classified under one category, “linguistic”. 2) They are somewhat outdated (especially in terms of classification methods) and do not take new techniques into consideration. 3) The authors of these review papers do not provide a standard or basis for the construction of their morphological analyzers or define the approaches that were used to analyze words.

Our aim is to bridge the gaps in the previous studies. Therefore, we have classified Arabic morphology in a more detailed and precise manner than previous studies, in terms of the units used in the analysis. This is based on the approach adopted (linguistic or data-driven lexicon) to adequately clarify the variation in work (see Fig. 2). We also limit ourselves to morphology work carried out after 2004, so that our research will complement the comprehensive survey conducted in this field by Al-Sughaiyer and Al-Kharashi [4] in 2004.

According to [4], the classification of Arabic morphological analysis techniques falls into four main approaches, namely, pattern-based, combinatorial, table lookup, and linguistic. This classification neglects the core unit of how to build lookup tables or linguistic rules (i.e. What should they be based on – root, stem, or lexeme?). As we know, Semitic languages are rich in morphology, and therefore the unit of Arabic used in the analysis must first be specified. Moreover, this classification ignores the machine learning approaches that have received more attention in the latest research. In addition, it does not differentiate between different levels of linguistics and does not take Arabic syntax into consideration. Lastly, it includes a pattern-based approach, which can be more accurately described as part of an approach rather than a separate approach in itself. In the next section, we present in greater detail the proposed classification, which is legitimate and covers all recent and traditional techniques.

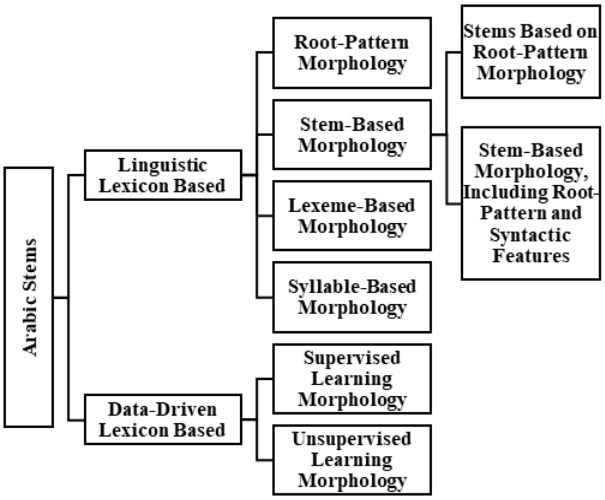

Arabic Stems
Linguistic Lexicon Based
Data-Driven Lexicon Based
Root-Pattern Morphology
Stem-Based Morphology
Lexeme-Based Morphology
Syllable-Based Morphology
Supervised Learning Morphology
Unsupervised Learning Morphology
Stems Based on Root-Pattern Morphology
Stem-Based Morphology, Including Root-Pattern and Syntactic Features

Fig. 2. Suggested Classification of Arabic Morphological Analysis Techniques.

IV. Survey of Arabic Morphological Techniques

This section reviews the main approaches to building Arabic morphological analyzers found in the existing literature. Additionally, it lists the morphological systems that have adopted these approaches. Table II provides a summary of the approaches surveyed.

TABLE. II. SUMMARY OF SURVEYED APPROACHES

Approach	Morphology	Author and reference	Date	Known as	Test Data	Result (%)	Language coverage
Data-driven	Supervised	Elghamry [10]	2005	A constraint-based algorithm	2,700 unique words	Percentage of correct root = 92%	Information not available (N/A)
		Daya et al. [11]	2008	Identifying Semitic roots	N/A	Precision: 87.92%; Recall: 92.19%	MSA
		Boudlal et al. [12]	2011	A Markovian approach	38,022 words	Percentage of correct root: Training set: 98%; Testing set: 93.81%	Non-vowelized
	Unsupervised	Rodrigues and Ćavar [13]	2007	Learning Arabic morphology using statistical constraints	10,000 words from BAMA1 dataset	Root predicted with 75% precision	Non-vowelized
		Snyder and Barzilay [14]	2008	Unsupervised multilingual learning	Snyder & Barzilay (S&B) dataset	Performance of automatic segmentation: Precision = 67.75% Recall = 77.29%	N/A
		Poon et al. [15]	2009	Unsupervised with log-linear models	– S&B dataset – Arabic Treebank (ATB)	– S&B : F1 = 90 – ATB: F1 = 80.2	N/A
		Botha and Blunsom [16]	2013	Adaptor grammar for learning	– BW corpus (without diacritics) – BW' with diacritics – Quranic Arabic (QA)	– Triliteral root identification accuracy: BW = 67.1% BW' = 0.7% – Segmentation: BW = 73.66% BW' = 74.54% – QA has a low performance (excluded from comparison)	Vowelized and non-vowelized
		Fullwood and O'Donnell [17]	2013	Learning nonconcatenative morphology	N/A	Accuracy = 92.3%	Vowelized
		Khaliq and Carroll [18, 19]	2013	Unsupervised induction of Arabic root and pattern lexicons	Quranic Arabic Corpus (QAC)	Root extraction accuracy = 87.2%	Non-vowelized
Linguistic	Root-pattern	Gridach and Chenfour [20]	2014	Developing a new system for Arabic morphological analysis and generation	ALECSO Corpus	Accuracy = 95.08%	MSA
	Lexeme	Habash and Rambow [21]	2006	MAGEAD	Penn Arabic Treebank (PATB), Levantine Arabic Treebank (LATB)	**MSA:** Context type recall (CTyR) = 52.9% Context token recall (CToR)= 60.4% **LEV:** CTyR = 95.4% CToR= 94.2%	MSA & Levantine
		Smrz [22]	2007	ElixirFM	N/A	N/A	N/A
		Habash [23]	2007	ALMOR	1m Arabic words from the United Nations Arabic-English corpus	Precision = 99.61% Recall = 87.78%	MSA
		Attia et al. [24]	2011	AraComLex	– 400,000 words from general news – 400,000 semi-literary words	– 87.13% coverage rate on words from the general news – 85.73% coverage rate on semi-literary words	MSA

		Habash et al. [25]	2012	CALIMA$_{\text{EGY}}$	Manually annotated EGY corpus	1 – Correct Answer = 84.1% 2 – Correct Answer = 92.1%	MSA, Dialectal Arabic (DA)
		Khalifa et al. [26]	2017	CALIMA$_{\text{GLF}}$	4,000 words from Emirati novels	Conventional Orthography for Dialectal Arabic (CODA) = 89.7%	MSA, DA
		Taji et al. [27]	2018	CALIMA$_{\text{star}}$	1m words from the Arabic Gigaword corpus	Coverage of 1.3 % out-of-vocabulary (OOV) rate	MSA, DA
	stems based on root-pattern morphology	Buckwalter [1, 28]	2004	BAMA2	N/A	N/A	MSA
		Maamouri et al. [29]	2010	SAMA 3	N/A	N/A	MSA
	stem-based morphology, including root-pattern and syntactic features	Sawalha et al. [7]	2013	SALMA	– 1000 words from Chapter 29 of the Qur'an, representing Classical Arabic (CA) – Corpus of Contemporary Arabic (CCA) representing MSA	Prediction accuracy of all features = 53.50% for the Qur'an 71.21% for the CCA	CA and MSA
		Boudchiche et al. [30]	2017	AlKhalil	– Tashkeela corpus – Nemlar corpus	99.31% coverage rate	Non-vowelized, partially or totally vowelized text

A. *Linguistic Lexicon-Based Approach*

In the linguistic lexicon-based approach, solid linguistic rules represented in the heavy lexicon are the core data upon which analysis depends. The lexicon contains two main sections: the first comprises word roots and/or patterns and/or stems, grouped in morphological ways, and the second contains any information related to these contents that the system shows in the results. This approach follows the steps in Fig. 3, with some variations depending on the lexicon and its analyses. The following shows the four basic linguistic lexicon-based approaches:

1) Root-pattern morphology: In brief, morphology is the study of the relationship between meaning and form. It is one of the most challenging tasks in Semitic languages like Arabic, Maltese, and Hebrew. For the most part, Arabic morphology is not concatenative (also called discontiguous or nonlinear). Arabic words are generated from their base roots [5]. In linguistics, there are several nonconcatenative methodologies for deriving the stems of words, because they provide linguistic information [6]. Root-pattern is one of these methodologies.

It is useful to briefly review one of the most important theories of nonconcatenative morphology. In 1979, McCarthy [31, 32] proposed a theorem accepted by linguists (especially computational) to form a stem through a derivational integration of roots and patterns. This mechanism is important for representing the structure of a word in Semitic language morphology.

McCarthy's [32] work depends on autosegmentalizing the vowels and placing them in a separate tier from the pattern. It has three tiers, as seen in Fig. 4, where C stands for Consonant, V for Vowel:

1) Root tier: refers to consonantal segments, including the meaning of a lexeme, such as (k t b ك ت ب), which means "write".

2) Pattern tier: refers to a prosodic template associated with a particular meaning or grammatical function such as ((katab) كَتَبَ = CVCVC =CaCaC), which means, "he wrote".

3) Vocalization tier: represents pronounced letters and involves grammatical information such as tense, number, and derivational functions.

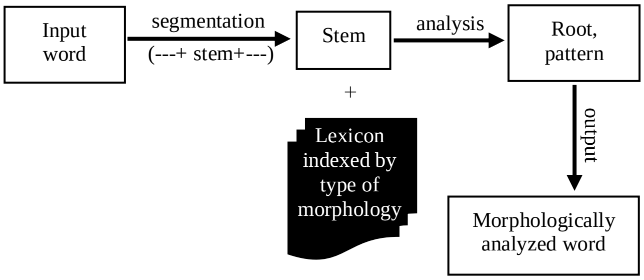


Fig. 3. The basic Steps of the Linguistic Lexicon-based Approach.

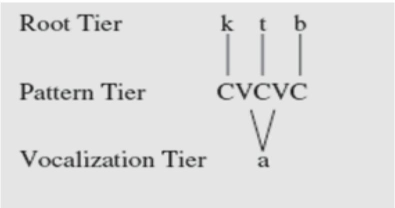


Fig. 4. An Example of McCarthy's [32] Work.

To form an abstract stem, association rules are matched between consonants from the root tier and the pattern tier, and between vowels from the vocalization tier and from the pattern tier. There have been many systems attempting to model Arabic morphology based on McCarthy's theorem. Most of these systems adopted finite-state language modelling tools [33].

Root-pattern morphology depends on the root and pattern of the word entered for analysis (see Table III). The method involves building lexicons of roots and patterns (or lists of Arabic roots and affixes to cover all prefixes, suffixes, and infixes). Continuous research is being done to extract words that belong to one of the entries in these lists. This process is meant to output analysis of stem forms. Fig. 1 illustrates the main steps followed in this morphology.

One of the earliest published works to adopt this morphology was a system proposed by Hlal [34] and Hegazi and El-Sharkawi [35, 36]. It was also adopted by the Xerox lexicon [37], whose entries depend on root and pattern morphemes. Gridach and Chenfour [20] adopted this morphology with some variations in building their lexicon, depending on XML-based morphological definition language (XMODEL) for its construction.

2) Stem-based morphology: Dichy and Farghaly [6] and Farghaly and Senellart [38] support the claim that building a stem-based lexicon is more intuitive, efficient, and easy to develop and extend compared to a lexicon based on roots.

On the other hand, earlier Arabic morphologies were only responsible for the analysis and/or generation of the correct formations of Arabic words. Many Arabic NLP systems, such as machine translations and automatic summarizations, need linguistic information related to each lexical entry to ascribe elaborate knowledge to each word, in order to become more efficient. This information involves the tense of the verb, number, gender, and part of speech (POS), as well as syntactic features such as the type of subject or object, the count of nouns, and so on. In this context, one adds semantic information, such as the categorization of the noun as human, time, place, and so on. This linguistic information is associated with the stems, which are neither roots nor patterns nor a combination of them [6].

According to the above, Arabic stem-based morphology can achieve a more effective morphological strategy by reducing the complexity of word formations and granting linguistic and semantic information to each entry, thus eliminating the greater lexical gaps.

TABLE. III. EXAMPLES OF ROOT-PATTERN MORPHOLOGY

Root	Pattern	In Arabic	Meaning
د ر س (d r s)	(CaCaCa)فَعَلَ	(darasa) دَرَسَ	study
	(CACiC)فَاعِل	(dAris) دَارِس	student
	(CaC~aCa)فَعَّلَ	(dar~asa) دَرَّسَ	he teaches
	(CACiC)فَاعِل	(dAriswn) دَارِسُون	group of students

Two approaches have been built based on this morphology: 1) stems based on root-pattern morphology; and 2) stem-based morphology, including root patterns and syntactic features.

a) Stems based on root-pattern morphology: Briefly, this morphology can be described as follows: each existing lexical entry is checked against candidate entries integrating root and pattern (to generate a stem), in addition to prefix or suffix combinations. Therefore, if the lexicon in this morphology contains, for example, X root and Y pattern, then the XY root-pattern virtual links represent all possible stems, which must be severely restricted to give a reasonable number of meaningful words [33].

The major difference between root-pattern morphology and stems based on root-pattern morphology lies in their analysis mechanisms (see Fig. 1 and 5). The former uses the root and pattern morphemes themselves, while the latter uses stems based on root and pattern morphemes [39].

In this regard, the most famous Arabic analyzer to adopt this morphology is the Buckwalter Arabic Morphological Analyzer (BAMA) [1, 28]. BAMA is based on Buckwalter's lexicon, which is integrated with the Xerox lexicon [38].

Currently, there are three main versions of BAMA. BAMA 1.0 is available for public use, while BAMA 2.0 and Standard Arabic Morphological Analyzer (SAMA) 3.0 [29] are available through the Linguistic Data Consortium (LDC).

b) Stem-based morphology, including root patterns and syntactic features: Dichy and Farghaly [33] present the significance of syntactic features in Arabic computational morphology in detail. Systems based on this method produce a higher level of morphological analyzers, called morpho-syntactic analyzers. As we know, there are six linguistic levels: phonetics, phonology, morphology, syntax, semantics, and pragmatics (see Fig. 6). This approach takes advantage of the features of the syntax and morpheme levels.

This morphology differs from previous approaches because it applies the additional grammatical features step to results such as prepositions "ب"<b> and "ك"<k>, which only appear in the genitive case with nouns. These features play an important role in ensuring proper insertion of lexical entries, especially the main ones, such as nouns and verbs.

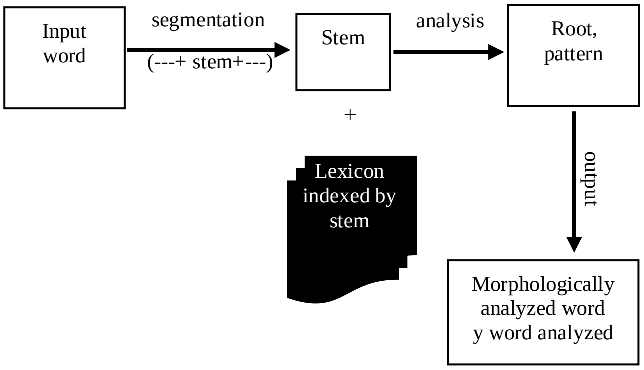


Fig. 5. Stems based on the Root-Pattern Morphology Process.

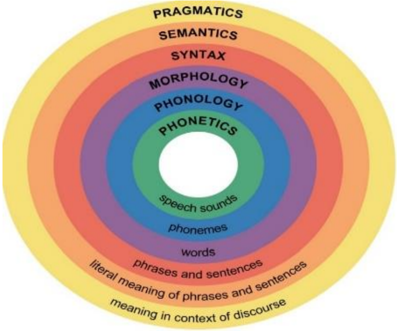


Fig. 6. Linguistic Levels [40].

Standard Arabic Language Morphological Analysis (SALMA) tools [7, 41] fall under this morphological approach. They include SALMA–Tagger, SALMA–ABCLexicon, and SALMA–Tag Set. AlKhalil morphological analyzer [30, 42] also depends on this morphology.

3) Lexeme-based morphology: Typically, lexemes differ only in inflection and cliticization (*الملحقات مثل: أل التعريف وحروف الجر المتصلة كالباء والكاف*). To put it simply, more than one word can be formed from one lexeme. For example, the lexeme *(*bayt*)* بَيْت includes *(*bayt*)* بَيْت , *(*lilbayt*)* لِلبَيْت, , and (buyuwt*)* بُيُوت. Therefore, the lexeme is not equivalent to a word in any language. It is considered an important abstraction used in linguistic morphology, and is the smallest part of the lexicon that has meaning (or semantic content). Additionally, a lexeme has a morphological form and syntactic category [2].

The claim that the stem is a morphological part with greater relevance to the lexeme is the premise underpinning lexeme-based morphology. This methodology depends on the crucial information of the stem, which must be extracted from the word in the right way. Soudi et al. [43] develop a lexeme-based morphology and present an Arabic version of a morphology rule compiled in the MORPHE tool (MORPHE is a general computational engine that works based on transformational rules and a discrimination hierarchy which must be constructed for each language).

In the lexeme-based methodology, the primary representation is made for the stem (including all operations on the stem, such as transformational rules applied to a stem to handle stem variation issues in several contexts of prefixes and/or suffixes). In other words, this methodology adopts a computational implementation of a non-sub-fragmented lexicon. Thus, this methodology differs from the root-pattern methodology, which gives equal consideration and separate lexicons to each constituent of a word (i.e., sub-lexicons for the root, for the pattern, and for vocalization) [5].

Many works on Arabic morphological analyzers adopt this methodology. Among these works are the following: a) a prototype lacking broad coverage, such as the MORPHE tool [43, 44]; and b) large-scale systems such as:

- ElixirFM [22], which reused the Buckwalter lexicon [1, 28].
- MAGEAD [21, 45] CALIMA, both of which handle Arabic dialects (MAGEAD entirely manually designed while CALIMA manually verified the annotated data lexicon using several computational techniques). There are three versions of CALIMA: CALIMAEGY [25], CALIMAGLF [26], and CALIMAstar [27]. Respectively, these cover Egyptian Arabic, Gulf Arabic, and all variants of MSA and Arabic dialects.
- AL-MORGEANA (abbreviated to ALMOR) [23], which extends the BAMA morphological databases with the lexeme and feature keys that are used in the analysis. For example, ALMOR uses the BAMA lexicon but changes the mode of analysis to produce a lexeme-and-feature format as output, rather than the stem-and-affix format, which is the Buckwalter output. It is important to mention here that ALMOR is the analyzer used in the MADA [46] tool. In addition, the new version of MADA is called MADAMIRA [47]. It is a Java NLP tool combining MADA with a shallow syntactic parser called AMIRA [48].
- AraComLex [24], which is based on the MSA lexical database[1], was specifically constructed for this purpose using a corpus of more than one million words.

4) Syllable-based morphology: Most syllable-based morphology work has been performed on European languages such as German, English, and Italian. Cahill [49] asserts the possibility of analyzing the Semitic languages using syllable-based morphology in a way that is not significantly different from that applied to European languages.

However, to our knowledge, there have been no attempts to build an Arabic morphological analyzer adopting this morphology to substantiate or reject this claim.

B. Data-Driven Lexicon-Based Approach

Machine learning techniques underpin these morphologies. These techniques are fast and do not require extensive linguistic knowledge because they depend on the annotated or unannotated corpus used in the training stage. Dinh et al. [50] claim that doubts could be raised around purely data-driven systems (which do not possess any linguistic base), but they are based on a hybrid. The new techniques prove this claim to be untrue. Recently, many supervised and unsupervised learning techniques have proved valuable in this area, as we will demonstrate in the two following subsections. Thus, we predict a promising future for these morphologies.

1) Supervised learning morphology: This approach attempts to infer parameter values from labeled resources without linguistic expertise about data. Supervised learning resources involve lexica of affixes and pairs of inflected words with their roots [51].

Supervised approaches are not famous in the domain of nonconcatenative morphology acquisition. These approaches require a massive lexicon in the training stage to achieve high precision. Some researchers take pride in their ability to avoid these massive lexica, but the disadvantages can be seen in their results, which have many limitations and are therefore not

[1] http://arabiconly.com/aracomlex/form_nominals.php

highly precise in general. However, this is the reality of any new technique. This method will become more promising as more annotated data becomes available.

The existing literature on Arabic morphology that uses this approach to identify Arabic roots is limited. There are two types which adopt some supervised learning: a) learning that is based on pre-existing dictionaries using Hidden Markov Models (HMM) [12] or neural network (NN) models [52], and b) learning that only uses rule constraints [10] or multi-class classifier models [11].

2) Unsupervised learning morphology: Unsupervised learning morphology, in essence, is the process of acquiring intra-word structures and the rules by which they merge to generate word forms [16]. In other words, morphology is induced without prior knowledge, based on training that uses large volumes of unannotated data, without supplying an example of the expected output. This research field began in the mid-1990s and continues today. Researchers consider unsupervised approaches attractive because of the large quantities of unlabelled data available on the Internet [15]. In recent years, unsupervised learning of concatenative language morphology (e.g., stem+affix morphology) has received more attention than nonconcatenative language morphology (e.g., root and pattern morphology) [53].

There are few studies in this field, but they vary according to the objectives of their algorithms. Some aim to learn segmentation [13-15], which means transforming a given word into its stem and affix(es), whereas others aim to learn lexica and patterns [16, 17], which means providing a list of the patterns and assigning each pattern the lexicon information related to all stems belonging to it.

In a significant contribution to this field of research, Khaliq and Carroll [18, 19] have built a morphological analyzer based on roots and patterns induced from the lexicon, based on learning from an unannotated corpus rather than linguistic rules, as noted in the section of this paper dealing with root-pattern morphology. This analyzer achieved good accuracy with root extraction, achieving 94% after many iterative reinforcement stages.

V. Discussion

As shown in the previous survey section, there are multiple morphological analyzers, with varying accuracy and features. No analyzer provides perfect performance, and none has been adopted as standard. Therefore, choosing one of these existing analyzers is difficult and represents a challenge in NLP tasks.

In this section, we compare the analyzers available for public use. Most relevant morphological analyzers achieved acceptable results (according to their developers) but were not available for reuse or evaluation.

To the best of our knowledge, the most recent and efficient morphological analyzers to achieve good accuracies for Arabic morphology are AlKhalil, AraComLex, and ALMOR. ALMOR is no longer available for download. It was distributed as part of MADA Distribution from Columbia University. A new version of MADA, called MADAMIRA, is now available. MADAMIRA is a morphological analyzer and a POS tagger (i.e., MADAMIRA operates *within* a word context while AlKhalil and AraComLex operate *outside* of a word context). Table IV compares these analyzers according to various attributes.

TABLE. IV. Comparison of Available Arabic Morphological Analyzers

Attribute name	**ALMOR**	**AraComLex**	**AlKhalil**
Different configurations (pre-/post-processing)	Yes	No	Yes
Performance metric	Precision & recall	Coverage rate	Coverage rate
Running through	Application Programming Interface (API)	API	Graphical User Interface (GUI)
Directionality	Analysis and generation	Analysis only	Analysis only
Expected input	Text only (works on diacritized text, but no consideration of these diacritics)	Non-vowelized word (does not work on a diacritized word)	Fully or partially diacritized text
Accuracy (sample of 50 words) [8]	88%	56%	90%
Engine	Code-based (Java and Perl languages)	Finite-state machinery	Code-based (Java and Perl languages)
Input format	Word or text	Just one word per query	Word or text
Output format	Text of (feature: value) pairs	In one line, separate between features by (+)	Table (like CSV file) of features
Tag set	About 36 basic tag sets	About 14 tag sets	About 118 tag sets
Transliteration schemes for results	Buckwalter	UTF-8	UTF-8
Last version	As a part of MADAMIRA 2014	AraComLex 2.1 (2018)	AlKhalil 2 (2016)

VI. Conclusion

Many scientific studies discuss Arabic morphological analysis techniques, reviews, and analyzer tools, but they lack a specific and accurate classification of traditional and recent methods. In fact, the linguistic lexicon-based and data-driven lexicon-based approaches are the two main approaches for morphological analysis techniques. All techniques found in the existing literature align with these approaches. This classification can guide us towards standard Arabic morphological analysis techniques.

A linguistic lexicon-based approach depends on solid linguistic rules derived from the lexicon. It covers four types of morphology based on analysis process terms: root-pattern, stem, lexeme, and syllable. The data-driven lexicon-based approach depends on an annotated or unannotated corpus to undergo a training process on data, in order to collect rules which are then used to output word forms.

Most of the systems mentioned in this survey are not available for public use. We highlighted the most recent available systems, and compared them on various aspects.

It is important that future research in Arabic morphological analysis investigate the following issues:

- Developing a gold standard Arabic corpus that can be used to compare morphological analysis systems.
- Developing a large annotated Arabic corpus to be used in the promising data-driven approach morphologies.
- Developing a hybrid approach using linguistic and data-driven morphologies to merge the advantages and strengths of these two approaches.
- Using a unified standard of performance metrics in evaluation systems to compare approaches.
- Building a multicomponent toolkit for Arabic morphological analyzers to integrate these analyzers' results and choose the one with the best performance.
- Building a multicomponent toolkit for Arabic morphological analyzers in order to facilitate a selection process for the one that best fits the researcher's/user's needs.

References

[1] T. Buckwalter, Buckwalter Arabic Morphological Analyzer Version 1.0. Philadelphia: Linguistic Data Consortium, 2002.

[2] N. Y. Habash, Introduction to Arabic Natural Language Processing. San Rafael: Morgan & Claypool, 2010.

[3] H. Alkalifah, م قدمة ف ي ال معال جة ال ط ب يع ية ل لغة ال عرب ية. Saudi Arabia: King Saud University Press, 2014.

[4] I. A. Al‐Sughaiyer and I. A. Al‐Kharashi, "Arabic morphological analysis techniques: A comprehensive survey," J. Am. Soc. Inf. Sci. Technol., vol. 55, pp. 189‐213, February 2004.

[5] A. Soudi, G. Neumann, and A. van den Bosch, "Arabic computational morphology: Knowledge-based and empirical methods," in Arabic Computational Morphology, A. Soudi, A. Bosch, and G. Neumann, Eds. Dordrecht: Springer, 2007, pp. 3–14.

[6] J. Dichy and A. Farghaly, "Roots & patterns vs. stems plus grammar-lexis specifications: On what basis should a multilingual lexical database centred on Arabic be built," in The MT-Summit IX Workshop on Machine Translation for Semitic Languages. New Orleans, 2003.

[7] M. Sawalha, E. Atwell, and M. A. M. Abushariah, "SALMA: Standard Arabic language morphological analysis," in 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), 2013, pp. 1–6.

[8] A. Alosaimy and E. Atwell, "Tagging classical arabic text using available morphological analysers and part of speech taggers," J. Lang. Technol. Comput. Linguist., vol. 32, pp. 1–26, December 2017.

[9] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," J. King Saud Univ. Comput. Inf. Sci., doi:10.1016/j.jksuci.2019.02.006, February 2019.

[10] K. Elghamry, "A constraint-based algorithm for the identification of Arabic roots," in Proceedings of the Midwest Computational Linguistics Colloquium. Bloomington: University of Indiana, 2005.

[11] E. Daya, D. Roth, and S. Wintner, "Identifying semitic roots: Machine learning with linguistic constraints," Comput. Linguist., vol. 34, pp. 429–448, September 2008.

[12] A. Boudlal, R. Belahbib, A. Lakhouaja, A. Mazroui, A. Meziane, and M. Bebah, "A markovian approach for Arabic root extraction," Int. Arab J. Inf. Technol., vol. 8, pp. 91–98, January 2011.

[13] P. Rodrigues and D. Cavar, "Learning Arabic morphology using statistical constraint-satisfaction models," Amst. Stud. Theory Hist. Linguist. Sci. 4, vol. 289, pp. 63–75, January 2007.

[14] B. Snyder and R. Barzilay, "Unsupervised multilingual learning for morphological segmentation," in Proceedings of ACL-08: HLT. Ohio: Association for Computational Linguistics, 2008, pp. 737–745.

[15] H. Poon, C. Cherry, and K. Toutanova, "Unsupervised morphological segmentation with log-linear models," in Annual Conference of the North American Chapter of the ACL. Boulder, Colorado: Association for Computational Linguistics, 2009, pp. 209–217.

[16] J. A. Botha and P. Blunsom, Adaptor Grammars for Learning Non-Concatenative Morphology. Stroudsburg: Association for Computational Linguistics, 2013.

[17] M. Fullwood and T. O'Donnell, "Learning non-concatenative morphology," in Proceedings of the 4th Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL). Sofia: Association for Computational Linguistics, 2013, pp. 21–27.

[18] B. Khaliq and J. Carroll, "Unsupervised induction of arabic root and pattern lexicons using machine learning," in Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013. Shumen: INCOMA Ltd., 2013, pp. 350–356.

[19] B. Khaliq and J. Carroll, "Induction of root and pattern lexicon for unsupervised morphological analysis of Arabic," in Proceedings of the 6th International Joint Conference on Natural Language Processing. Japan: Asian Federation of Natural Language Processing, 2013, pp. 1012–1016.

[20] M. Gridach and N. Chenfour, "Developing a new system for Arabic morphological analysis and generation," in Proceedings 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP). Thailand: IJCNLP, 2011, pp. 52–57.

[21] N. Habash and O. Rambow, "MAGEAD: A morphological analyzer and generator for the Arabic dialects," in Proceedings of 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Sydney: Association for Computational Linguistics, 2006, pp. 681–688.

[22] O. Smrz, "ElixirFM – implementation of functional arabic morphology," in Proceedings of 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources. Prague: Association for Computational Linguistics, 2007, pp. 1–8.

[23] N. Habash, "Arabic morphological representations for machine translation," in Arabic Computational Morphology: Knowledge-Based and Empirical Methods, A. Soudi, A. van den Bosch, and G. Neumann, Eds. Dordrecht: Springer Netherlands, 2007, pp. 263–285.

[24] M. Attia, P. Pecina, A. Toral, L. Tounsi, and J. van Genabith, "An open-source finite state morphological transducer for modern standard Arabic," in Proceedings of 9th International Workshop on Finite State Methods and Natural Language Processing. France: Association for Computational Linguistics, 2011, pp. 125–133.

[25] N. Habash, R. Eskander, and A. Hawwari, "A morphological analyzer for Egyptian Arabic," in Proceedimgs of 12th Meeting of the Special Interest Group on Computational Morphology and Phonology. Canada: Association for Computational Linguistics, 2012, pp. 1–9.

[26] S. Khalifa, S. Hassan, and N. Habash, "A morphological analyzer for Gulf Arabic verbs," in Proceedings of Third Arabic Natural Language Processing Workshop. Stroudsburg: Association for Computational Linguistics, 2017, pp. 35–45.

[27] D. Taji, S. Khalifa, O. Obeid, F. Eryani, and N. Habash, "An arabic morphological analyzer and generator with copious features," in Proceedings of 15th Workshop on Computational Research in Phonetics, Phonology, and Morphology. Belgium: Association for Computational Linguistics, 2018, pp. 140–150.

[28] T. Buckwalter, Buckwalter Arabic Morphological Analyzer: Version 2.0. Philadelphia: Linguistic Data Consortium, 2004.

[29] M. Maamouri, D. Graff, B. Bouziri, S. Krouna, A. Bies, and S. Kulick, LDC Standard Arabic Morphological Analyzer (SAMA), Version 3.1. Philadelphia: Linguistic Data Consortium, 2010.

[30] M. Boudchiche, A. Mazroui, M. O. A. O. Bebah, A. Lakhouaja, and A. Boudlal, "AlKhalil morpho Sys 2: A robust Arabic morpho-syntactic analyzer," J. King Saud Univ. Comput. Inf. Sci., vol. 29, pp. 141–146, April 2017.

[31] J. J. McCarthy, Formal Problems in Semitic Phonology and Morphology, Ph.D. Thesis. Cambridge: Massachusetts Institute of Technology, 1979.

[32] J. J. McCarthy, "A prosodic theory of nonconcatenative morphology," Linguist. Inq., vol. 12, pp. 373–418, January 1981.

[33] J. Dichy and A. Farghaly, "Grammar-lexis relations in the computational morphology of Arabic," in Arabic Computational Morphology, A. Soudi, A. Bosch, and G. Neumann, Eds. Dordrecht: Springer, 2007, pp. 115–140.

[34] Y. Hlal, "Morphology and syntax of the Arabic language," in Computers and the Arabic language, A.M. Pierre, Ed. Bristol: Taylor & Francis/Hemisphere, 1990, pp. 201–207.

[35] N. H. Hegazi and A. A. El-Sharkawi, "An approach to a computerized lexical analyzer for natural Arabic text," in Proceedings of the Arabic Language Conference. Kuwait, 1985.

[36] N. H. Hegazi and A. A. El-Sharkawi, "Natural Arabic language processing," in Proceedings of the National Computer Conference. Riyadh, 1986, pp. 10–15–11–10–15–17.

[37] K. R. Beesley, "Finite-state morphological analysis and generation of Arabic at Xerox Research: Status and plans in 2001," in ACL Workshop on Arabic Language Processing: Status and Perspective, 2001, pp. 1–8.

[38] A. Farghaly and J. Senellart, "Intuitive coding of the Arabic lexicon," in SYSTRAN, MT, Summit IX Workshop, Machine Translation for Semitic Languages: Issues and Approaches, Tuesday September. New Orleans: Citeseer, 2003.

[39] T. Buckwalter, "Issues in Arabic morphological analysis," in Arabic Computational Morphology, N. Ide, J. Veronis, A. Soudi, A. van den Bosch, and G. Neumann, Eds. Netherlands: Springer, 2007, pp. 23–41.

[40] Wikimedia Commons, Major levels of linguistic structure. 2019. Available: https://commons.wikimedia.org/wiki/File:Major_levels_of_linguistic _structure.svg.

[41] M. Sawalha and E. S. Atwell, "ت وظ يف ق واعد ال نحو وال صرف ف ي ب ناء محلل صرف ي ل لغة ال عرب ية" (Adapting language grammar rules for building a morphological analyzer for Arabic text)," in Proceedings of ALECSO Arab League Educational Cultural and Scientific Organization workshop on Arabic morphological analysis. Damascus, 2009.

[42] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, M. Bebah, and M. Shoul, "Alkhalil morpho sys1: A morphosyntactic analysis system for arabic texts," in International Arab Conference on Information Technology. Benghazi, 2010, pp. 1–6.

[43] A. Soudi, V. Cavalli-Sforza, and A. Jamari, "A computational lexeme-based treatment of Arabic morphology," in Proceedings of the Arabic Natural Language Processing Workshop, Conference of the Association for Computational Linguistics (ACL 2001), 2001, pp. 50–57.

[44] V. Cavalli-Sforza, A. Soudi, and T. Mitamura, "Arabic morphology generation using a concatenative strategy," in Proceedings of the 1st North American chapter of the Association for Computational Linguistics Conference. Stroudsburg: Association for Computational Linguistics, 2000, pp. 86–93.

[45] N. Habash, O. Rambow, and G. Kiraz, "Morphological analysis and generation for Arabic dialects," in Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages. Ann Arbor: Association for Computational Linguistics, 2005, pp. 17–24.

[46] N. Habash, O. Rambow, and R. Roth, "MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization," in Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR). Egypt, 2009, pp. 102–109.

[47] A. Pasha, M. Al-Badrashiny, M. Diab, A. E. Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. Roth, "MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic," in Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014). Reykjavik: European Language Resources Association (ELRA), 2014, pp. 1094–1101.

[48] M. Diab, "Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking," in 2nd International Conference on Arabic Language Resources and Tools, 2009.

[49] L. Cahill, "A syllable based account of Arabic morphology," in Arabic Computational Morphology, A. Soudi, G. Neumann, and A. van den Bosch, Eds. Netherlands: Springer, 2007, pp. 45–66.

[50] D. Dinh, H. Kiem, and E. Hovy, "BTL: A hybrid model for English-vietnamese machine translation," in Proceedings of the IXth MT Summit. New Orleans, 2003, pp. 87–94.

[51] J. A. Botha, Probabilistic modelling of morphologically rich languages. 2015. Available: https://ui.adsabs.harvard.edu/\#abs/2015arXiv150804271B.

[52] H. Al-Serhan and A. Ayesh, "A triliteral word roots extraction using neural network for Arabic," in 2006 International Conference on Computer Engineering and Systems. Cairo, Egypt: IEEE, 2006, pp. 436–440.

[53] H. Hammarström and L. Borin, "Unsupervised learning of morphology," Comput. Linguist., vol. 37, pp. 309–350, June 2011.

Using Social Network Analysis to Understand Public Discussions: The Case Study of #SaudiWomenCanDrive on Twitter

Zubaida Jastania[1], Rabeeh Ayaz Abbasi[3*]
Kawther Saeedi[4]
Faculty of Computing and Information Technology
King Abdul-Aziz University
Jeddah, Kingdom of Saudi Arabia

Mohammad Ahtisham Aslam[2]
Faculty of Computing and Information Technology
King Abdul-Aziz University
Jeddah, Kingdom of Saudi Arabia

***Abstract*—Social media analytics has experienced significant growth over the past few years due to the crucial importance of analyzing and measuring public social behavior on different social networking sites. Twitter is one of the most popular social networks and means of online news that allows users to express their views and participate in a wide range of different issues in the world. Expressed opinions on Twitter are based on diverse experiences that represent a broad set of valuable data that can be analyzed and used for many purposes. This study aims to understand the public discussions that are conducted on Twitter about essential topics and developing an analytics framework to analyze these discussions. The focus of this research is the analytical framework of Arabic public discussions using the hashtag #SaudiWomenCanDrive, as one of the hot trends of Twitter discussions. The proposed framework analyzed more than two million tweets using methods from social network analysis. The framework uses the metrics of graph centrality to reveal essential people in the discussion and community detection methods to identify the communities and topics used in the discussion. Results show that @SaudiNews50, @Algassabinasser, and @Abdulrahman were top users in two networks, while @KingSalman and @LoujainHathloul were the top two users in another network. Consequently, "King Salman" and "Loujain Hathloul" Twitter accounts were identified as influencers, whereas "Saudi News" and "Algassabi Nasser" were the leading distributors of the news. Therefore, similar phenomena could be analyzed using the proposed framework to analyze similar behavior on other public discussions.**

Keywords—Social network analysis; twitter; public discussion; network science

I. Introduction

Understanding human social behavior is important to understand the sophisticated processes that happen in real life. Social networks, as media of human communication, can be considered as a primary source of social behavior analysis [1]. Social network analysis (SNA) that was originated from graph theory has become one of the essential tools for studying social movements. Twitter is one of the most popular social networks and means of online news that allows users to express their views and participate in a wide range of different issues in the world. In the contemporary news flow, Twitter is among the channels that allow individual users to interact in a more engaged and self-driven way. Twitter has also become a platform by which different views are shared and spread through society, and it plays a vital role in current social events and processes.

The "Saudi Women Driving" was a debate for a long time [2][3], in which researchers study and analyze to see the trends and factors that affect people's decisions [4]. The purpose of this analysis is to find which of the users who shared their opinion were more critical in this discussion; whether they formed groups based on their opinion; and whether the more dominant influence is popularity—for example, in terms of the number of followers. The analysis also digs into further analysis of user behavior, including what is retweeted, or who is co-mentioned.

The rest of the paper is organized as follows: Section II discusses the background and related work, Section III presents the methodology used in this paper, Section IV presents the results, Section V discusses the results and Section VI concludes the paper.

II. Background and Related Work

Social networks form an environment where people express their opinions and collaborate to instigate discussions which affect governments to take requested decisions. Moreover, it is also an environment where people express their gratitude and thankfulness to others. In addition, they are also used for marketing. Therefore, studying and analyzing social networks could reveal information flow and patterns that could be used for decision making [5][6][7]. Often Social networks are studied based on graph theory [8].

In mathematics, graph theory is the study of graphs, used to model pairwise relations between objects. Therefore, a graph is made up of vertices (or nodes), which are connected by edges (or links). Based on the graph theory many concepts have emerged that were used to resolve issues in Social Network Analysis. The centralization of any network is a measure of how central its most central nodes are [9]? There are several metrics for centralization including degree, eigenvectors, closeness, and PageRank. The in-degree and out-degree give an indication to the centralization in a directed graph.

Twitter is a way to express patterns and significant topics discussed online. Twitter information can be classified based

*Corresponding Author.

on patterns of information flow [2][10]. Researchers used the metrics of density, modularity, centralization, and the fraction of isolated users [2]. Consequently, they identified categories which include Broadcast and Support; Polarized and In-Group, and Brand and Clustered Community networks. However, the output depends on the search string, and the model was tested on a limited set of topics including health, politics, leisure, academic, and commercial.

The Twitter network structure is affected by geographical and language diversity. Based on a network analysis of 2,500 users, the model in [3] shows the structure of the network's relationships and users' position in the discussion online. They found that people who discuss the same language tend to have similar clusters. They found that defining digital humanities help understand the behavior of people, where a small number of individuals and institutions are discussing. However, this model, using centrality metrics of in-degree, out- degree, betweenness, and eigenvector does not include other features of retweet networks which limits the gatekeeper and hubs of users' information.

AlFarasani et al. [11] proposed a road incident prediction model using an Arabic lexicon. The predicted incident classes from the city of Riyadh, Saudi Arabia over 10,000 tweets provide an accuracy of 82%. Kuroshima1 and Tian [12] mined 10,000 tweets to demonstrate a practical case of utilizing social media in identifying customer opinions to build an effective drug detection system. They applied a set of machine learning algorithms (Decision Trees, Random Forest, Support Vector Machines, K-Nearest Neighbors, and Naïve Bayes) to classify a set of selected drugs (Advil, Aleve, Motrin, Tylenol) based on a set of keywords features. The work in [13] used the concept of homophily to design an affiliation graph to identify the most popular community. They have used data mining techniques and embeddedness, betweenness, and graph occupancy to detect the evolutionary dynamics of social behavior.

The use of case studies in a research like the one proposed herein is strategic for multiple reasons. Reviewing different articles that are conducting similar research allow for the utilization of information that would normally not be easily attainable. Twitter, for example, is such a big data source, with a myriad of topics and issues being addressed, extracting specific data would require a specific framework in order to comprehend the analytical results. Therefore, the use of case studies and different literature materials submitted by different researchers who use different social media platforms as sources of their data provide consequential information that would have otherwise been difficult to obtain. This includes research conducted by Thom and Kruger [4], who were investigating the possibility whether Twitter can save lives. To achieve their objective, they considered the 2013 German Floods as their case study and collected all the relevant Twitter data on the floods. Even though the researchers would be able to acquire paramount data on the floods, the case study is still relatively very general, as it would yield an insurmountable amount of data that would still require filtering. A related research explored the use of Twitter for saving lives by analyzing blood donation requests on Twitter [14]. A recent study used machine learning to identify tweets related to the empowerment of women in Saudi Arabia by a startup and determining the sentiments of these tweets [15].

A more specific case study yielded valuable big data for analysis in the case of the Brexit referendum. Grčar et al. [16] investigated the influence of Twitter users regarding the Brexit referendum, which showed a possibility to establish the relationship between the Twitter mood and the outcome of the referendum. Despite following a single topic, these researchers faced specific challenges concerning the demographics of Twitter users concerning that of the real voters. These concerns are genuine and warrant the need for a case study for a practical framework to analyze social media data to be developed. Similar challenges are distinguished by Wan and Gao [17], who were analyzing Twitter data for airline services analysis. They concluded that for an overall accuracy while analyzing Twitter data, a classification system adoption would be necessary. Such a classification system for the research at hand would analyze a hashtag as the case study.

A number of Twitter analysis models could be used to analyze the impact on a patient's life. A study of patient-driven innovation in diabetes management resulted in building and sharing knowledge around a do-it-yourself (DIY) open-source artificial pancreas systems (OpenAPS). The developers OpenAPS [18] examined Twitter data to understand how patients, caregivers, and care partners perceive OpenAPS. They found a different cluster of people has several forms that provide a widely accepted way to manage day-to-day glucose levels and quality of life. However, the work did not provide details of social network metrics calculations. Also, in the medical domain, the study of people with dementia was analyzed as some people with dementia publicly share their experiences on Twitter [19]. Talbot et al. [19] have manually identified six themes relating to collaboration, experience, community, and stories of dementia. While the study raises awareness, challenge stigma, it was limited to 3 countries with only 2,774 tweets.

As part of the role of education network reform in the US, Rosenberg [20] studied the debate changes in teaching and learning by analyzing tweets of Next Generation Science Standards chat hashtag (#NGSSchat). They examined users' profile data, locations, and tweets with regression model regression analysis for the exploration of influence to study the diverse participants on the topic in terms of occupation, location, and activity. Therefore, they identified what explains the significant interaction between conversing and endorsing participants. As a result, they found that administrators and teachers appear to be central, though there does not appear to be robust clustering between groups.

Based on the graph theory, many concepts have emerged that were used to resolve issues in Social Network Analysis (SNA). Social network analysis has been used in a variety of applications including but not limited to archeology [21], education [22][23], influence propagation [24], [25], event detection [26]–[29], altmetrics [30], disaster management [31], community detection [32], linguistics [33], analyzing money laundering [34], tourism [35], and software engineering [36]. The betweenness-centrality is a measure of centrality based on shortest paths [9], where the shortest path is the minimum number of edges between two vertices. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex. In twitter, betweenness centrality finds those who are on the most paths between others in the network [37]. Therefore, people with high betweenness tend to be the innovators and brokers in any network [38].

III. Materials and Methods

A. Dataset

The data was crawled using the search API for the Arabic keywords related to the topic "Saudi women can drive" (Arabic hashtags were used related to the discussions). As a result, the dataset contained two million tweets, with more than 680,000 users. The statistics of the dataset are shown in Table I. Sample tweets from the dataset are given in Table II. After getting the tweets, networks were extracted. To illustrate the networks extracted from the data, Fig. 1 to 3 show various networks. Fig. 1 shows the retweet (RT) network. According to the figure, each user who has retweeted someone else appears as the initiator of a link – between him/her and the retweeted one (with the end pointing towards the retweeted one).

Fig. 2 shows the mention network extracted from the sample tweets of Table I. The figure shows which users are mentioned by the initial user in his/her tweet, thus creating a link between both (again, the point is directed towards the one mentioned). Similarly, the hashtag network is constructed based on hashtag co-occurrence, i.e., when two hashtags appear in a tweet, thus generating a link between them. An example of hashtag network is shown in Fig. 3. It is an undirected network. Similar to the hashtag network, the co-mention network is based on the co-occurrences of the people mentioned together in a tweet.

Therefore, with many types of networks, the analysis of Twitter data is comprehensive. For example, the Retweet network can be a measure of how much attention the retweeted user attracted through his post. To what extent he/she provoked a reaction in the discussion in general. The Mention network is pointing to some participants that might be interested in or used as an example in proving a point. The hashtags network, for example, can be an estimate of how many themes were intertwined and how this discussion was formed by the common phrases of the Saudi women driving.

B. Networks

In the course of the analysis, the original dataset was transformed into several types of networks:

- Retweet: a link is formed when a user retweets another user (directed network);
- Mention-Network: a link is formed when a user mentions another user (directed network);
- Co-Mention-Network: when two users are mentioned in the same tweet (undirected network); and
- Hashtag-Network: when two hashtags appear in the same tweet, and there will be an edge between them (undirected network).

Table III shows the total number of nodes and edges in each network.

C. Methods

Analyzing large volume of information can be a considerable effort, so, whenever possible, scientists look for ways that can reveal clues of the bigger picture. Essentially, this is the case with the current study, which tries to get at the essence of more than two million tweets to highlight patterns, draw out expected behaviors, and reveal indispensable persons and areas for further analysis. The centralization of any network is a measure of how central are its nodes [9]. There are several metrics for centralization, including the degree, eigenvector, closeness, and PageRank. The in-degree and out-degree indicate the centralization of a Twitter user, while the eigenvector is a measure of node importance in a network based on a node's connections. When information flows in a hierarchal way, it provides information about hierarchical or egalitarian, which indicates the level to which people are hubs to sharing information or gatekeepers of information. Table IV summarizes commonly used SNA metrics.

TABLE. I. Dataset Statistics

Entity	Frequency
Tweets	2,132,379
Hashtags	2,121,087
Retweets	1,616,543
Users	630,824

TABLE. II. Sample Tweets of the Dataset of Two Million Tweets

User	Tweet
user-1	RT @user-10: The only consistent in Saudi history is change. - Saudi FM @AdelAljubeir□□#SaudiWomenCanDrive #MyArabia https://t.co/07kNKkNW9Y
user-2	RT @FordMiddleEast: Hi @user-11, we'd like to give you your dream car. #MustangSahar #SaudiWomenCanDrive https://t.co/Tln8aiWUNU
user-3	RT @FordMiddleEast: Hi @user-11, we'd like to give you your dream car. #MustangSahar #SaudiWomenCanDrive https://t.co/Tln8aiWUNU
user-4	RT @user-12: Let's not forget who fought to make this possible @manal_alsharif #المرأة_ل قيادة_ي ن تصر__الملك #SaudiWomenDriving https:…
user-5	RT @user-13: Now we all know what it meant. @user-14 #المراه_ق ياده_ي رف ض_ال شعب #SaudiWomenDriving https://t.co/Tan1xH9GHp
user-6	RT @user-13: Now we all know what it meant. @user-14 #المراه_ق ياده_ي رف ض_ال شعب #SaudiWomenDriving https://t.co/Tan1xH9GHp
user-7	RT @user-15: Buckle up, ladies! @user-16 writing on the big news from #SaudiArabia this week https://t.co/PpoXWPnogC #SaudiWomenC…
user-8	RT @user-17: The day after #KSA announced that #SaudiWomenCanDrive, @SAP attracts young female professionals, to fuel next wave of growth.…
user-9	RT @user-13: Now we all know what it meant. @user-14 #المراه_ق ياده_ي رف ض_ال شعب #SaudiWomenDriving https://t.co/Tan1xH9GHp

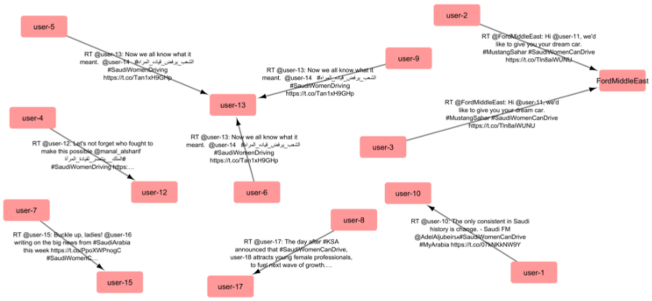

user-5
user-2
user-9
FordMiddleEast
user-13
user-4
user-3
user-12
user-6
user-10
user-7
user-8
user-15
user-17
user-1
RT @user-13: Now we all know what it meant. @user-14 #الشعب_يرفض_قياده_المراه #SaudiWomenDriving https://t.co/Tan1xH9GHp
RT @FordMiddleEast: Hi @user-11, we'd like to give you your dream car. #MustangSahar #SaudiWomenCanDrive https://t.co/Tln8aiWUNU
RT @user-12: Let's not forget who fought to make this possible @manal_alsharif #الملك__ينتصر_لقيادة_المرأة #SaudiWomenDriving https:…
RT @user-15: Buckle up, ladies! @user-16 writing on the big news from #SaudiArabia this week https://t.co/PpoXWPnogC #SaudiWomenC…
RT @user-17: The day after #KSA announced that #SaudiWomenCanDrive, user-18 attracts young female professionals, to fuel next wave of growth….
RT @user-10: The only consistent in Saudi history is change. - Saudi FM @AdelAljubeirد#SaudiWomenCanDrive #MyArabia https://t.co/07kNKkNW9Y

Fig. 1. Retweet Network Extracted from the Sample Tweets Shown in Table I.

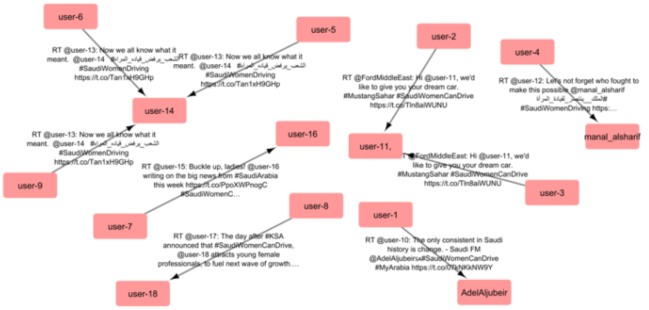

user-6
user-5
user-2
user-4
user-14
user-16
user-11,
manal_alsharif
user-9
user-7
user-8
user-1
user-3
user-18
AdelAljubeir
RT @user-13: Now we all know what it meant. @user-14 #الشعب_يرفض_قياده_المراه #SaudiWomenDriving https://t.co/Tan1xH9GHp
RT @FordMiddleEast: Hi @user-11, we'd like to give you your dream car. #MustangSahar #SaudiWomenCanDrive https://t.co/Tln8aiWUNU
RT @user-12: Let's not forget who fought to make this possible @manal_alsharif #الملك__ينتصر_لقيادة_المرأة #SaudiWomenDriving https:…
RT @user-15: Buckle up, ladies! @user-16 writing on the big news from #SaudiArabia this week https://t.co/PpoXWPnogC #SaudiWomenC…
RT @user-17: The day after #KSA announced that #SaudiWomenCanDrive, @user-18 attracts young female professionals, to fuel next wave of growth….
RT @user-10: The only consistent in Saudi history is change. - Saudi FM @AdelAljubeirد#SaudiWomenCanDrive #MyArabia https://t.co/07kNKkNW9Y

Fig. 2. Mention Network Extracted from the Sample Tweets Shown in Table I.

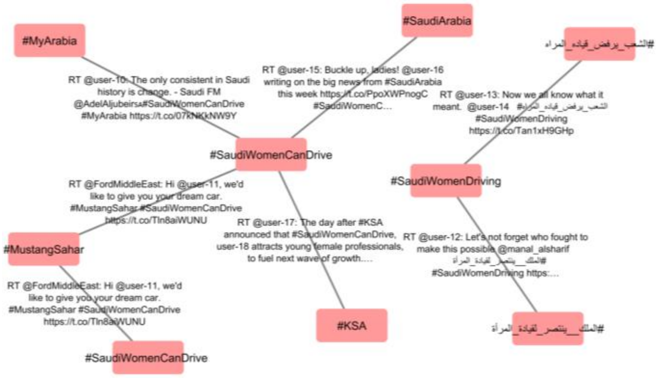

#MyArabia
#SaudiArabia
#الشعب_يرفض_قياده_المراه
#SaudiWomenCanDrive
#SaudiWomenDriving
#MustangSahar
#KSA
#الملك__ينتصر_لقيادة_المرأة
#SaudiWomenCanDrive
RT @user-10: The only consistent in Saudi history is change. - Saudi FM @AdelAljubeirد#SaudiWomenCanDrive #MyArabia https://t.co/07kNKkNW9Y
RT @user-15: Buckle up, ladies! @user-16 writing on the big news from #SaudiArabia this week https://t.co/PpoXWPnogC #SaudiWomenC…
RT @user-13: Now we all know what it meant. @user-14 #الشعب_يرفض_قياده_المراه #SaudiWomenDriving https://t.co/Tan1xH9GHp
RT @FordMiddleEast: Hi @user-11, we'd like to give you your dream car. #MustangSahar #SaudiWomenCanDrive https://t.co/Tln8aiWUNU
RT @user-17: The day after #KSA announced that #SaudiWomenCanDrive, user-18 attracts young female professionals, to fuel next wave of growth….
RT @user-12: Let's not forget who fought to make this possible @manal_alsharif #الملك__ينتصر_لقيادة_المرأة #SaudiWomenDriving https:…
RT @FordMiddleEast: Hi @user-11, we'd like to give you your dream car. #MustangSahar #SaudiWomenCanDrive https://t.co/Tln8aiWUNU

Fig. 3. Hashtag Network Extracted from the Sample Tweets Shown in Table I.

TABLE. III. NETWORKS STATISTICS

	Number of Nodes	Number of Edges
Retweet network	516,248	1,356,538
Mention network	520,670	1,384,388
Co-Mention network	5,875	7,536
Hashtag network	14,305	59,638

There are several ways to analyze Twitter content: social network metrics based on graph theory [18] [20] and machine learning models [39] [40]. Literature reported several types of community detection algorithms: discordant algorithms that detect inter-community links and remove them from the network [16], recursive algorithms that agglomerate similar communities recursively [17][41] and optimization-based algorithms that maximize of an objective function [42].

TABLE. IV. MOST COMMONLY USED SNA METRICS

Metric	Description	Objective
Degree Power Law Diameter	The degree of a node is the number of edges that are adjacent to the node. The Degree Power Law measures how closely the degree distribution of a network follows a power-law scale. The diameter metric measures the maximal distance between all pairs of nodes.	Node to node distance functions
Average Clustering Coefficient	The clustering coefficient when applied to a single node, is a measure of how complete the neighborhood of a node is. When applied to an entire network, it is the average clustering coefficient over all of the nodes in the network.	The clustering coefficient, along with the mean shortest path, can indicate a "small-world" effect.
Average Path Length	The average of shortest path lengths between all pairs of nodes.	How dense and lengthy is the graph. How long are chains of communications?
Eigenvector Centrality	A measure of node importance in a network based on a node's inwards connections.	Node centrality measure

IV. RESULTS

A. Influential People in the Discussion

This section answers the following research question.

RQ: How to identify influential people in the discussion?

It is crucial to consider the right measures to find out which are the most critical nodes taking into consideration the multitude of tweets and users in this conversation. Social network analysis suggests multiple approaches regarding this matter. To answer the research question, we used degree centrality in the retweet, mention, and the co-mention networks.

1) Retweet network: The retweet network is a directed, weighted graph where nodes represent Twitter users, while edges are formed whenever a retweet occurred. The direction of the link reflects the retweet mechanism, with the direction pointing to the user that was retweeted. Table V and Fig. 4 depict this network. Before constructing the retweet network, the number of followers was used to determine the importance of the node. A plausible hypothesis might be that a Twitter account with 100,000 followers is highly influential. However, if only 1,000 users retweet their tweets, then this user might be not so influential as a user with 20,000 followers but 7,000 retweets his/her message. The higher rate of retweeting suggests that the second user is better nested and more active in terms of Twitter communication.

The in-degree interval of this network falls between 0 and 4,430, and the out-degree is between 0 and 1,066. As the degree centrality measure suggests, we can first focus on the nodes with the highest in-degree (the user who has been retweeted the most) and highest out-degree (the user who retweets the most). Table VI shows the top ten in-degree users.

The data in Table VI is summarized with the highest in-degree users: @SaudiNews50 and @AjelNews24: the certified accounts of well-known news sources in Saudi Arabia. @Algassabinasser: the official account of Nasser Al Gassabi, a Saudi actor. @abdulrahman: Abdulrahman bin Musa'id bin Abdul Aziz is a Saudi Arabian-French businessman. @FordMiddleEast: official account for Ford in the Middle East. @AzzamAlDakhil: the official account of Azzam Al Dakhil, Minister of Education of Saudi Arabia, since 2015. @AmeerahAltaweel: The official account of Ameera Al-Taweel Al-Otaibi, a Saudi princess and philanthropist. Results showed that the highest in-degree user is SaudiNews50 (44,330), which is followed by more than 10 million followers. Other prominent nodes with high in-degree are algassabinasser, with an in-degree of 22,827 and more than 1 million followers, and abdulrahman, with an in-degree of 15,891 and more than 7 million followers.

The analysis of Table VI shows that the number of followers is not the only factor contributing to a user being retweeted because there are nodes with significantly fewer followers that received many retweets. It is also notable that the top in-degree nodes retweeted others rarely, somewhat marginally. Therefore, users suggest that they act more as authority and information generators. These accounts represent the most influential nodes in the retweet network, as they have a huge follower base, and their posts are the most retweeted in the network. Another layer of the data is revealed through the top out-degree nodes. By definition, these would be the participants that retweeted most, which means they acted as transmitters of messages produced by others. Fig. 5 and 6 shows the in-and out-degree distribution of this network, which follows the power-law distribution.

2) Mention network: The mention network is constructed by observing when a user mentions another user. The direction of the link is towards the mentioned user. Being a directed network, which reveals the in-degree versus out-degree statistics of the network. The mention network gives a clue of who might have more significant influence. Table VII and Fig. 7 represent this network.

In and out-degree of this network reflect important people in the discussion. Given that the mention network also consists of a large number of nodes, hence it requires to filter the network. If we focus on the meaning of this representation of discourse, we can conclude that the higher the edge weight, the more semantic value it has, considering that the most mentions/retweets occur between the two accounts involved. Similarly, nodes with lower in-degree value are contributing less to our analysis, since a small number of accounts mentions them. Consequently, we decided to exclude the nodes and edges of lower semantic or analytical significance from our graph model. Using subjective judgment, we set the threshold to keep edge weights and in-degrees higher than 2. Table VIII shows the statistics of the network before and after filtering, while Fig. 8 shows the filtered mention network. Fig. 9 and 10 show the in-degree and out-degree distribution of this network, which follow a power law distribution, similar to the retweet network. Fig. 11 shows the prominent users in the mention network.

As shown in Table IX, the list of discussion participants with the high in-degree measure is substantially different from that of the ones with high out-degree (as it was with the retweet network). This difference is because, in the context of Twitter communication, high in-degree means that more observers are following and mentioning the specific member's tweets, hence this member has more influence on others. On the other hand, high out-degree measure means that a user is mentioning a large number of accounts, thus only consuming information and not creating it or influencing others in a significant way, which is similar to the finding in [13]. The most distinctive feature of the two groups is that the high in-degree group is formed by news channels and public figures, while individual users form the high out-degree. Here is a summary of them: @SaudiNews50, @Sabqorg, @AjelNews24, and @Alekhbariyatv: these are all news accounts. @Sabqorg: the official account of the electronic magazine Sabq. @Alekhbariyatv: the official account of the Saudi channel Al Ekhbariya. @KingSalman: the official account of King Salman. @SaudiNews50, @Sabqorg, @AjelNews24, and @Alekhbariyatv: these are all news accounts. @Sabqorg: the official account of the electronic magazine Sabq. @Alekhbariyatv: the official account of the Saudi channel Al Ekhbariya. @KingSalman: the official account of King Salman. @Algassabinasser: the official account of Nasser Al Gassabi, a Saudi actor.

In order to reveal the opinion-makers and to filter users, we filtered the graph by in-degree and number of followers. Fig. 11 shows the users with high in-degree and the highest number of followers. The dark red nodes have high number of followers, and the size refers to in-degree. There are many well-known accounts for tweeting and spreading the latest and important news, such as: @SaudiNews50, one of the major players; @AjelNews24, with fewer followers than @SaudiNews50; and @SabqOrg, with the highest number of followers but less in-degree than the others.

TABLE. V. RETWEET-NETWORK STATISTICS

Retweet-Network	
Number of nodes	516,248
Number of edges	1,356,538
Average degree	2.628
Avg. weighted network	3.131
Network diameter	25
Modularity	0.71
Number of communities	1,033

Fig. 4. Retweet Network

TABLE. VI. DATA LABORATORY OF THE HIGHEST IN-DEGREE USERS IN THE RETWEET-NETWORK

Label	User_name	In-degree	Out-degree	Followers count
SaudiNews50	أخبار السعودية	44,330	3	44,333
algassabinasser	ناصر القصبي	22,827	1	22,828
abdulrahman	عبدالرحمن بن مساعد	15,891	5	15,896
FordMiddleEast	Ford Middle East	13,214	-	13,214
AjelNews24	خبر عاجل	12,283	3	12,286
AzzamAlDakhil	عزام الدخيّل	10,033	2	10,035
Al_khalden8	الخالدي	9,568	-	9,568
AmeerahAltaweeL	أميرة الطويل	8,796	3	8,799
badughaish	Faisal BaDughaish	7,838	1	7,839
faare8	فارس الهلال	7,498	1	7,499

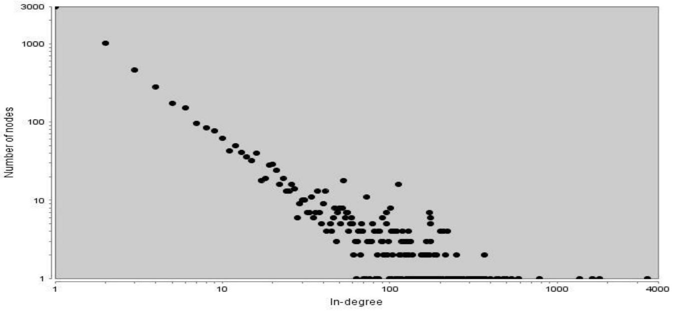


Fig. 5. In-Degree Plot of the Retweet Network.

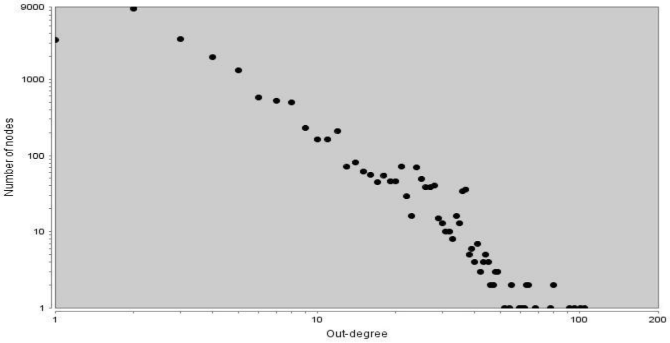


Fig. 6. Out-Degree Plot of the Retweet Network.

TABLE. VII. MENTION NETWORK STATISTICS

Mention-Network	
Number of nodes	520,670
Number of edges	7,536
Average degree	2.565
Avg. weighted network	28.823
Network diameter	17
Modularity	0.808
Number of communities	1,268

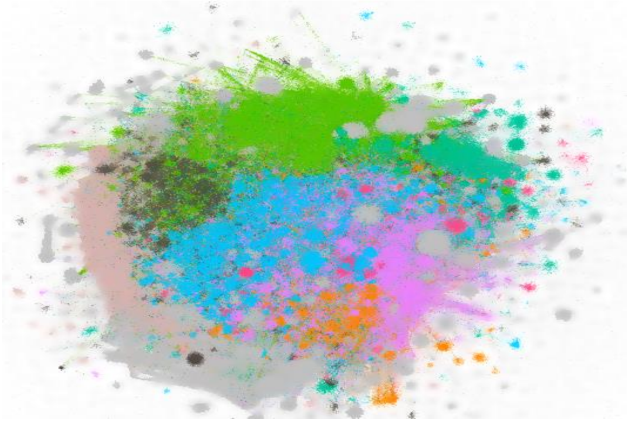


Fig. 7. Mention Network.

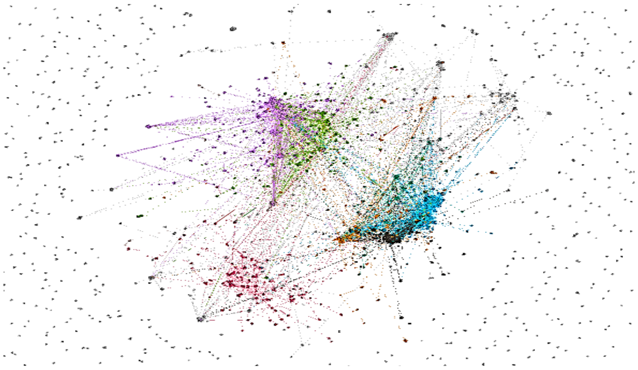


Fig. 8. Mention-Network Graph after Filtering.

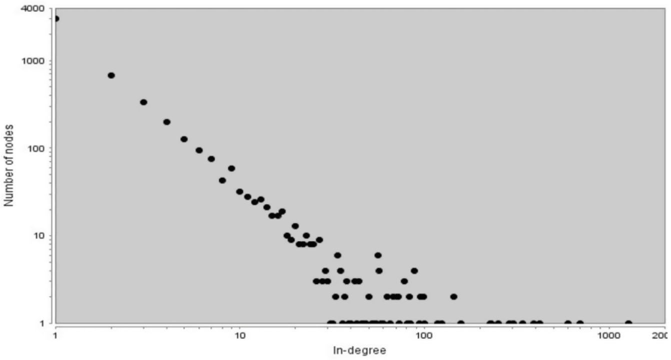


Fig. 9. In-Degree Plot of the Mention Network.

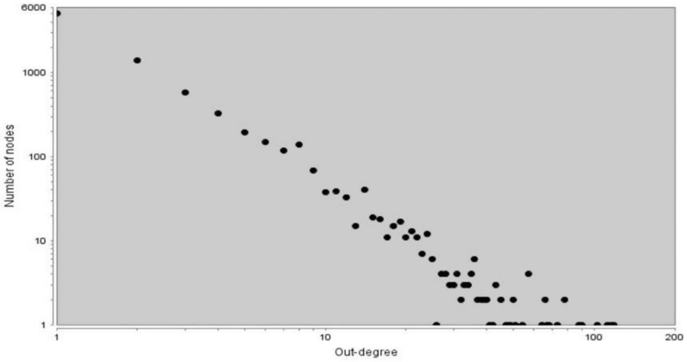


Fig. 10. Out-Degree Plot of the Mention Network.

TABLE. VIII. FILTERED MENTION NETWORK

Mention-Network – Original		**Mention-Network – Filtered**	
Average degree	2.565	Average degree	3.666
Avg. weighted network	2.596	Avg. weighted network	1.971
Network diameter	28	Network diameter	22
Graph Density	0	Graph Density	0.0
Modularity	0.712	Modularity	0.731
PageRank	0.85	PageRank	0.85

TABLE. IX. IMPORTANT NODES IN THE MENTION NETWORK

Top in-degree nodes			**Top out-degree nodes**		
#	***Label***	***In-degree***	**#**	***Label***	***Out-degree***
1	SaudiNews50	1274	1	alkinha505	119
2	Algassabinasser	697	2	Asss2019	116
3	Abdulrahman	600	3	mode_smart2030	112
4	Sabqorg	419	4	Waeel05550Waeel	103
5	AjelNews24	393	5	sltan_656	90
6	KingSalman	342	6	asd_sunah	88
7	KSA_620	303	7	Aaaa06259528	78
8	hosbah_tweet	289	8	asd025056	78
9	Alekhbariyatv	251	9	TqU6lSOGEFjRJcu	73
10	faare8	229	10	l680352	68

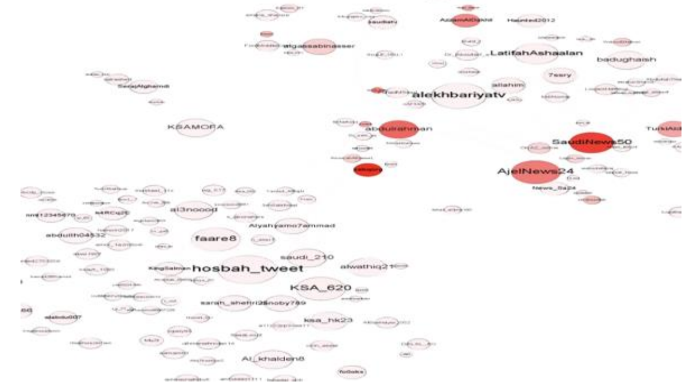


Fig. 11. Network of the users with a High in-Degree and Highest Number of Followers in Mention Network.

We could say that these drives and, to a certain extent, shape the whole conversation by being the top Arabic news sources on Twitter. The distribution of the users varies due to location and language on Twitter. Most of the users do not mention their locations. Also, the language used can be either Arabic or English, which adds to the disparity in the distribution of the users. From Fig. 12, we can see that most of the users do not state a location (indicated by the color purple). Others are from Saudi Arabia and England.

In summary, there are many similarities regarding the leading players in the degree centrality of the retweet and Mention networks. These results are because the dataset is the same. So, even though the links are formed differently, the leading players re-appear.

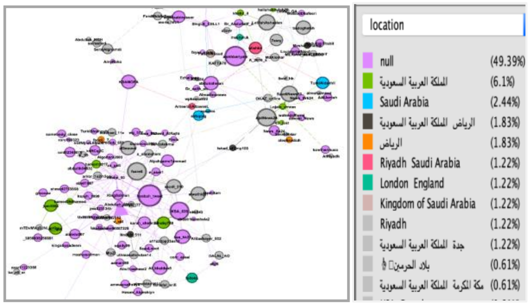

location
null (49.39%)
المملكة العربية السعودية (6.1%)
Saudi Arabia (2.44%)
الرياض المملكة العربية السعودية (1.83%)
الرياض (1.83%)
Riyadh Saudi Arabia (1.22%)
London England (1.22%)
Kingdom of Saudi Arabia (1.22%)
Riyadh (1.22%)
جدة المملكة العربية السعودية (1.22%)
بلاد الحرمين (0.61%)
مكة المكرمة المملكة العربية السعودية (0.61%)

Fig. 12. Location of the users in the Mention Network.

3) Co-mention network: The co-mention network is created based on the fact that users are mentioned together in tweets. For example: if @user1 and @user2 are both mentioned in a tweet, they will be connected through an edge in the network. The co-users appearing together does not represent information flow, and this determines this network as undirected. Table X and Fig. 13 represent the co-mention network.

Table XI shows the top ten users as follows:@KingSalman, the official account of the King of Saudi Arabia. @LoujainHathloul is a Saudi women's rights activist, a social media figure. @protectmax is a corporate account that announced a prize for users retweeting tweets with specific hashtags related to Women Driving, so, logically, the company received many mentions. @algassabinasser, or Nasser Algassabi is a famous Saudi actor. @SaudiNews50 and @KSA24 are personal accounts for spreading news, but they are highly prevalent in Saudi Arabia. @MOISaudiArabia is the official account of the Ministry of the Interior in Saudi Arabia. @ssa_at is the official account of the Council of Senior Scholars. @abdulrahman: Abdulrahman bin Musa'id bin Abdul Aziz is a Saudi Arabian-French businessman. @manal_alsharif: Manal al-Sharif is a Saudi Arabian women's rights activist who helped start a women's right to drive campaign in 2011.

It is clear from Fig. 14 that @KingSalman has edges with many other top accounts. Two of them are the official accounts @MOISaudiArabia and @ssa_at. These form a triangle. We can also observe that @manal_alsharif and @LoujainHathloul have a thick edge between them because they are considered endorsers of the decision, and it is highly probable that users mention both of them in their tweets. @protectmax does not have an edge with any of these most prominent accounts, but it is still one of the top users because of the interest the company triggered by encouraging users to cite particular hashtags about women driving.

TABLE. X. CO-MENTION NETWORK STATISTICS

Co-Mention network	
Number of nodes	5,875
Number of edges	1,384,388
Average degree	2.659
Avg. weighted network	3.131
Network diameter	28

TABLE. XI. THE TOP TEN DEGREE USERS IN THE CO-MENTION-NETWORK

Label	User_name	Degree	Count
KingSalman	ع بدال عزيز ز ب ن سلمان	286	5,155
LoujainHathloul	الهذلول هذلول لجين	111	687
protectmax	ماكس ب روت كت شركة	99	361
algassabinasser	ال قصبي ناصر	87	1,031
SaudiNews50	ال سعودية أخ بار	70	718
manal_alsharif	ال شريف مسعود منال	70	658
MOISaudiArabia	ال داخ ل ية وزارة	68	449
KSA24	SA الأخ بار موجز	66	122
ssa_at	ال ع لماء ك بار هيئة	66	519
abdulrahman	مساعد ب ن ع بدال رحمن	57	521

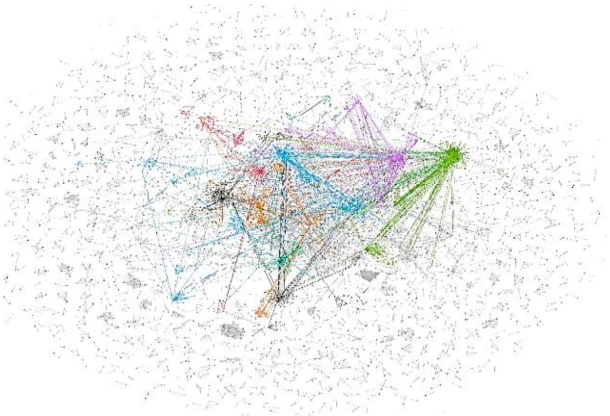


Fig. 13. Co-Mention Network.

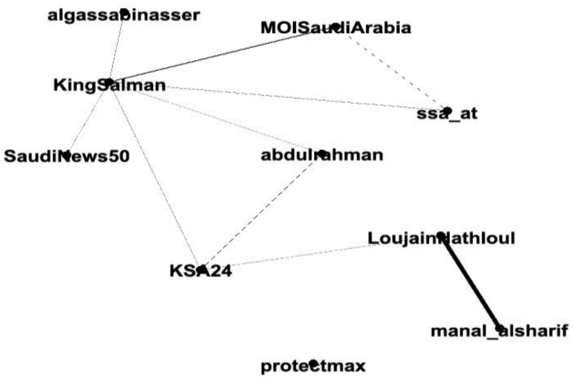


Fig. 14. Top Ten users in the Co-Mention Network.

V. DISCUSSION

The most influential people in the discussion are listed in Table XII and Table XIII shows the top 5 users in the Mention, Retweet, and Co-mention networks:

The most critical users in mention and Retweet network are SaudiNews50, Sabqorg, AjelNews24, and Alekhbariyatv, who are news accounts. While KingSalman's account (the King of Saudi Arabia) has the highest degree in the co-mention network since his account was mentioned along with other users to discuss the topic of women driving decision, namely Loujain Hathloul. Loujain Hathloul is one of the biggest supporters of Women Driving in Saudi Arabia, which makes her one of the top co-mentioned accounts. Similarity, algassabinasser, presumably got many mentions due to his popularity and to his stance in favor of allowing Saudi women to drive.

TABLE. XII. MOST SIGNIFICANT PEOPLE IN THE STUDIED DATASET

Account owner	Accounts
The official account of King Salman.	@KingSalman
The official account of the news TV channel	@AlArabiya_Brk , @Alekhbariyatv
certified accounts of well-known news sources in Saudi Arabia	@SaudiNews50 , @AjelNews24, and @KSA24
Islamic scholar	@Dr_alqarnee, @MohamadAlarefe and @Dr_almosleh
Among the first supporters of the “Women Driving” decision, even before the decision’s official declaration.	@LoujainHathloul, @manal_alsharif

TABLE. XIII. TOP FIVE USERS IN ALL NETWORKS

Top users in Mention-Network	Top users in Retweet-Network	Top users in Co-Mention-Network
SaudiNews50	SaudiNews50	KingSalman
Algassabinasser	Algassabinasser	LoujainHathloul
Abdulrahman	Abdulrahman	protectmax
Sabqorg	AjelNews24	algassabinasser
AjelNews24	Alekhbariyatv	manal_alsharif

However, there are some differences between all the three networks. For example – SaudiNews50 is among the top users in the mention and retweet networks, which means that many users referred to it as a news source by either for sharing the decision or proving a point. In this sense – both retweet and mention networks reveal the most important news outlets or information disseminators. However, the co-mention network is more targeted towards connected persons, as in its top users, there are no news outlets. Therefore, we can say that when it comes to sharing of information, retweet and mention networks tend to provide the most influential persons, including news sources. Whereas influential people closely associated with the discussions can be identified through the co-mention network, it may exclude popular users like news dissemination accounts. Results showed that @SaudiNews50, @Algassabinasser, and @Abdulrahman were top users in both retweet and co-mention networks, while @KingSalman and @LoujainHathloul were the top two users in the co-mention network. Consequently, KingSalman and LoujainHathloul were identified as influencers, whereas SaudiNews and Algassabinasser were the leading distributors of the news.

VI. CONCLUSIONS AND RECOMMENDATIONS

This research analyzed a dataset of two million tweets discussing the topic of women driving cars in Saudi Arabia. Using a large dataset of #SaudiWomenCanDrive, this research has proposed a framework that consists of four networks: hashtag, retweet, mention, co-mention networks. The framework helps in identifying influential people. For example, the framework showed that the account of KingSalman admired people by his decisive decision; therefore, the account appears on top in co-mention network. Moreover, the research identified the collective behavior of people while expressing their opinions on different topics. This research found that people's influence was related to their effect on social media in terms of Twitter network structure as distributors or hubs of news. The proposed framework could provide to analyze similar behavior on other discussions.

References

[1] W. Tan, M. B. Blake, I. Saleh, and S. Dustdar, "Social-network-sourced big data analytics," IEEE Internet Comput., vol. 17, no. 5, pp. 62–69, 2013.

[2] A. Alotaibi, "Why the panic? Gendered moral panics and the Saudi ban on women driving," The George Washington University, 2017.

[3] E. Alhussein, "Triangle of change: the situation of women in Saudi Arabia," Exec. Summ., 2014.

[4] I. Chaudhry, "Arab Revolutions: Breaking fear|# hashtags for change: Can Twitter generate social progress in Saudi Arabia," Int. J. Commun., vol. 8, p. 19, 2014.

[5] M. Boukes, "Social network sites and acquiring current affairs knowledge: The impact of Twitter and Facebook usage on learning about the news," J. Inf. Technol. Polit., vol. 16, no. 1, pp. 36–51, 2019.

[6] D. Knoke and S. Yang, Social network analysis, vol. 154. SAGE Publications, Incorporated, 2019.

[7] L. Vega and A. Mendez-Vazquez, "Detecting of topic-specific leaders in social networks," Procedia Comput. Sci., vol. 151, pp. 1188–1193, 2019.

[8] J. A. Barnes and F. Harary, "Graph theory in network analysis," 1983.

[9] L. C. Freeman, "A set of measures of centrality based on betweenness," Sociometry, pp. 35–41, 1977.

[10] J. A. Morente-Molinera, G. Kou, K. Samuylov, R. Ureña, and E. Herrera-Viedma, "Carrying out consensual Group Decision Making processes under social networks using sentiment analysis over comparative expressions," Knowledge-Based Syst., vol. 165, pp. 335–345, 2019.

[11] A. LFarasani, T. AlHarthi, and S. AlHumoud, "ATAM: Arabic Traffic Analysis Model for Twitter," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 3, 2019.

[12] D. Kuroshima and T. Tian, "Detecting Public Sentiment of Medicine by Mining Twitter Data," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 10, 2019.

[13] H. Al-Qaheri and S. Banerjee, "Measuring Homophily in Social Network: Identification of Flow of Inspiring Influence under New Vistas of Evolutionary Dynamics," in International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Extended Papers from Science and Information Conference, 2013.

[14] R. A. Abbasi et al., "Saving lives using social media: Analysis of the role of twitter for personal blood donation requests and dissemination," Telemat. Informatics, vol. 35, no. 4, 2018.

[15] B. Alotaibi, R. A. Abbasi, M. A. Aslam, K. Saeedi, and D. Alahmadi, "Startup Initiative Response Analysis (SIRA) Framework for Analyzing Startup Initiatives on Twitter," IEEE Access, vol. 8, pp. 10718–10730, 2020.

[16] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," Proc. Natl. Acad. Sci., vol. 101, no. 9, pp. 2658–2663, 2004.

[17] P. Pons and M. Latapy, "Computing communities in large networks using random walks.," J. Graph Algorithms Appl., vol. 10, no. 2, pp. 191–218, 2006.

[18] M. L. Litchman, D. Lewis, L. A. Kelly, and P. M. Gee, "Twitter analysis of# OpenAPS DIY artificial pancreas technology use suggests improved A1C and quality of life," J. Diabetes Sci. Technol., vol. 13, no. 2, pp. 164–170, 2019.

[19] C. V Talbot, S. T. O'Dwyer, L. Clare, J. Heaton, and J. Anderson, "How people with dementia use twitter: A qualitative analysis," Comput. Human Behav., vol. 102, pp. 112–119, 2020.

[20] J. Rosenberg, "Exploring the Next Generation Science Standards Chat (# NGSSchat) Professional Network on Twitter Through Social Network Analysis," 2019.

[21] B. J. Mills, "Social Network Analysis in Archaeology," Annu. Rev. Anthropol., vol. 46, no. 1, pp. 379–397, Oct. 2017.

[22] F. Haneef et al., "Using network science to understand the link between subjects and professions," Comput. Human Behav., vol. 106, 2020.

[23] R. Isba, K. Woolf, and R. Hanneman, "Social network analysis in medical education," Med. Educ., vol. 51, no. 1, pp. 81–88, Jan. 2017.

[24] N. E. Friedkin, A. V. Proskurnikov, R. Tempo, and S. E. Parsegov, "Network science on belief system dynamics under logic constraints," Science (80-.)., vol. 354, no. 6310, pp. 321–326, Oct. 2016.

[25] N. Alduaiji, A. Datta, and J. Li, "Influence Propagation Model for Clique-Based Community Detection in Social Networks," IEEE Trans. Comput. Soc. Syst., vol. 5, no. 2, pp. 563–575, Jun. 2018.

[26] Z. Saeed, R. A. Abbasi, I. Razzak, O. Maqbool, A. Sadaf, and G. Xu, "Enhanced Heartbeat Graph for emerging event detection on Twitter using time series networks," Expert Syst. Appl., vol. 136, pp. 115–132, 2019.

[27] Z. Saeed et al., "What's Happening Around the World? A Survey and Framework on Event Detection Techniques on Twitter," J. Grid Comput., vol. 17, no. 2, pp. 279–312, Jun. 2019.

[28] Z. Saeed, R. A. Abbasi, A. Sadaf, M. I. Razzak, and G. Xu, Text stream to temporal network - A dynamic heartbeat graph to detect emerging events on twitter, vol. 10938 LNAI. Springer Verlag, 2018, pp. 534–545.

[29] Z. Saeed et al., "Event Detection in Twitter Stream Using Weighted Dynamic Heartbeat Graph Approach," IEEE Comput. Intell. Mag., vol. 14, no. 3, pp. 29–38, Aug. 2019.

[30] A. Said, T. D. Bowman, R. A. Abbasi, N. R. Aljohani, S.-U. Hassan, and R. Nawaz, "Mining network-level properties of Twitter altmetrics data," Scientometrics, Apr. 2019.

[31] J. Kim and M. Hastak, "Social network analysis: Characteristics of online social networks after a disaster," Int. J. Inf. Manage., vol. 38, no. 1, pp. 86–96, Feb. 2018.

[32] A. Said, R. A. Abbasi, O. Maqbool, A. Daud, and N. R. Aljohani, "CC-GA: A clustering coefficient based genetic algorithm for detecting communities in social networks," Appl. Soft Comput. J., vol. 63, 2018.

[33] C. S. Q. Siew and M. S. Vitevitch, "The phonographic language network: Using network science to investigate the phonological and orthographic similarity structure of language," J. Exp. Psychol., vol. 148, no. 3, pp. 475–500, 2019.

[34] A. Fronzetti Colladon and E. Remondi, "Using social network analysis to prevent money laundering," Expert Syst. Appl., vol. 67, pp. 49–58, Jan. 2017.

[35] C. Casanueva, Á. Gallego, and M.-R. García-Sánchez, "Social network analysis in tourism," Curr. Issues Tour., vol. 19, no. 12, pp. 1190–1209, Oct. 2016.

[36] J. Kanwal, O. Maqbool, R. Abbasi, and A. Q. A. Q. Abbasi, "Network analysis of software change history for understanding software evolution," in 17th IEEE International Multi Topic Conference 2014, 2014, pp. 229–234.

[37] U. Brandes, "A faster algorithm for betweenness centrality," J. Math. Sociol., vol. 25, no. 2, pp. 163–177, 2001.

[38] F. Imran, R. A. Abbasi, M. A. Sindhu, A. S. Khattak, A. Daud, and T. Amjad, "Finding research areas of academicians using clique percolation," in 2018 14th International Conference on Emerging Technologies, ICET 2018, 2019.

[39] M. Congosto, P. Basanta-Val, and L. Sanchez-Fernandez, "T-Hoarder: A framework to process Twitter data streams," J. Netw. Comput. Appl., vol. 83, pp. 28–39, 2017.

[40] D. M. Best, J. Bruce, S. Dowson, O. Love, and L. McGrath, "Web-based visual analytics for social media," in Sixth International AAAI Conference on Weblogs and Social Media, 2012.

[41] Z. Liu and Y. Ma, "A divide and agglomerate algorithm for community detection in social networks," Inf. Sci. (Ny)., vol. 482, pp. 321–333, 2019.

[42] S. Rahimi, A. Abdollahpouri, and P. Moradi, "A multi-objective particle swarm optimization algorithm for community detection in complex networks," Swarm Evol. Comput., vol. 39, pp. 297–309, 2018.

Cross-Language Plagiarism Detection using Word Embedding and Inverse Document Frequency (IDF)

Hanan Aljuaid
Computer Sciences Department, College of Computer and Information Sciences
Princess Nourah bint Abdulrahman University (PNU), 84428 Saudi Arabia, Riyadh

Abstract—**The purpose of cross-language textual similarity detection is to approximate the similarity of two textual units in different languages. This paper embeds the distributed representation of words in cross-language textual similarity detection using word embedding and IDF. The paper introduces a novel cross-language plagiarism detection approach constructed with the distributed representation of words in sentences. To improve the textual similarity of the approach, a novel method is used called CL-CTS-CBOW. Consequently, adding the syntax feature to the approach is improved by a novel method called CL-WES. Afterward, the approach is improved by the IDF weighting method. The corpora used in this study are four Arabic-English corpora, specifically books, Wikipedia, EAPCOUNT, and MultiUN, which have more than 10,017,106 sentences and uses with supported parallel and comparable assemblages. The proposed method in this paper combines different methods to confirm their complementarity. In the experiment, the proposed system obtains 88% English-Arabic similarity detection at the word level and 82.75% at the sentence level with various corpora.**

Keywords—NLP; cross-language plagiarism detection; word embedding; similarity detection; IDF

I. Introduction

Plagiarism is a major problem today. Cross-lingual plagiarism (CLP) is a type of plagiarism that occurs when texts are translated from one language to another without citing the original sources. Monolingual plagiarism analysis, which detects plagiarism in documents written in the same language, has been executed by many researchers, but CLP remains a challenge. Earlier studies have used approaches such as cross-lingual explicit semantic analysis (CL-ESA), syntactic alignment using character N-grams (CL-CNG), dictionaries and thesauruses, statistical machine translation, online machine translators [1] [6], and more recently, semantic networks and word embedding [7]. However, these approaches are specific to bilingual plagiarism detection tasks and are normally not sufficient for limited resource languages.

Conversely, word embedding is a significant representation theory used to represent sentence units used in natural language processing (NLP) applications [15]. This process depends on the low-dimensional vector representation of words, and it can easily measure the syntax vs. semantic relationship. Currently, a variety of NLP applications are contingent on two-word embedding models: the word2vec model [12] and the GloVe model [17]. The word2vec model is a neural network that includes three layers: one input layer, one output layer and one hidden layer. However, the GloVe word embedding model uses a global vector for word representation [21].

In this paper, we explore the performance of the distributed representation of word embedding to propose novel cross-lingual similarity procedures for similarity detection. We use word embeddings with the IDF weighting method.

II. Related Work

Word embedding is used in natural language processing as a representation of the vocabulary of a document. This method depends on identifying the context of a word (syntactic and semantic similarities) relative to other words using vector representation and involves two models: the word2vec and GloVe models. Recently, these two-word embeddings models have been used in various natural language processing applications [21].

However, this processing starts by converting words into vectors. Consequently, the cosine similarity is used to measure the semantic similarity between two words [13]. The previous method for representing a word vector was a “one-hot” representation, where the number of dimensions of each vector is matched to the number of dimensions of the vocabulary. Modern word embeddings are accessible for the study of semantic and syntax similarities.

Word2vec is one type of neural network with three layers: an input layer, hidden layer, and output layer. The number of dimensions of the vector that represents a word is the same as the number of neurons in the hidden layer. Typically, the word2vec model applies big datasets in the training phase to optimize the syntax and semantics correctly. Word2vec mathematically detects similarities to cluster the vectors of similar words together in vector space. The created vectors detect the word features by distributed arithmetic representations without human mediation. Additionally, using the given data, word2vec can determine highly accurate solutions about a word’s meaning based on past sentences. Those solutions can be used to launch a word’s connection with other words or cluster documents and classify them by topic (for example, “man” is to “boy”, and “woman” is to “girl”). In addition, those clusters can be used in a sentiment analysis, where each item in the vocabulary has a vector attached to it and can be fed into a deep-learning networked or analysed to discover the relations between words.

The main approaches of word2vec are the skip-gram model and the bag-of-words model (BOW), and both of these models have achieved developments in computational cost and accuracy. In these two approaches, the same hyperparameters are used, such as the window size denoted by C and the vocabulary size (represents the number of words in the corpus) denoted by |v|. In the next paragraph, these two approaches are explained briefly.

Conversely, the continuous bag-of-words technique (CBOW) inputs the context of each word using a linear classifier and predicts the middle word corresponding to the adjacent features in that context [10][21]. The deeper analysis of CBOW can show that the input words comprise a one-hot encoded CxV dimension matrix of the context words, and the output layer comprises a vector with the elements being the softmax values of V length; the hidden layer contains N neurons and takes an average over all the C context input words, as shown in Fig. 1.

The continuous skip-gram approach or skip-gram technique (the second approach of the word2vec model) is very similar to the CBOW model. However, the difference between the two approaches exists in the input and output layers. The input in CBOW is the context words, and the output is the middle word, whereas the opposite occurs in the skip-gram model, where the input is the present word, and the output is the context words.

Fig. 2 shows that the skip-gram model has three layers. The input layer includes the input vector with length V for only one word. The hidden layer has the same definition as it does in the CBOW model, where h in formula (1) denotes the relationship between the input and hidden layers, i.e., h is simply transposed onto a row with two layers with weight matrix, W, which is supplementary to the input word wI:

h=WT:=vT, (1) (k,·) wI (1)

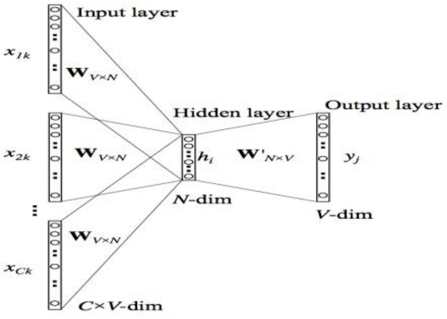


Fig. 1. CBOW Model Architecture [19]; [10].

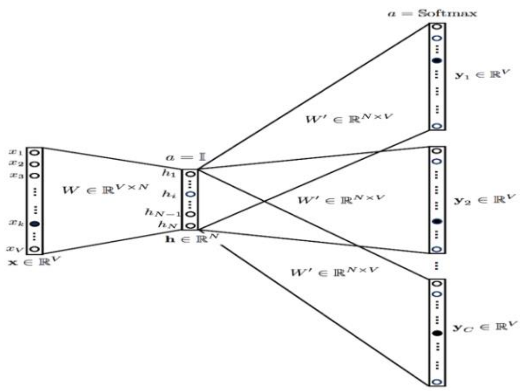

a = Softmax
$y_1 \in \mathbb{R}^V$
$W' \in \mathbb{R}^{N\times V}$
x_1
x_2
x_3
x_k
x_V
$\mathbf{x} \in \mathbb{R}^V$
$W \in \mathbb{R}^{V\times N}$
$a = \mathbb{I}$
h_1
h_i
h_{N-1}
h_N
$\mathbf{h} \in \mathbb{R}^N$
$W' \in \mathbb{R}^{N\times V}$
$y_2 \in \mathbb{R}^V$
$W' \in \mathbb{R}^{N\times V}$
$y_C \in \mathbb{R}^V$

Fig. 2. Skip-Gram Model Architecture [19].

For the output layer of the model outputting C probability distributions, each context position has C probability distributions with V probabilities (one for each word) [19].

The skip-gram model is efficient when training small datasets with irregular words. However, the CBOW model is proficient when used with common words [15]. Moreover, the considerable challenge with both word2vec representations is learning the output vectors. To appropriately learn the output vectors, the proposed hierarchical softmax and negative sampling algorithms can be used [13]. The first algorithm (hierarchical softmax) is centred on the Huffman tree (a binary tree), which uses word frequencies to estimate the words in a tree. Then, the algorithm uses normalization in each step from the root to the target word [15]. The second algorithm, negative sampling, targets the noise distribution to update the samples of the output vectors. Correspondingly, negative sampling is used in the case of low-dimension vectors with more common words, whereas hierarchical softmax is used in the case of irregular words.

III. Preprocessing Management

A. Dataset

The dataset used throughout our study is the new dataset familiarized by Aljuaid [2]. The characteristics of this dataset are as follows:

- written in English and Arabic;
- united at different levels (the document, sentence, and word chunk levels);
- uses supported parallel and comparable assemblages;
- conceals several subjects;
- translates automatically or by humans, regardless of whether the translations are performed by professionals;
- collected from more than 3,000 random documents that were checked manually.

Table I shows the details of the dataset and presents the number of aligned units. Table II presents the different characteristics of the dataset within each corpus.

B. Outline of State-of-the-Art Methods

Cross-language plagiarism estimates the textual similarity between two languages in two textual units. In this section, the state-of-the-art methods that are used in this paper are discussed.

TABLE. I. Corpora Description of our Dataset

Corpus	Language	#document	# sentences	# word chunks
Books	English/ Arabic	≈ 6,000	≈ 120,000	≈ 720,000,0
Wikipedia	English/ Arabic	≈ 10,000	≈ 800,000	≈ 480,000,00
EAPCOUNT	English/ Arabic	≈341	≈ 53,000	≈ 5,392,491
MultiUN	English/ Arabic	≈1659	≈1,124,609	≈ 300,000,000

TABLE. II. CORPORA CHARACTERISTICS OF OUR DATASET

corpus	Alignment	Written by	Translated by
Books	Parallel	Computer scientists	Professional translators
Wikipedia	Comparable	Anyone	Student translators
EAPCOUNT	Parallel	Politicians	Machine translated
MultiUN	Parallel	Politicians	Machine translated

Cross-language character n-gram (CL-CnG) is dependent on the comparison of dual textual units according to their n-gram vectors based on the [11].

Cross-language conceptual thesaurus-based similarity (CL-CTS) is used to extract the roots of the textual units to measure the semantics of the words [16].

Cross-language alignment-based similarity analysis (CL-ASA) is used as a bilingual unigram dictionary to determine the ability of one textual unit to translate to another textual unit and their probabilities extracted from a parallel corpus [18].

Cross-language explicit semantic analysis (CL-ESA) denotes the meaning of a document by a vector based on concepts derived from Wikipedia according to the explicit semantic analysis [8].

Translation + monolingual analysis (T+MA) involves translating elements in two different languages into the same language to perform monolingual identification among the elements [3]. This state-of-the-art method is discussed in depth in our previous paper [2].

IV. PROPOSED METHODS

A. Model used

The word embedding representation is achieved and is compatible with the corpus context. Words with similar contexts should be projected onto a continuous multidimensional space. However, word embedding can be used to detect and calculate similarities between sentences in the same or different languages.

Consequently, we used the word2vec CBOW approach toolkit offered by MultiVec [4]. To build and train the vectors, we use the large collection corpus discussed in [2].

To train the CBOW embedding system, some parameters are selected to affect the resulting vectors. The selected parameter has a vector size of 100 with a window size of 5, and a number of negative examples in training 10 are shown in Table III.

TABLE. III. THE ARABIC CBOW MODEL PARAMETERS FOR TRAINING THE CONFIGURATION PARAMETERS

Parameter	Significance
Window	5
Vector size	100
Negative	10
Sample	1e − 5
Frequency threshold	0.02

B. Textual Similarity

We introduce a new method to identify the similarity among textual words. However, the lexical resource in the cross-language conceptual thesaurus-based similarity (CL-CTS) is replaced with the distributed representation of words. To construct the words with the BOW model, we used the CBOW model to detect pairs of two words, wi and wj. Each word is represented by vectors vi and vj, respectively. The similarity between wi and wj is obtained by comparing their vectors vi and vj that were evaluated using cosine similarity. We call this new implementation CL-CTS-CBOW, and this method is used to improve textual similarity.

Then, we implement a method that performs a comparison between two sentences S and S’ in different languages. We call this method CL-WES, which uses the cosine similarity of the embedded vectors of all units among the sentences to represent the distribution of the sentences [6], where S’ = w1,w2...,wi and S” = w1′, w2′,...,wj′, with two textual units U’ and U” in two different languages. Then, CL-WES builds the bilingual corpus of the two different languages. The two representation vectors V’ and V” utilize cosine similarity.

The calculation of the distributed representation V around a textual unit U is:

$$V = \Sigma\_(i=1)\^{}n▒(ui) \qquad (2)$$

where V is the vector of the function that gives the word embedding, and ui is the textual unit. Fig. 3 shows our proposed system.

C. Syntax Similarity

In this section, the CL-WES model is improved by adding the syntax aspect, as discussed in Section 4.2, where U is a textual unit with n words, as shown in formula (1). However, we start by applying the part of speech tagger (POS) to syntactically tag U, which is used to weight every word in the sentence representation, classifying it into its morphosyntactic category. Then, we normalize the tags using the universal tagset [20]. Then, a weight is assigned to each tag according to this formula:

$$V = \sum_{k=1}^{i} \text{Pos weight}(\text{Poswk}) * \text{vk} \qquad (3)$$

where Poswk is the function used to determine the weight of the POS tagging of wk [14].

Moreover, if U_1 and U_2 are two textual units with different languages, their representation vectors V_1 and V_2 are built using formula (4); then, cosine similarity is applied between them.

$$V = \sum_{i=1}^{n} (\text{weight}(\text{pos}(\text{ui})). \text{vector}(\text{ui})) \qquad (4)$$

where the variable weight is a function that determines the weight of a POS, and the variable vector is a function that outputs the word embedding vector.

D. Combining Multiple Methods

To improve our method’s performance in detecting cross-language similarity in English and Arabic languages, we combine our method with the IDF weighting method, where during weight processing, the similarity score of each method is assigned, and the composite score is calculated (weighted), as shown in Fig. 3. The distribution of the weights is optimized with the Bersini method[5]. However, one fold of every corpus is used to train the IDF weights, so the other evaluates the IDF method.

1) IDF weighting method: The IDF method constructs a compound weight of every word in a sentence. The IDF weight operates as a measurement term related to the absolute similarity between documents.

However, the Salton and [9] method is employed, where one fold of each corpus is used as an input to be semantically verified. To compute the *IDF weight* for every word, the other folds in the corpus are used as the background quantity. Moreover, the IDF is calculated with the following formula:

$$\mathrm{idf}(w)=\log\left(\frac{S}{ws}\right) \quad (5)$$

where S is the number of sentences in the corpus written in the two languages of Arabic and English, and WS is the number of sentences containing *w*. Then, the cosine similarity between V1 and V2, cos(V1, V2), in L_1 and L_2 is calculated to obtain the similarity between S1 and S2:

$$\begin{cases} V_1 \sum_{k=1}^{i} idf(w_k)v_k \\ V_2 \sum_{k=1}^{i} idf(w'_k)v'_k \end{cases} \quad (6)$$

where *idf (wk)* is the weight of w*k* in the background.

Regarding the state-of-the-art methods for clustering capacity, the similar and different terms are correctly separated, and their ability to predict a (mis)match is determined. We combine these methods with IDF weighting to reduce uncertainties in the classification and exploit the complementarities of these methods. However, we find that these methods are processed differently according to their features. Some of them are lexical syntax-based, others are semantic-based and process the aligned words, and others capture the context with word vectors.

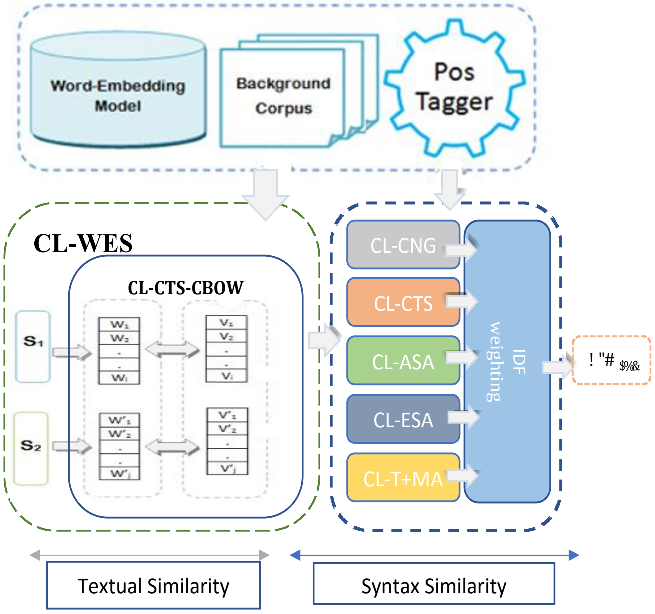


Fig. 3. The Proposed System Architecture.

V. Experiments and Results

A. Evaluation Indicators

To evaluate our method, a distance matrix of size NxM is built, where M=1,000 and N is the evaluated sub-corpus we previously denoted as (S). However, to operate S, every textual unit is matched with its consistent units in the intentioned language (i.e., to detect the similarity in the cross-lingual analysis); in addition, it is compared to M-1, which is a unit randomly selected from S. In the comparison, each obtained matching score leads to the distance matrix. To identify the threshold of the matrix, the best F-score is used and defined as the symmetrical mean of precision and recall, where precision is the number of matches in similar units that is retrieved using all of the matches. All of the methods are applied to the Arabic-English corpus at the word and sentence levels. In every construction, a particular method is applied to the sup-corpus for training and evaluation when considering a particular level. The evaluation folds are supported by varying the M selected units. The formulas for calculating the F-score, precision and recall are shown in formulas (7) -(9), respectively.

$$\text{precision} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (8)$$

$$F = \frac{2\times\text{precision}\times\text{Recall}}{\text{precision}\times\text{Recall}} \quad (9)$$

where *TP* is the number of samples with positive similarity. *TN* is the number of samples with negative similarity. *FP* is the number of samples that have a negative similarity tagged as a positive similarity, and *FN* is the number of samples that have a positive similarity tagged as a negative similarity.

1) Use of word embedding evaluation: The F-score, which presents the distributed representation of words compared with lexical resources, improves the *CL-CTS-WE* performance to 78% at the word level, which is better than the performance of the C*L-CTS* method, which obtains a 59% performance at the word level and 54% performance at the sentence level, as shown in Table IV. However, the use of *CL-WES* improves the performance at the word level to 86%, which is higher than the state-of-the-art method performances, as shown in Fig. 4. Focusing on the state-of-the-art methods, we found that the best performance is from the CL-ASA method at the word and sentence levels, but the overall performance of the method is lower than the CL-WES performance, which is the best single method evaluated.

2) IDF evaluation: The results of the IDF method are recorded at both the word and sentence levels in Table IV and Fig. 5. In each case, we combine five state-of-the-art approaches and the proposed novel approach. The IDF weighting method is better than the state-of-the-art approaches and the embedding-based approaches at all levels. At the word level, the IDF method has an F-score of 88%. However, the best single method achieves an F-score of 86.5%. At the sentence level, the IDF method also obtains a trend of 82.75 against the CL-WES trend (81.5), which was recorded as the best single method. The results obtained in Table IV confirm that the altered approaches proposed experience enhanced performance. Additionally, the obtained results in Table IV indicate that the embeddings are practical for Arabic-English cross-language similarity detection.

Finally, the performances of the methods indicate their capabilities with the dataset. In Fig. 6, we find that the precision improved by 1.54% in the Wikipedia and MultiUN corpora; the recall increased to 1.23%, and the F-score also increased by 2.05 in the Wikipedia and MultiUN corpus. By combining the performances of each method for the dataset, we find that the effect of the IDF method is better than that of the state-of-the-art methods, as discussed previously.

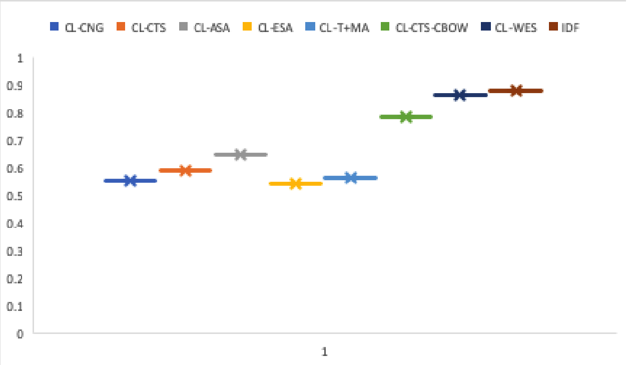


Fig. 4. Comparison of State-of-the-Art Method Performances and the Proposed Method Performance.

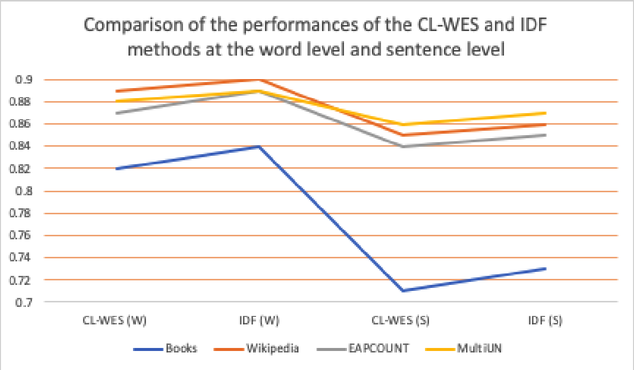


Fig. 5. Comparison of the Performances of the CL-WES and IDF Methods at the Word Level and Sentence Level.

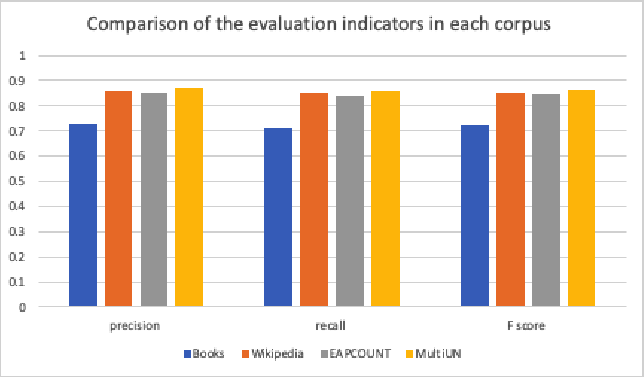


Fig. 6. Comparison of the Evaluation Indicators in each Corpus.

TABLE. IV. THE PERFORMANCES OF CROSS-LANGUAGE SIMILARITY DETECTION METHODS ON ARABIC-ENGLISH CORPORA

Word level					
Methods	***Books (%)***	***Wikipedia (%)***	***EAPCOUNT(%)***	***MultiUN (%)***	***Overall (%)***
CL-CNG	***0.44***	***0.61***	***0.58***	***0.57***	***0.55***
CL-CTS	*0.58*	*0.65*	*0.57*	*0.56*	*0.59*
CL-ASA	***0.56***	***0.74***	***0.66***	***0.63***	***0.6475***
CL-ESA	***0.47***	***0.57***	***0.53***	***0.60***	***0.5425***
CL-T+MA	***0.54***	***0.59***	***0.54***	***0.58***	***0.5625***
CL-CTS-CBOW	***0.75***	***0.80***	***0.79***	***0.80***	***0.785***
CL-WES	***0.82***	***0.89***	***0.87***	***0.88***	***0.865***
IDF	***0.84***	***0.90***	***0.89***	***0.89***	***0.88***
Sentence level					
Methods	***Books (%)***	***Wikipedia (%)***	***EAPCOUNT(%)***	***MultiUN (%)***	***Overall (%)***
CL-CNG	***0.44***	***0.61***	***0.58***	***0.57***	***0.55***
CL-CTS	***0.48***	***0.55***	***0.57***	***0.56***	***0.54***
CL-ASA	***0.54***	***0.67***	***0.64***	***0.65***	***0.625***
CL-ESA	***0.51***	***0.53***	***0.65***	***0.66***	***0.5875***
CL-T+MA	***0.56***	***0.59***	***0.54***	***0.58***	***0.5675***
CL-WES	***0.71***	***0.85***	***0.84***	***0.86***	***0.815***
IDF	***0.73***	***0.86***	***0.85***	***0.87***	***0.8275***

VI. CONCLUSION AND FUTURE WORK

A novel approach for a word embedding-based system is presented in this paper to measure similarities in two cross-linguistic plagiarism. This method could be used for different cross-language similarities and in the training and evaluation phases applied in the Arabic-English corpus as a special case. The proposed methodology improves upon a syntactically weighted distribution representation that operates using the cosine similarity of imbedded vectors (*CL-WES*). The CL-WES model dominates all of the top state-of-the-art methods. Conclusively, the outcomes achieved from the proposed system confirmed that all methods are complementary and that their IDF weights are beneficial to the performance of cross-language textual similarity detection. The IDF method indicates an overall F-score of 88% at the word level; however, the CL-WES method obtains an 86.5% F-score at the word level, whereas the best single method obtains an F-score of only 64.75%. Additionally, at the sentence level, the methods show the same trends.

Our future work will be to improve the *CL-WES* method by exploring the syntactic and semantic weights according to the plagiarist's stylometry. Additionally, a smart hybridization between both IDF weighting and POS tagging procedures will be applied to improve the results.

VII. Funding

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Programme.

References

[1] Al-Suhaiqi M, Hazaa MAS, Albared M (2018) Arabic english cross-lingual plagiarism detection based on keyphrases extraction, monolingual and machine learning approach. Asian J Res Comput Sci 2:1-12. https://doi.org/10.9734/ajrcos/2018/v2i330075.

[2] Aljuaid H. (2020) Arabic-English corpus for cross-language textual similarity detection. In: 10th International Conference on Information Science and Applications, ICISA 2019; Seoul; South Korea; 16 December 2019 through 18 December 2019; Information Science and Applications, Lecture Notes in Electrical Engineering, Springer Nature, Volume 621, 2020, Pages 527-536.

[3] Barron-Cedeno A (2012) On the mono- and cross-language detection of text re-use and plagiarism. PhD thesis, Universitat Politenica de Velenica, Span.

[4] Berard A, Servan C, Pietquin O, and Besacier L. (2016.). MultiVec: a Multilin- gual and Multilevel Representation Learning Toolkit for NLP. . In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portoroz, Slovenia,: European Language Resources Association (ELRA).

[5] Berghen FV, Bersini H (2005) CONDOR, a new parallel, constrained extension of powell's UOBYQA algorithm: experimental results and comparison with the DFO algorithm. J Comput Appl Mathemat 181:157-175. https://doi.org/10.1016/j.cam.2004.11.029.

[6] Ferrero J, Agnès F, Besacier L, Schwab D (2017) CompiLIG at semeval-2017 Task 1: cross-language plagiarism detection methods for semantic textual similarity. arxiv preprint arxiv:1704.01346.

[7] Franco-Salvador M, Rosso P, Montes-Y-Gómez M (2016) A systematic study of knowledge graph analysis for cross-language plagiarism detection. Inf Process Manag 52:550-570. https://doi.org/10.1016 /j.ipm.2015.12.004.

[8] Gabrilovich E, Markovitch S (2007) Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on artifical intelligence (IJCAI'07), Hyderabad, India, pp 1606–1611.

[9] Gerard Salton and Christopher Buckley. 1988. Term- weighting approaches in automatic text retrieval. In- formation processing & management, 24(5):513– 523.

[10] Karani D (2018) Towards data science. https://towardsdatascience. com/introduction-to-word-embedding-and-word2vec-652d0c2060fa.

[11] McNamee P, Mayfield J (2004) Character N-gram tokenization for european language text retrieval. Inf Retri 7:73-97. https://doi.org/ 10.1023/B:INRT.0000009441.78971.be.

[12] Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[13] Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013b) Distributed representations of words and phrases and their compositionality. Adv Neural Inform Process Syst 26:9.

[14] Nagoudi ES (2017) Semantic similarity of arabic sentences with word embeddings. In: Proceedings of the third Arabic natural language processing workshop. Association for Computational Linguistics, Valencia, Spain, pp 18–24.

[15] Naili M, Chaibi AH, Ben Ghezala HH (2017) Comparative study of word embedding methods in topic segmentation. Proced Comput Sci 112:340-349. https://doi.org/10.1016/j.procs.2017.08.009.

[16] Pataki M (2012) New approach for searching translated plagiarism. In: Proceedings of the 5th international plagiarism conference, Newcastle, UK.

[17] Pennington J, Socher R, Manning C (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1532-1543.

[18] Pinto D, Civera J, Barrón-Cedeño A, Juan A, Rosso P (2009) A statistical approach to crosslingual natural language tasks. J Algorithms 64:51-60. https://doi.org/10.1016/j.jalgor.2009.02.005.

[19] Rong X (2016) word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.

[20] Slav P, Dipanjan D, Ryan M (2012) A universal part-of-speech tagset. In: Proceedings of the eight international conference on language resources and evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey, pp 2089-2096.

[21] Suleiman D, Awajan A (2018) Comparative study of word embeddings models and their usage in Arabic language applications. In: The 19th internationnal Arab conference on information technology – ACIT 2018. IEEE, Werdanye, Lebanon, Lebanon, pp. 0857-1812

A Study of LoRa Performance in Monitoring of Patient's SPO_2 and Heart Rate based IoT

Puput Dani Prasetyo Adi[1], Akio Kitagawa[2]
Electrical Engineering Department (UNMER), University of Merdeka Malang-Indonesia, Malang-East Java, Indonesia[1]
Micro Electronics Research Laboratory (MeRL), Kanazawa University, Kanazawa, Ishikawa, Japan[1, 2]

***Abstract*—In this research, a sensor that will be equipped with blood oxygen saturation function (SPO_2) blood and Heart Rate is MH-ET Live max30102 Sensor with Library Max30105. The advantage of this sensor is compatible with ATmega 328P, which is the Arduino board, the first experiment using Arduino Uno. Therefore, MH-ET Sensor data is integrated with Wireless Sensor Network (WSN) devices, e.g, LoRa (Long Range) 915 MHz and calculate WSN path loss when sending sensor data in mountainous areas, the model used to represent signal analysis and measurements in this study is the Ground Reflection (2-ray) model. therefore, the conditions that can be explained are patients who will send their data over hilly areas and hospitals or medical treatments called receiving nodes or coordinator nodes in much lower areas, in the same situation adding routers is expected to be a comparison of whether the data sent faster or even no impact. Furthermore, in this study, it is expected to provide clear results on the function of the router as the sender of pulse sensor data. The point is patients who are in a higher area with the level of impossibility in bringing the patient due to the condition of the patient so that the SPO_2 data transmission and heart rate of the patient are expected to be known quickly by the medical authorities through the sensor node device attached to the patient's body. The use of the Adaptive Data Rate (ADR) Algorithm is used to optimize data rate, time on air (ToA) or airtime and energy consumption in the network. Therefore, the End Device (ED) in the ADR algorithm must be static (non-mobile). In the process of measuring the ADR algorithm in the position of sending data (uplink) n-bits to n-gateway. Next, the application server used is ThingsSpeak or The Things Network (TTN).**

Keywords*—*Pulse; heart rate; adaptive; data rate; long range; bitrate

I. Introduction

The development of the medical world continues to grow rapidly, medical devices are sophisticated and light, flexible increasingly widespread and Speed in getting health data. Today's technology, known as the industrial revolution 4.0, can bring to a fast-paced, fast-paced world, one of which is the existence of the Internet of Things (IoT). IoT has become a mainstay in various fields of life e.g., health, education (Virtual Reality (VR), Automation, Robotic [1],[2]), industry, Search and Resque (SAR) application [3]. Especially for industries, for example, Programmable Logic Controller (PLC) now uses the term PLC-IoT. specifications in the Wireless Sensor Network, in the process of communication between nodes, between TX and RX, involve several components i.e., Tx Power or Power Consumption for transmitters, each wireless technology device is different. To create an efficient sensor node for power consumption, an algorithm, e.g. Adaptive Data Rate (ADR), Automatic Sleep mode, and other algorithms are needed for efficiency [4]. therefore, Cellular devices have the largest Tx Power (mW) which is ~ 500 mW. Then Tx Power WIFI ~ 80 mW, while LoRa ~ 20 mW and Bluetooth ~ 2.5 Mw [5]. Bluetooth has the smallest Tx Power Consumption of all Wireless devices, but the disadvantage is that the short distance is only ~ 10 m. while LoRa reaches 13 km in Free Space Path Loss (FSPL). Therefore, WiFi and Cellular require a large Power Consumption but also limited by the distance that can be up to ~ 5 km on WiFi. The next advantage of LoRa is a small data rate (bps) when sending data.

In this research, the Wireless Sensor Network that is built is based on LoRa (Long Range) Radio Frequency, according to LoRa, has a different type of frequency based on ISM (Industrial, Scientific, and Medical) Band, this frequency distribution is based on the location of the continent or region of each country. e.g., Europe 867-869 MHz, North America 902-928 MHz, China 470-510 MHz, Korea, and Japan 920-928 MHz, and India 865-867 MHz. This is an example of the region's division of the Frequency value in the ISM Band [6], furthermore, the details as in the regional document parameters of the LoRa Alliance. As in research [7], LoRa and LoRaWAN are Wireless devices that have the farthest data transfer capability of ~ 13 km [8].

Therefore, the Low Power Wide Area (LPWA) technology or Low Power Wide Area Network (LPWAN) [9], it has the farthest data sending capability with low power consumption, e.g, in FSPL with the smallest data bit rate and low power consumption. therefore, Three characteristics of Wireless Sensor Network or End-node are Range (m), the speed of data transmission (data rate or bit rate (bps)), and Power Consumption (mW) [10], under conditions of the number of small nodes or nodes in large numbers, e.g, Bluetooth, ZigBee [11][12], WiFi, and LoRa. When compared to the Long Range capability, LoRa is preferable than Bluetooth, ZigBee or WiFi. However, LoRa cannot transmit large data or LoRa bit rates of only ± 250 bps, but the LoRa Power Consumption is low when compared to other radio devices. When compared with ZigBee (250 kbps), Bluetooth (± 3 Mbps) or WiFi (± 11 Mbps), however, Zigbee, Bluetooth, and WiFi are only for close distances and require a large Power Consumption. Therefore, it is difficult to transmit Long Range data at high data rates. It was concluded that the best performance of the characteristics of Wireless Sensor Network devices is seen from 3 sides, i.e, Power Consumption, Range and Speed of

Data Transmission. Fig. 1 shows the WSN best performance triangle, which is the Long Range (km), High-Speed Data Transmission (Mbps) [13].

Table I shows a comparison of i.e radio technology, Bluetooth, WiFi, 3G / 4G and LoRa with reliability ranges. LoRa and LoRaWAN [14] are Wireless Sensor Network technologies that are specifically used for long distances because in addition to being able to send data up to ~ 15 km in FSPL (Free Space Path Loss) or Line of Sight (LOS) conditions and Tx Power reaches low ~ 20 mW [15].

From the data Table I, LoRa is excellent in the data transmission range and is small in Tx Power, but the smallest in terms of transmission speed data (bps). as Fig. 1 almost impossible, which is almost close to LoRa, but LoRa cannot send large data up to Mbps in size. Therefore, this research will use LoRa at Frequency 915 MHz to send SPO_2 and Heart Rate data for monitoring the health of patients in mountainous locations. accordingly, the theory that will be used in this research is the Two Ray Ground theory using Matlab software. it is used mountainous locations as research locations so it uses the transmitter height parameter (*Ht*). therefore, the Radio Signal Path Loss LoRa in mountainous locations. Transmitter height factor (*Ht*), receiver or Base Station (*Hr*). and the distance between the two or turnover distance. When the position of the transmitter (H_t) is above the mountains it means that it is possible for transmitting data to run well or at least reduce the large Path Loss (dB) due to Diffraction, reflection, and Scattering. furthermore, when discussing the diffraction factor, reflection and scattering will go deeper into the type of material, whether buildings or buildings, trees, and material forms that cause the scattering process.

Accordingly, this research will be focused on the SPO_2 and Heart-beat sensor data from RF96 Chip LoRa communication. furthermore, the LoRa Communication can get the result of the Receive Signal Strength Indicator (-dBm) and the Signal Noise Ratio (SNR) (dB). Furthermore, this research will be developed with the Communication system topologies and the Transmission methods of the node sensor and the Gateway with Uplink and Downlink data (bps).

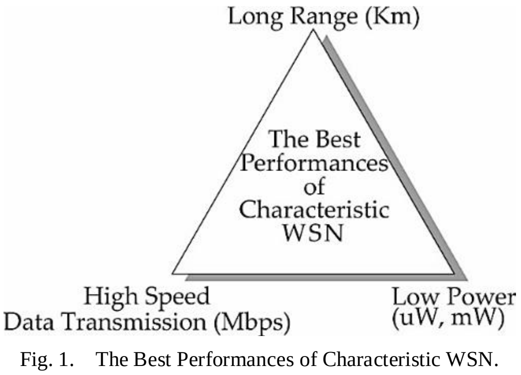

Long Range (Km)
The Best Performances of Characteristic WSN
High Speed Data Transmission (Mbps)
Low Power (uW, mW)

Fig. 1. The Best Performances of Characteristic WSN.

TABLE. I. THE RANGE, POWER AND TRANSMISSION SPEED OF WIRELESS TECHNOLOGY

Technology	**Wireless Communication**	**Range**	**Tx Power**	**The Transmission Speed (bps)**
Bluetooth	Short Range	~10 m	2.5 mW	±3 Mbps
ZigBee / IEEE 802.15.4	Short Range	~ 120 m (LoS)	~2 mW	250 kbps
WiFi	Short Range	~50 m	80 mW	± 11 Mbps
3G/4G	Cellular	~ 5 km	5000 mW	± 12.5 Mbps
LoRa	LPWAN	~ 2-5 km (urban) ~5-15 km (rural) > 15 km (LOS)	20 mW	±250 bps

II. Related Works

Dong-Hoon Kim, Eun-Kyu Lee, and Jibun Kim, in their research [16], conducted a LoRa performance test to obtain the path loss value of LoRa PHY at a distance of 630 - 1344 m with a variety of Spreading Factor values. This research also uses a dynamic back-off Algorithm to improve LoRa MAC performance. And the Multi Gateway approach is also carried out so that redundant communication of data can be studied, this is an interesting research topic about LoRa and LoRaWAN in the future. At another researcher give the conclusion of a LoRa gateway supports up to 6000 nodes with PRR requirement of >70% [17], In other studies sensor nodes were added up to 1000 nodes per gateway and the losess will be up to 32% [18].

Philip A.Catherwood, David Steele, Mike Little, Stephen Mccomb, and James Mclaughlin, with his research entitled "A Community-Based IoT Personalized Wireless Healthcare Solution Trial", have conducted a LoRaWAN Performance test at a distance of 1.1 - 6.0 Km and in this experiment apply for monitoring HealthCare or Medical, furthermore, the result of measurement obtained PathLoss radio value of 119-1141 dB [19].

III. Methodology and Devices used

A. The Sensor used Type

The type sensor used is MH-ET Live MAX30102, this is a sensor used to detect the Pulse Oximetry and Heart rate monitor transmission the module of the sensor. The type of microscope used is the Keyence VHX Digital Microscope. Furthermore, this sensor has dimensions x and y of x = 11,964 mm and y = 10.16 mm. The way this sensor works is to read pulse and Blood Oxigen saturation using different transmittance when the blood vessel beats.

$$SaO_2 = \frac{C_{HbO_2}}{C_{HbO_2}+C_H} \text{ x } 100\ \% \qquad (1)$$

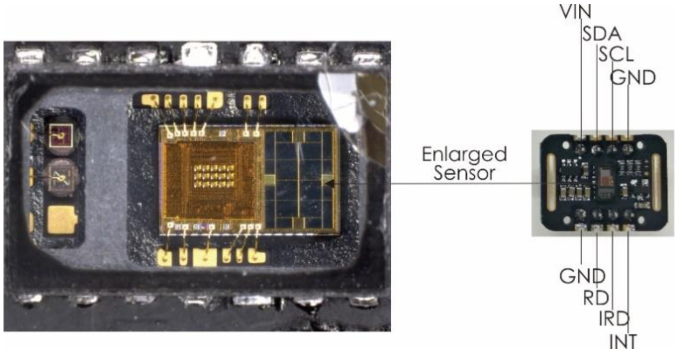


Fig. 2. Pulse Oximetry and Heart Rate Sensor Pins.

The light source, a specific wavelength of light-emitting diode selective for oxyhemoglobin (HbO_2) and hemoglobin (Hb) in arterial blood.

Light transmittance is converted into an electrical signal, the change in the volume of the arterial pulsation causes the light transmittance of the light to change. At this time the light reflected by the human tissue is received by the photoelectric transducer, converted into an electrical signal, and amplified and output. The equation to measurement the SaO_2 is represented at equation 1.

Table II shows the MH-ET Live and Dragino LoRa Pins, there are 4 main pins used by the MH-ET Sensor i.e, Vin, SDA, SCL, and GND. Therefore for Dragino LoRa uses a voltage of 5 Volt DC for MH-ET Sensor, A4 for SDA and A5 for SCL. therefore, GND, RD, IRD and INT pins on one side of the MH-ET Sensor [Fig. 2] are not used. More can be seen in Table II. The plot of the heart rate display is shown in Fig. 21. There are three different conditions shown, Fig. 21 is a condition where the Finger is not placed on the sensor so there is no detection of arterial blood. Fig. 22 is when the Finger is detected by the sensor precisely, in this condition the IR value increases. Although using objects other than fingers, will not be detected normally, only InfraRed detects the movement of objects near the sensor, so the Hearth rate data is not accurate, as shown in Fig. 23.

B. The Shannon-Hartley Theorem

Shannon-Hartley Theorem concluded that the magnitude of Channel Capacity (C) in units of bits per second (bps) is determined by bandwidth (B), signals received (S) at a certain bandwidth and noise (N) or interference over the bandwidth.

$$C = B\ log_2 \left(1 + \frac{S}{N}\right) \tag{2}$$

TABLE. II. MH-ET LIVE SENSOR AND DRAGINO LORA PINS CONNECTIVITY

Pin Number	Pulse Oximetry Sensor, MH-ET pin	Dragino LoRa Pin
1	Vin	5 Volt DC
2	SDA	A4 (Analog 4)
3	SCL	A5 (Analog 5)
4	GND	GND

So Shannon-Hartley Theorem can be written with equation 2 .This means that the signal strength is influenced by the Signal Noise Ratio (SNR). Furthermore, SNR becomes a parameter in determining the radio frequency level of a Long Range (LoRa) radio frequency.

C. The Chirp Spread Spectrum (CSS)

According to the theory, modulation is the process of carrying analog information or digital information through a carrier signal, as the research [7] discussed about the Chirp Spread Spectrum (CSS), it was applied to Radar technology [24], There are 3 types of modulation types in signal modulation in i.e analog information, Amplitude Modulation, Frequency Modulation and Phase modulation and in digital information, i.e., Amplitude Shift Keying (ASK), Frequency Shift Keying (FSK), and Phase Shift Keying (PSK). According to research [7], there are 2 types of chirp, namely up-chirp and down-chirp. It can also be up-chirp and down-chirp referred to as part of the preamble which shows the nature or shape change of the chirp signal in the encoded (data) or up and down condition of the chirp. Furthermore, this CSS will be seen in the different Spreading Factor values. Furthermore, Chirp is stated in Real Spectogram in Fig. 3. To detect LoRa signal or Chirp signal, this research uses Tektronix RSA 34088. Span 5 MHz to produce the appropriate Chirp signal form, in the trials in this research, LoRa signal type up -chirp and down-chirp containing CRC preamble, payload and Payload.

D. Spreading Factor (SF)

Spreading Factor (SF) is a factor that affects the strength of Radio Frequency Long Range Signal Frequency. The value of the Spreading Factor is 7,8,9,10,11 and 12 [7]. Spreading Factor determines the value of the symbol rate and chirp rate. Comparation Spreading Factor with different bandwidth (125 kHz, 250 kHz and 500 kHz) showed at Fig. 4, 5 and 6.

Spectrogram from Spreading Factor (SF) on Fig. 4, 5 and 6, was created using the MatLab Software in Program 1. In Program 1, for example, Bandwidth 500 kHz with Spreading Factor 7, produces a spectrogram different from other bandwidths (125 kHz, 250 kHz). furthermore, the Comparison can be made by comparing the spectrogram with the Spreading Factor 7,8,9,10,11 and 12.

With the combination of Bandwidth (BW). The red lines are called Preambles. The height of the preamble or amplitude of each preamble differs based on the value of the Bandwidth (BW) while the time difference is seen from the Spreading Factor (SF), the greater the Spreading Factor, the longer the time needed for one preamble. From the data taken from the calculation of the results of the comparison of characteristics between Bandwidth, Time on-air, Bitrate and SF obtained 2 graphs in Fig. 11 and Fig. 12.

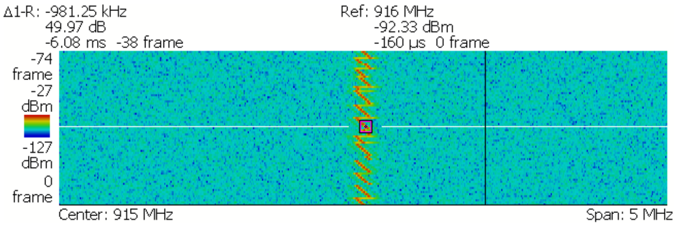


Fig. 3. Chirp Signal from Signal Analyzer.

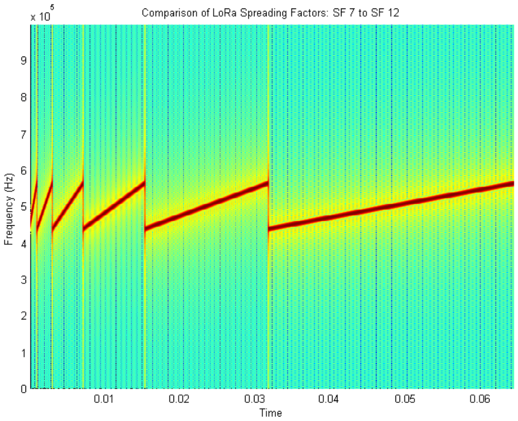


Fig. 4. Comparation Spreading Factor (SF) on 125 kHz Frequency.

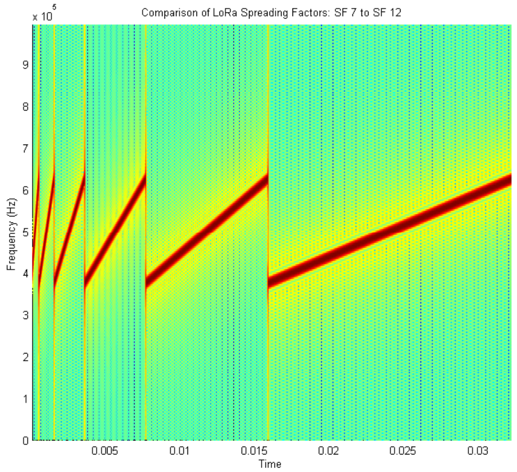


Fig. 5. Comparation Spreading Factor (SF) on 250 kHz Frequency.

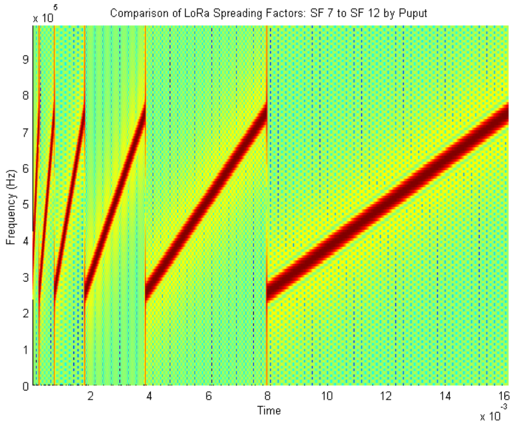


Fig. 6. Comparation Spreading Factor (SF) pada 500 kHz Frequency.

```
BW = 500000; % 500 kHz Bandwidth
Fs = 10^6;    % Sampling Frequency
inverse = 0;     % inverse = 1 for inverse chirps,
inverse = 0       % for normal chirps
% Case 1
SF = 7;
num_samples = Fs*(2^SF)/BW;  % Number of samples
[out_preamble1] =
LoRa_Modulation(SF,BW,Fs,num_samples,0,inverse);
outp = [out_preamble1 out_preamble2 out_preamble3
out_preamble4 out_preamble5 out_preamble6];
samples = length(out_preamble1)/4;
spectrogram(outp,samples,samples-
1,samples*2,Fs,'yaxis');
title('Comparison of LoRa Spreading Factors: SF 7
to SF 12');
grid on;
axis tight;
------------------ Program 1 ----------------
```

E. Sensitivity of LoRa

To calculate the power level or sensitivity of the Receiver, 3 parameters need to be known, i.e., Bandwidth (BW), Noise Figure (NF) and SNRlimit, Sensitivity (dBm), [Fig. 9] is abbreviated with S as equation 3. with a formula like in equation 1. Accordingly the theory, S is Sensitivity (-dBm), BW is Bandwidth (Hz), Noise Figure (NF) are measures of degradation Signal to Noise Ratio (SNR), Noise Figure (NF) of LoRa differ for each device, in general, the value of the LoRa NF is 6. While the SNRlimit, refer to Table III.

$$S = -174 + 10\, Log_{10}(BW) + NF + SNR_{limit} \qquad (3)$$

The greater the Spreading Factor (SF) causes the data speed to be smaller, the greater the range of distance (indicated by ToA), and the greater the sensitivity (*S*), complete information can be seen in Table III.

Furthermore, the Sensitivity of LoRa is also influenced by SNR_{limit}, the greater the Spreading Factor causes the value of SNR_{Limit} is also greater, in SF 7, the SNR_{Limit} value is -7.5 dB, and in SF 12, the SNR_{Limit} value reaches -20 dB.

F. Link Budget of LoRa

Link Budget shows the ability of LoRa in signal Propagation at a certain distance. Many factors that affect Link Budget include Power Transmitter, Transmitter and Receiver gain, Obstacle on Signal Propagation and Sensitivity factor. The relationship between sensitivity and Link Budget is shown in equation 4 and Fig. 10.

$$Link\ Budget\ (dB) = TX\ Power - S\ (Sensitivity) \qquad (4)$$

TABLE. III. LoRa Spreading Factor Comparation with SNR Limit

Spreading Factor (SF)	Chips /Symbol	SNR_{Limit}	Time on Air (ToA) (10 byte Packet)	Bitrate
7	128	-7.5	56 ms	5470 bps
8	256	-10	103 ms	3125 bps
9	512	-12.5	205 ms	1758 bps
10	1024	-15	371 ms	977 bps
11	2048	-17.5	741 ms	537 bps
12	4096	-20	1483 ms	293 bps

G. Bit Rate or Data Rate (R_b)

Bit Rate or Data Rate (Rb) is several bits sent in units of time (s) or expressed in bits per second (bps). There is an important link between Spreading Factor and Bandwidth to determine how much Bit rate (Rb) or data rate (bps) is generated. furthermore, the LoRa bit rate is expressed as in equation 5. Furthermore, the ratio of bandwidth (BW) and Spreading Factor (SF) to bit rate is shown in graph 1 (b) bits rate (bps) with SF.

$$R_b = \mathrm{SF}.\frac{\left[\frac{4}{4+CR}\right]}{\left[\frac{2^{SF}}{BW}\right]}.1000 \tag{5}$$

Where CR is a code rate of 1 to 4, and LoRa has bandwidth specifications ranging from 125 kHz, 250 kHz and 500 kHz. Based on from Libelium, LoRa or LoRaWAN has a configuration on Spreading Factor, Bandwidth and BitRate which can be seen in Table IV and Fig. 11.

H. Time on Air (ToA)

Time on Air (ToA) is the time used by Radio Transmission LoRa in one time sending data from the Transmitter (T_x) to Receiver (R_x) consisting of $T_{Preamble}$ and $T_{Payload}$. Therefore, Time on Air is shown in Equation 6, Equation 7 and Fig. 12.

$$ToA = T\ Preamble + T\ Payload \tag{6}$$

With $T_{preamble}$ = Nb Preamble (8)+symbols added by radio (4.25) x T_{symbol} , $T_{payload}$ = NbPayloadSymbol x T_{symbol}.

$$n_{payload} = 8 + \max\left(ceil\ \left[\frac{(8PL-4SF+28+16CRC-20IH)}{4(SF-2DE)}\right]\ (CR+4), 0\right) \tag{7}$$

I. LoRa Symbol Symbol Duration (Ts) or T_{sym}

LoRa Symbol is the time used by LoRa within 1 second to transmit data or signals, this signal is a Chirp signal consisting of Preamble, Payload and Payload CRC. Furthermore, LoRa Symbol can be avowed as in equation 8 [25].

$$T_{sym}\ or\ T_S = \frac{2^{SF}}{BW} \tag{8}$$

J. Signal to Interference Ratio (SIR)

Signal Noise Ratio or Signal to interference ratio (SIR) is transmit Power (P_i) or (P_t) multiply with Direct Channel or Gain on the transmitter (G_{ii}) divide by Power receiver (P_j) or (P_r) multiply with Direct Channel or Gain on receiver (Gij) plus noise on the transmitter (ni), which is affected by interference or noise.accordingly, equation 9 shows the value of Signal to Interference Ration (SIR) when transferring LoRa 915 MHz data. Furthermore, the percentage of SIR1 and SIR2 can be showed by equation 10.

$$SIR_i = \frac{P_{i.}G_{ii.}}{\sum_{j\neq i}\ P_{j.}G_{ij} + n_i} \tag{9}$$

$$SIR_1 = \frac{P_{1.}G_{11}}{P_{2.}G_{12} + n_1}\ \text{dan}\ SIR_2 = \frac{P_{2.}G_{22}}{P_{1.}G_{21} + n_2},\ \text{etc.} \tag{10}$$

K. Bit ErrorRate (BER)

Bit error rate or bit error ratio is the number of digital bits in the transmission network, in this case, LoRa 915 MHz where the total number of error bits is divided by the number of bits sent in a certain time (t), e.g, Bits sent 0110001011, while those received are 0010101001, from 10 data bits sent, there are 3 error bits, so the percentage is 3/10 or 0.3 or 30% BER, BER can be showed by equation 11[23].

$$BER = \frac{N\ error}{N\ bits} \tag{11}$$

L. Packet ErrorRate (PER)

Packet Error Rate (PER)% is the total number of Packets Error divided by the total packet received at a certain time during the Uplink process, i.e, the transmission sensor data from End Devices to the Gateway (GWs). Packet Error Rate (PER)% can be stated in equation 12. furthermore, Packets sent from EDs to GWs will be recorded by GWs and will be compared between Packets sent and Packets that Error, so that the percentage or amount of error arises from the transmitting data process [20].

$$P_p = 1 - (1 - P_e)^N = 1 - e^{N\log(1-P_e)} \tag{12}$$

E_b/N_o and C/N

in data transmission, Eb / No is energy per bit to noise power spectral density ratio, is the same as the normalized measurement of Signal Noise Ratio (SNR), also known as SNR per bit, Eb / No can be showed at Equation 13, 14 and 15.

$$\frac{C}{N} = \frac{Eb}{N_O}.\frac{fb}{B_W} \tag{13}$$

Where, C / N is Carrier to noise ratio (dB), Eb / N_O (dB) is the ratio of energy per bit (Eb) to the spectral noise density (N_O), Fb is bit rate (bps) and B_W is the receiver noise bandwidth (bps).

Noise Power is computed using Boltzmann's equation: $\mathrm{N = kTB}$, where, $\mathrm{k = Boltzmann's\ constant} = 1.380650\mathrm{x}\ 10^{23}\ J/K$; T = is effective temperature in Kelvin, and B is the receiver bandwidth. Furthermore, the equation of $\frac{C}{N}\ (dB)\ and\ \frac{E_b}{N_o}\ (dB)$.

$$\frac{C}{N}\ ratio\ (dB) = \frac{E_b}{N_o}\ (dB) + 10\ Log\ (\frac{fb}{Bw}) \tag{14}$$

and,

$$\frac{E_b}{N_o}\ (dB) = \frac{C}{N}(dB) + 10\ Log\ (\frac{Bw}{fb}) \tag{15}$$

TABLE. IV. CONFIGURATION OF SPREADING FACTOR, BANDWIDTH AND BITRATE IN JAPAN

Index Number	Spreading Factor (SF)	Bandwidth (BW)	Bit Rate (bps)
0	12	125 kHz	250 bps
1	11	125 kHz	440 bps
2	10	125 kHz	980 bps
3	9	125 kHz	1760 bps
4	8	125 kHz	3125 bps
5	7	125 kHz	5470 bps
6	7	250 kHz	11000 bps

Furthermore, to calculate BER, with an approach to the energy factor per bit to noise power spectral density ratio or Eb / No, then, according to equation 16.

$$BER = Q\ (\frac{log_{12}(SF)}{\sqrt{2}}\ \frac{E_b}{N_o}) \tag{16}$$

M. SNR (-dB)

The relationship between SNR (dB) and $\frac{E_b}{N_o}$ (dB), can be showed at equation 17.

$$SNR\ (dB) = \frac{E_b}{N_o} + 10.\ log_{10}(R_S) + 10.\ log_{10}(k) + 10.\ log_{10}(R) - 10.\ log_{10}(BW_n) \qquad (17)$$

Where R_S is the symbol rate; k is the number of information bits per symbol; R is code rate; and BW_n is the noise Bandwidth.

N. *Coding Rate (CR)*

Code Rate or Coding Rate (CR) is used to handle Packet Error Rate (PER) due to interference, with a formula as shown in equation 18. where n is {1, 2, 3, 4}.

$$CR = 4/(4+n) \qquad (18)$$

O. *Symbol Rate (R_s)*

In digital communication Symbol Rate (R_s) also called Boudrate, the value of R_s is shown in equation (*r*). the relationship between R_s, SF, R_c is shown in equation 19.

$$SF = \ Log_2\left(\frac{Rc}{Rs}\right) \qquad (19)$$

Where Rs is symbol rate, symbol rate (Rs) equal $\frac{1}{Ts}$ which showed by equation 20.

$$Rs = \frac{1}{Ts} \qquad (20)$$

Rs juga dapat ditentukan dengan perbandingan Bandwidth dengan 2^{SF} seperti ditunjukkan pada equation 21.

R_s can also be determined by comparing Bandwidth with 2^{SF} as shown in equation 21.

$$Rs = \frac{BW}{2^{SF}} = \frac{Rc}{2^{SF}} \ \text{symbols/s} \qquad (21)$$

Fig. 13 shows the characteristics of the Symbol Rate (Rs) in different *BW* and Spreading Factor (*SF*).

P. *Bandwidth or Chip Rate (Rc)*

Chip Rate (*Rc*) is the number of chips / second. And the value of Chip rate (*Rc*) equal by Bandwidth (*BW*) value. e.g, Bandwidth is 125 kHz, then Rc is 125000 chips / second, or in other words, the same as the Chirps Spread Spectrum (CSS) reaches 125000 chips / second. The relationship between Chip rate (R_c) and Bit rate (R_b) showed on equation 22.

$$Rc = 2^{SF}.Rb \qquad (22)$$

Then, The relationship between Chip rate (R_c) with *Symbol Rate (R_s)* by unit Chip/sec, showed on equation 23 and 24.

$$Rc = 2^{SF}.Rs \qquad (23)$$

Where, R_c equal *BW.*

$$R_c = \text{BW chips/s} \qquad (24)$$

Where *Rb* is Data rate or Bit Rate in bits per second (*bps*), *BW* is Bandwidth in KHz (10.4, 15.6, 20.8, 31.25, 41.7, 62.5, 125.250, 500) used by LoRa is Bandwidth 125, 250, 500 kHz, *CR* is Code Rate (1,2,3,4) and *SF* is Spreading Factor (6,7,8,9,10,11,12). Code or Coding rate (*CR*) equal to 4 / (4 + n), with n ∈ {1,2,3,4}. Furthermore, if the bandwidth is 125 kHz, and SF 7, the bit rate result is 5.46 *kbps* or 5460 *bps.*

Q. *RSSI (dBm)*

The *RSSI* (Receive Signal Strength Indicator) is the amount of Power Signal in units (*dBm*), accordingly the theory, *RSSI* can be generated from equation 25. Complete *RSSI* is classified in Table V which shows Signal Level Range (*dBm*).

$$\text{RSSI (dBm)} = 10 \log (\text{Pr}) \qquad (25)$$

R. *Long Range Radio Propagation for LoRa FSPL*

LoRa uses an RP-SMA Male type antenna which has a gain of 2.dBi with 50 Ω impedance. Table VI shows the LoRa Antenna specifications used in this research. The FSPL equation for LoRa propagation can be showed on Equation 26 and 27. Furthermore, RP-SMA Male type antenna can be showed at Fig. 7.

$$FSPL = \left(\frac{4\pi d}{\lambda}\right)^2 \text{ equal } \left(\frac{4\pi df}{c}\right)^2\text{, because } \lambda = \frac{c}{f} \qquad (26)$$

$$FSPL\ (dB) = 20\ log_{10}\ (d) + 20\ log_{10}\ (f) - 147.55 \qquad (27)$$

Where, LoRa 915 MHz Wavelength (λ) = 0.30327642030 m, *c* = Speed of light (299,792.458 m/s), *d* is distance (*m*), frequency is Hertz (*Hz*) and π = 3.14159265358979. In several studies, the calculation and design of circuit and Power efficiency using LTSpice Software [21]. From the comparison of the specifications of the LoRa module and the distances, the *FSPL* of LoRa 915 Module data is obtained as in Fig. 20. LoRa has a different Frequency for every Continent as shown in the table about United Stated frequency allocations in the Radio Spectrum. As 433 MHz and 868 MHz (Europe), 915 MHz (Australia and North America), and 923 MHz (Asia), for Japan 920-923 MHz.

S. *A Type of Obstacle Materials During Radio Propagation*

At the time of propagation of the signal through the obstacle, the signal attains attenuation due to the material in its path [22]. Table VII shows the material type and thickness and PathLoss value. Furthermore, the equation used is the n PathLoss Exponent which shows the value of n based on the material conditions that are passed by Tx and Rx.

TABLE. V. SIGNAL STRENGTH CLASSIFICATION

Signal Level Range (dBm)	**Classification**	**Score**
-120 to -95	Extremely Bad	1
-95 to -85	Bad	2
-85 to -75	Average	3
-75 to -65	Good	4
-65 to -55	Very Good	5
-54 to -30	excellent	6

Fig. 7. RP-SMA Male Type Antenna.

TABLE. VI. LoRa Antenna Spesification

Connector Type	Gain	Polarization	Impedance
RP-SMA Male	2.0 dBi	Linear	50 Ω

TABLE. VII. Material and Thickness and Nilai PathLoss (dB)

No	Material and thickness	Path Loss (dB)
1	Glass (6 mm)	0.8
2	Glass (13 mm)	2
3	Wood (76 mm)	2.8
4	Brick (89 mm)	3.5
5	Brick (178 mm)	5
6	Brick (267 mm)	7
7	Concrete (102 mm)	12
8	Stone wall (203 mm)	12
9	Brick Concrete (192 mm)	14
10	Stone wall (406 mm)	17
11	Concrete (203 mm)	23
12	Reinforced Concrete (89 mm)	27
13	Stone wall (610 mm)	28
14	Concrete (305 mm)	35

T. Two-Ray Ground Reflection (2-ray) Model

The Ground Reflection *(2-ray*) model is a model that predicts Path Loss when sending data from the Transmitter Antenna (*Tx*) to the Receiver (*Rx*) Antenna with Line of Sight (LOS) or facing each other. in general, both antennas have different heights (*ht* and *hr*). ht is the height of the transmitter antenna in meters (*m*) and Hr is the height of the receiving antenna in meters (*m*). consequently, a signal has reflected the ground before the signal is received by the receiving antenna (*Rx*), while the d (distance) is the distance between the sending and receiving antennas in meters (*m*). at the mountains area, the Signal transmitting from *Tx* antenna is far above the hill, therefore, the theory of ground signal reflection can occur so that the Ground Reflection (*2-ray*) of this model is used. on the 2-Ray ground reflection propagation have two wave components that arrive at the Receiver (*Rx*), i.e. Line of Sight (LOS) and reflected from the ground and the reflection coefficient or Fresnel Coefficient ($\Gamma = -1$). The Reflection Coefficient for i-wave can be stated in the equation 28.

$$\Gamma_i = \frac{cos\,\theta_i - q\sqrt{\varepsilon_c - sin^2\,\theta_i}}{cos\,\theta_i + q\sqrt{\varepsilon_c - sin^2\,\theta_i}} \quad (28)$$

ε_c is the complex Permittivity of the ground, θ_i is the incident angle with the normal to the ground, q is a polarization-dependent factor, which is q = 1 for horizontal polarization and q = 1 /ε_c for vertical polarization. a ZigBee has a frequency of 2.4 GHz so that the ZigBee wavelength (λ) is obtained is 0.124913 meters, and LoRa 915 MHz Wavelength (λ) is 0.30.327642030 meters, this value from $\lambda = c\ /\ f$. therefore, to get the Path loss (*PL* (dBm) value), it is necessary to find the strength of the transmitter (*Pt*) and receiver (*Pr*), in general, the calculation formula for the Power Receiver (*Pr*) is as the equation 29.

$$Pr = Pt + Gt + Gr - PL \quad (29)$$

While the model of 2-ray ground reflection propagation formula from *Pr* added the ht and hr parameters due to the relationship with the height of the ht transmitter antenna and different hr receiving antenna which Affects the signal strength level the added angle formed from the reflected process the distance from x to x' therefore, the *Pr* formula becomes as the equation 30.

$$Pr\ (dBm)\ = \frac{\lambda^2}{(4\,\pi\,d)^2}\left[2\,sin\,\frac{2\,\pi}{\lambda}\,\frac{ht.hr}{d}\right]\mathrm{G_t\,.G_r\,.P_t} \quad (30)$$

Overall the variables used in the calculation are described as below:

Pt = Power Transmitter (*dBm)*

Pr = Power Receiver (*dBm*)

λ = Wavelength (*m*)

c = Speed of light (299,792,458 *m/s*)

f = Radio Wave Frequency (*Hz*)

d = Distance (*m*)

ht = Height of transmitter antenna (*meters*)

hr = Height of receiver antenna (*meters*)

Gt = Transmitter antenna gain (*dBi*)

Gr = Receiver antenna gain (*dBi*)

PL = Path Loss (*dBm*)

FSPL = Free Space Path Loss (*dBm*)

RSSI = Receive Signal Strength Indicator (*dBm*)

π = 3.14159265358979

On the Ground Reflection Model, Path Difference (Capital Delta) Δ is the result of a reduction of in reflection or (d'') and in Line of Sight (LOS) or (d'), therefore, the formula from the The Difference Path is shown in the equation 31.

$$\Delta = d'' - d' = \sqrt{(h_t + h_r)^2 + d^2} - \sqrt{(h_t - h_r)^2 + d^2} = \frac{2h_t h_r}{d} \quad (31)$$

Furthermore, Phase Difference $\theta_\Delta$is shown in equation 32, according to the equation, the difference in Phase results can be shown from the Difference Path value.

$$\theta_{\Delta} = \frac{2\,\pi\,\Delta}{\lambda} \qquad (32)$$

as for the correlation with Time delay shown in equation 33. According to the equation 33, Time delay is generated from the division of the Difference Path and the speed of light (c), and equal to Phase Difference divided by $2\pi f_c$, where f_c is the carrier frequency.

$$t_d = \frac{\Delta}{c} = \frac{\theta_{\Delta}}{2\pi f_c} \qquad (33)$$

furthermore, the Power Receiver *(Pr)* can be formulated as in equation 34.

$$P_r = P_t G_t G_r \left(\frac{h_t h_r}{d^2}\right)^2 \qquad (34)$$

Referring to equation 26 and 27, the comparison of the PathLoss (-dBm) results can be seen in the graph Fig. 20, the path loss value is influenced by the transmitter location, in the experiment on the hill, the transmitter is placed in different positions, which affects the signal reception power from the Fig. 14 it can be seen that the higher the transmitter is the greater the strength of the P_{rx} Signal. The first analysis is on 2-ray ground propagation models using the Matlab software, by looking at equation 30, wavelength (λ) value is 0.125 m, with 3 sender height (*Ht*) different that is H_{t1} is 5m, H_{t2} is 20 m and Ht3 is 40 m with Receiver height (*Hr*) is 0.5 m and 50 m. This analysis functions to find out PathLoss if it is based on the sender height (*Ht*) and Receiver Height (*Hr*). from the analysis it was found that if the height of *Ht* and *Hr* is almost comparable, the sinusoidal wave will dock, meaning the signal can be received properly or there is still a response from the receiver even though the power or consequently, *Pr* (Power Receiver) has been decreasing, from the *Hr* = 0.5 m, the simulation it appears that the biggest Power Receiver is -54.3 dBm on *Ht* 5m.

U. Block Diagram

The method used in this research is the Adaptive Data Rate (ADR) Algorithm to Sensor Node 1 (ED_1) to Sensor Node End Device (ED_n). Adaptive Data Rate (ADR) is proven by looking at the indicators on the ToA (Time on Air), the effectiveness of the Bit Rate (bps) and the remaining energy (mW) of the Battery at the sensor node or EDs. Adaptive Data Rate (ADR) flowchart can be seen in the Research reference [7],[26].

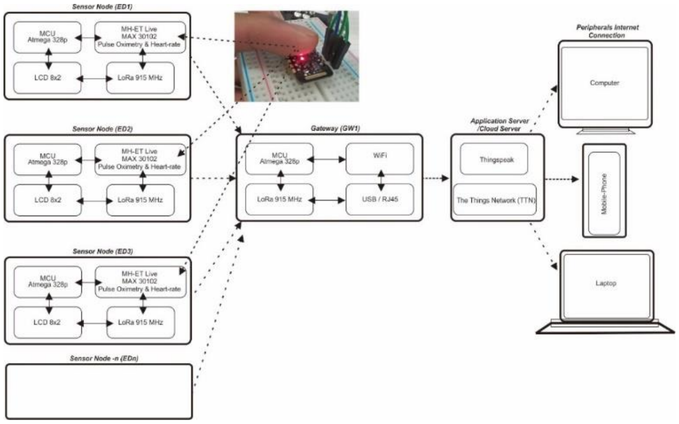


Fig. 8. Blog Diagram on this Research.

Fig. 8 is a blog diagram showing this research, this blog diagram explains how the research works from start to finish and how sensors work on EDs, furthermore, EDs transmit sensor data (transmitting) MH-ET Live Max30102 data to LoRa Gateway (GW). Furthermore, Gateway will store sensor data bits from n EDs and calculate Uplink and downlink values on the Application Server (TTN or Thingspeak) or Cloud Server LoRa. Furthermore, the MH-ET Live Max30102 sensor data provides an IP Address that can be accessed by internet-connected devices.

IV. Result and Discussion

A. Comparison of LoRa Parameters

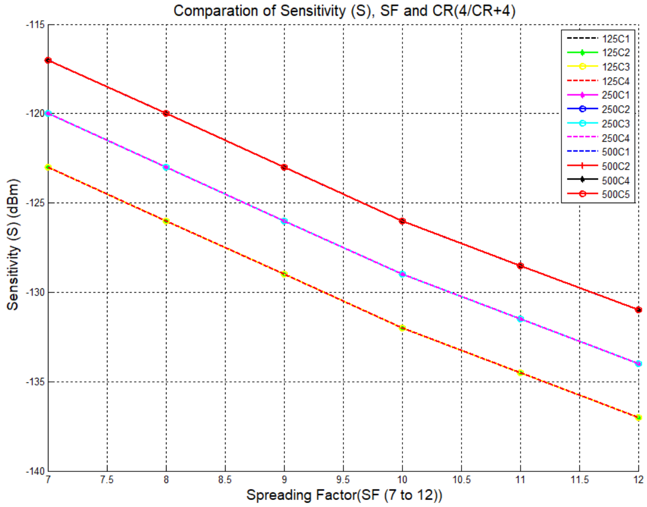


Fig. 9. Sensitivity of LoRa.

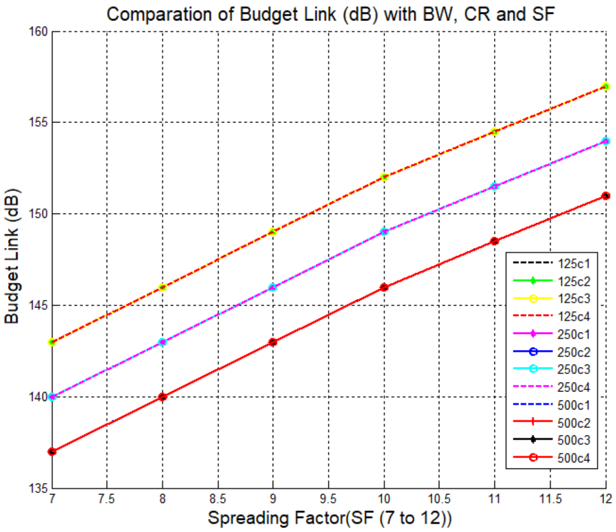


Fig. 10. Comparation of Budget Link (dB), SF and CR (4/CR+4).

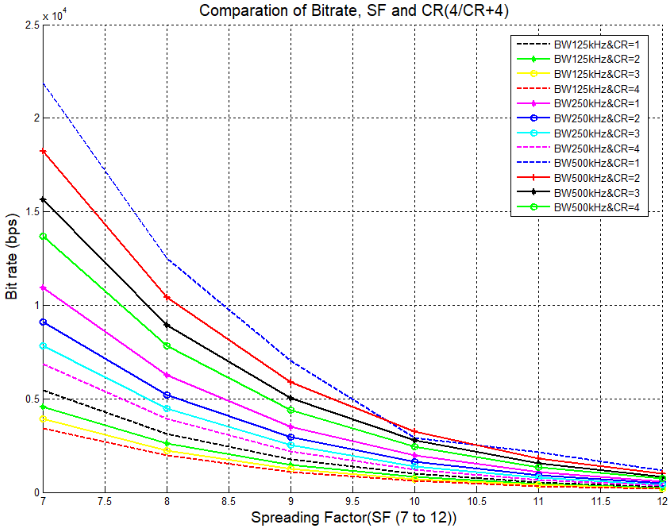


Fig. 11. Comparation of Bitrate (BW), SF and CR (4/CR+4).

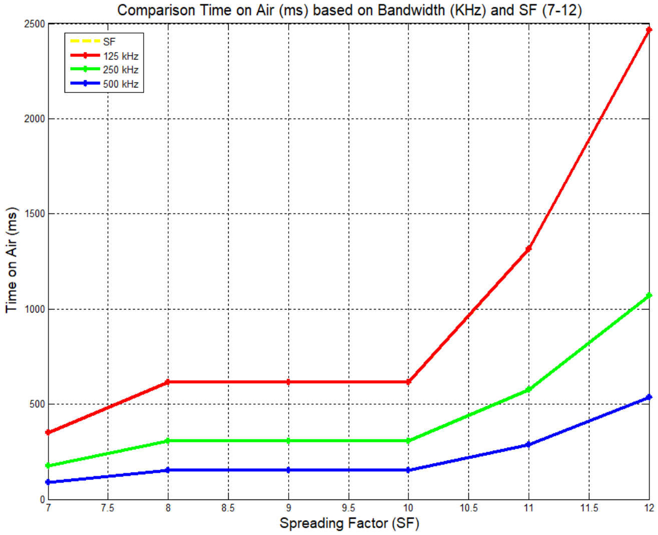


Fig. 12. ToA (ms) with SF.

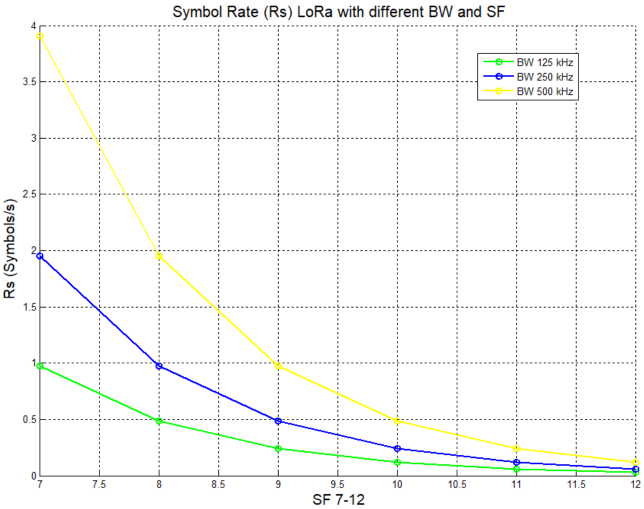


Fig. 13. Comparison of Symbol rate (Rs), BW and SF.

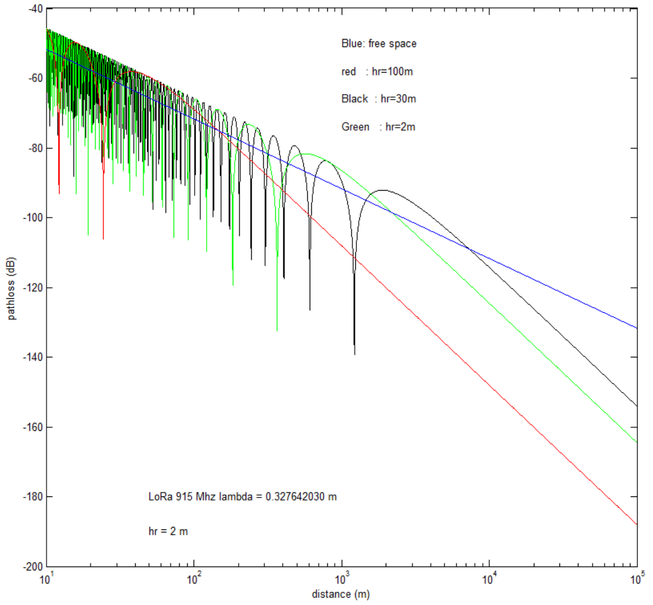


Fig. 14. LoRa 915 MHz Pathloss [dBm] of 2-Ray Ground Reflection Model.

B. Testing uses a Serial Monitor and LoRa Library

The first trial is sending the MH-ET Live sensor data max30102 sensor, point to point from the LoRa transmitter (Tx) to the LoRa receiver (Rx), from this step, it will be known that LoRa can communicate well, furthermore, the GW, Lora Tx and LoRa Rx can be showed on Fig. 15. Furthermore, Fig. 16, 17, 18, 19 and 20 are the output produced from LoRa Tx and LoRa Rx, at a certain distance on Arduino Serial Monitor, in LoRa Rx the RSSI, SNR and Packet Frequency Error values are indicated. According to him, in previous experiments the value of RSSI and SNR would experience attenuation signal based on the increasingly longer distance between Tx and Rx.

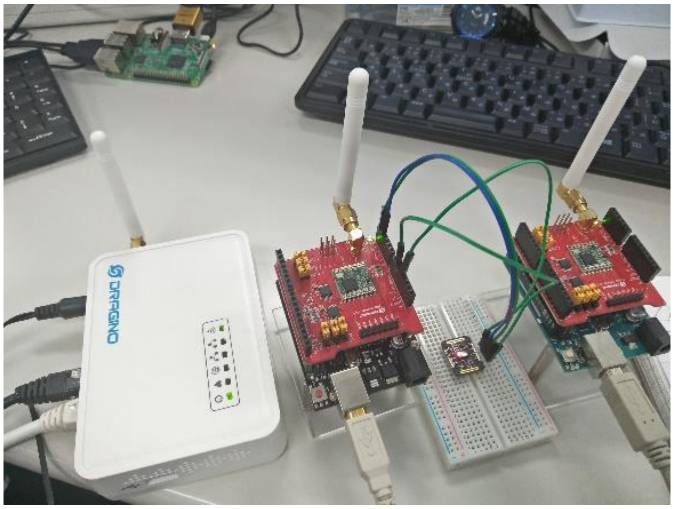
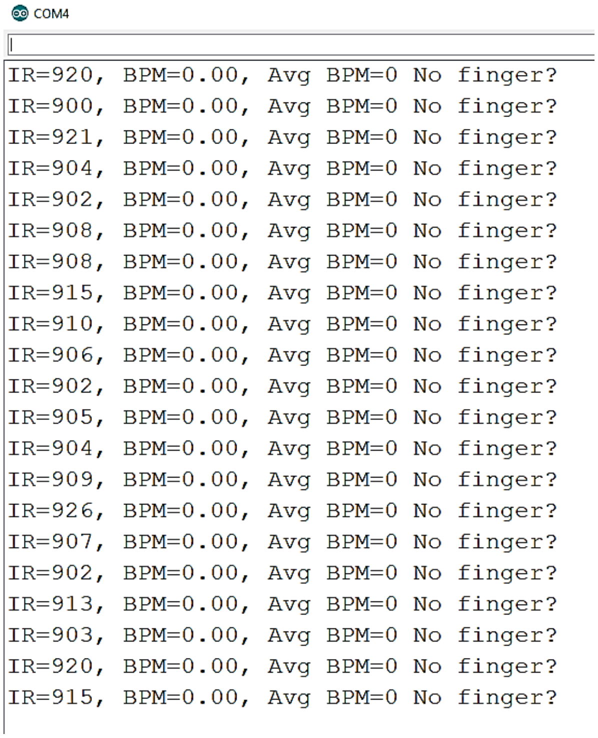
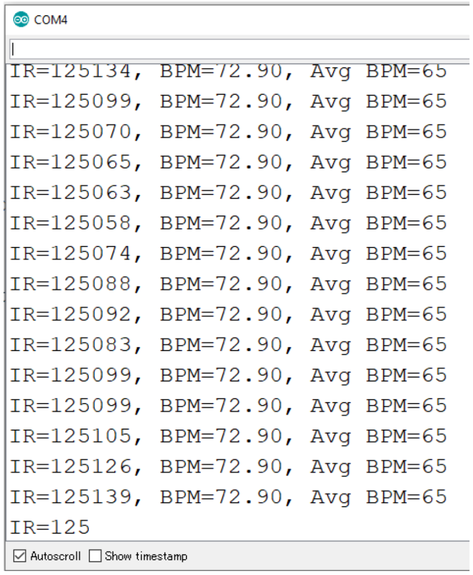


Fig. 15. LoRa Gateway, LoRa Tx and LoRa Rx.



```
IR=920, BPM=0.00, Avg BPM=0 No finger?
IR=900, BPM=0.00, Avg BPM=0 No finger?
IR=921, BPM=0.00, Avg BPM=0 No finger?
IR=904, BPM=0.00, Avg BPM=0 No finger?
IR=902, BPM=0.00, Avg BPM=0 No finger?
IR=908, BPM=0.00, Avg BPM=0 No finger?
IR=908, BPM=0.00, Avg BPM=0 No finger?
IR=915, BPM=0.00, Avg BPM=0 No finger?
IR=910, BPM=0.00, Avg BPM=0 No finger?
IR=906, BPM=0.00, Avg BPM=0 No finger?
IR=902, BPM=0.00, Avg BPM=0 No finger?
IR=905, BPM=0.00, Avg BPM=0 No finger?
IR=904, BPM=0.00, Avg BPM=0 No finger?
IR=909, BPM=0.00, Avg BPM=0 No finger?
IR=926, BPM=0.00, Avg BPM=0 No finger?
IR=907, BPM=0.00, Avg BPM=0 No finger?
IR=902, BPM=0.00, Avg BPM=0 No finger?
IR=913, BPM=0.00, Avg BPM=0 No finger?
IR=903, BPM=0.00, Avg BPM=0 No finger?
IR=920, BPM=0.00, Avg BPM=0 No finger?
IR=915, BPM=0.00, Avg BPM=0 No finger?
```

Fig. 16. Output Data from LoRa Tx No Finger.

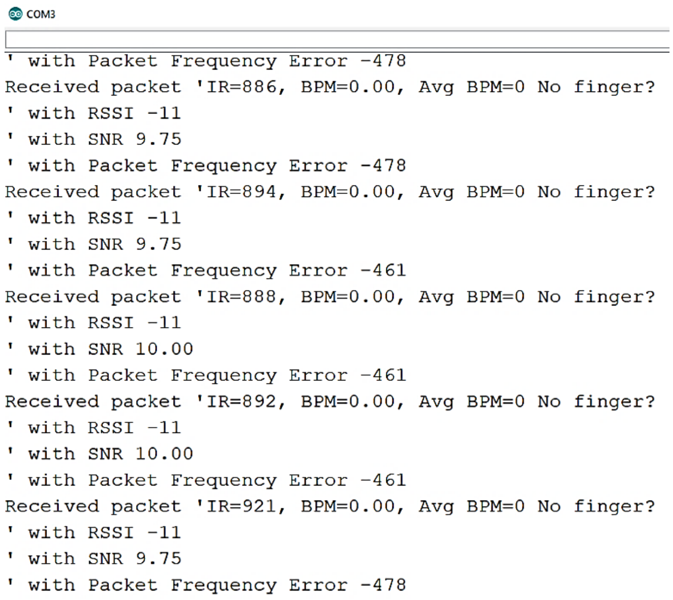


```
IR=125134, BPM=72.90, Avg BPM=65
IR=125099, BPM=72.90, Avg BPM=65
IR=125070, BPM=72.90, Avg BPM=65
IR=125065, BPM=72.90, Avg BPM=65
IR=125063, BPM=72.90, Avg BPM=65
IR=125058, BPM=72.90, Avg BPM=65
IR=125074, BPM=72.90, Avg BPM=65
IR=125088, BPM=72.90, Avg BPM=65
IR=125092, BPM=72.90, Avg BPM=65
IR=125083, BPM=72.90, Avg BPM=65
IR=125099, BPM=72.90, Avg BPM=65
IR=125099, BPM=72.90, Avg BPM=65
IR=125105, BPM=72.90, Avg BPM=65
IR=125126, BPM=72.90, Avg BPM=65
IR=125139, BPM=72.90, Avg BPM=65
IR=125
```

Fig. 17. Output Data from LoRa Tx with Finger Detected.

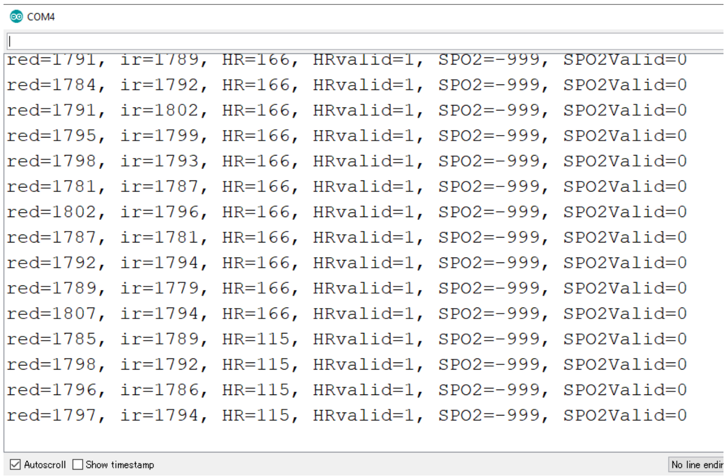
C. Sensor Output

There are three examples of sensor output, on Fig. 21 is a graph from Arduino Serial Plotter, when Finger is not in the position of the sensor properly, so that the signal is irregular and does not show a precise value. Furthermore, Fig. 22 is when the finger is in the right position on the sensor, thus producing a precise HeartBeat value and the sensor detects Arterial Blood, while Fig. 23, the signal shows an inaccurate and changing value because the object used is not Finger, but the object, in other words, cannot detect Arterial Blood.



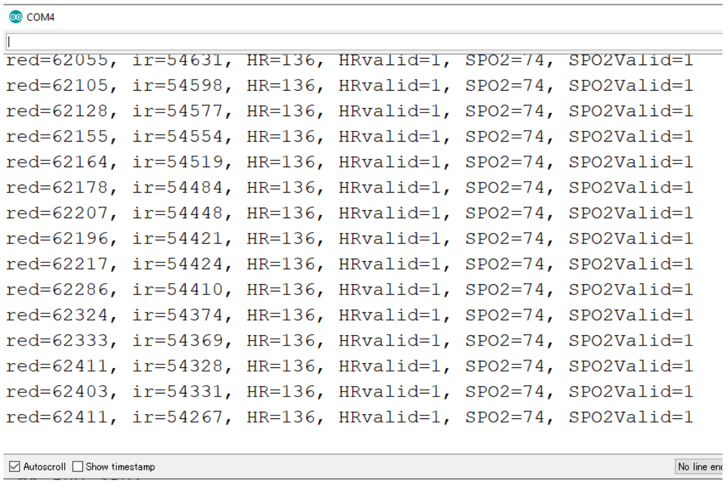
```
' with Packet Frequency Error -478
Received packet 'IR=886, BPM=0.00, Avg BPM=0 No finger?
' with RSSI -11
' with SNR 9.75
' with Packet Frequency Error -478
Received packet 'IR=894, BPM=0.00, Avg BPM=0 No finger?
' with RSSI -11
' with SNR 9.75
' with Packet Frequency Error -461
Received packet 'IR=888, BPM=0.00, Avg BPM=0 No finger?
' with RSSI -11
' with SNR 10.00
' with Packet Frequency Error -461
Received packet 'IR=892, BPM=0.00, Avg BPM=0 No finger?
' with RSSI -11
' with SNR 10.00
' with Packet Frequency Error -461
Received packet 'IR=921, BPM=0.00, Avg BPM=0 No finger?
' with RSSI -11
' with SNR 9.75
' with Packet Frequency Error -478
```

Fig. 18. Output RSSI and SNR Data from LoRa Rx.



```
red=1791, ir=1789, HR=166, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1784, ir=1792, HR=166, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1791, ir=1802, HR=166, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1795, ir=1799, HR=166, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1798, ir=1793, HR=166, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1781, ir=1787, HR=166, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1802, ir=1796, HR=166, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1787, ir=1781, HR=166, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1792, ir=1794, HR=166, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1789, ir=1779, HR=166, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1807, ir=1794, HR=166, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1785, ir=1789, HR=115, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1798, ir=1792, HR=115, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1796, ir=1786, HR=115, HRvalid=1, SPO2=-999, SPO2Valid=0
red=1797, ir=1794, HR=115, HRvalid=1, SPO2=-999, SPO2Valid=0
```

Fig. 19. Output SPO_2 Sensor Data (Not Valid).



```
red=62055, ir=54631, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62105, ir=54598, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62128, ir=54577, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62155, ir=54554, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62164, ir=54519, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62178, ir=54484, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62207, ir=54448, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62196, ir=54421, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62217, ir=54424, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62286, ir=54410, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62324, ir=54374, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62333, ir=54369, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62411, ir=54328, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62403, ir=54331, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
red=62411, ir=54267, HR=136, HRvalid=1, SPO2=74, SPO2Valid=1
```

Fig. 20. Output SPO_2 Sensor Data (Valid).

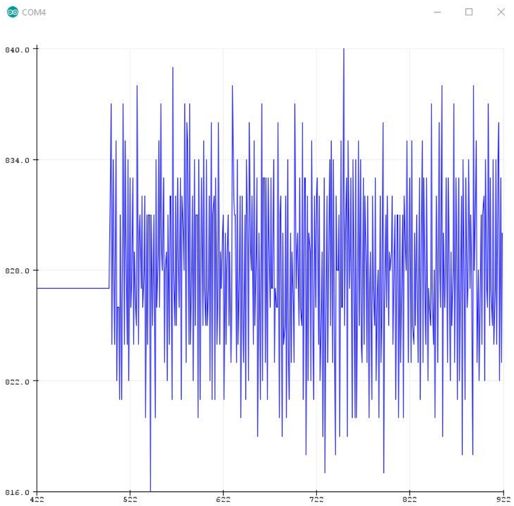


Fig. 21. The Finger is Not Placed on the Sensor, there is no Detection of Arterial Blood.

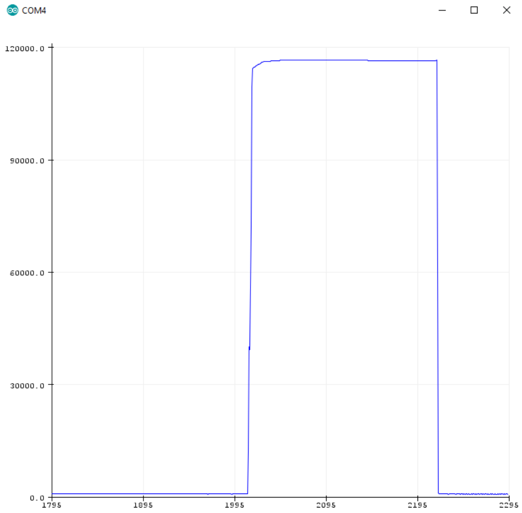


Fig. 22. The Finger is Detected by the Sensor Precisely.

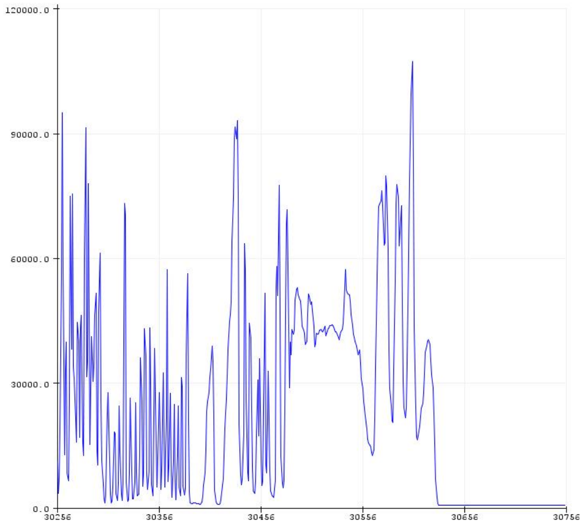


Fig. 23. Using Objects other than Fingers,the Hearth Rate Data is not Accurate.

In Fig. 24, it shows the Power Receiver (-dB) output in the Free Space Path Loss condition at the LoRa 915 MHz Frequency. At a test distance of 1 to 5000 m, the Power Receiver experiences attenuation from ~ 30 dB to ~ 105 dB at a distance of 5 km. furthermore, Fig. 25 is using a different frequency, resulting in the conclusion that with a Frequency of 433 MHz at the same distance can experience attenuation (-dB) which is smaller than other frequencies. Whereas in Fig. 26, that the types of material on the obstacle affect the attenuation signal, the smallest is shown with glass material 6 mm and 13 mm, and the greatest influence on the attenuation signal is Stone Wall and Concrete.

D. Observations using the Signal Analyzer

The MH-ET Live max30102 sensor data is sent using the 915 MHz Dragino LoRa board. 2 output displays are using the Arduino IDE Microcontroller serial monitor and using the LoRa Output with the command. LoRa.beginPacket, furthermore, in this case, the delay must be disabled so that the MH-ET Live max30102 sensor data is sent continuously and Chirp Signal data can be obtained. Fig. 27 and Fig. 28 are an example of a 915 MHz LoRa Spectrum using the Textronix Signal Analyzer. Fig. 28 is accompanied by a process signal demodulation. Furthermore, Fig. 29 is the LoRa Chirp signal in a position with the 10 MHz Span. Chirp shows the different ToA (ms) based on the value of Tpayload and preamble.

Fig. 30 shows the SNR (-dB) and RSSI (-dBm) values in the Receiver. The RSSI (-dBm) value attains attenuation if the distance is further (km).

Fig. 31 is the RSSI of LoRa on indoor Obstacles, these Obstacles i.e, 1 Glass Door, 2 Glass Doors, 3 Glass Doors based on increasingly longer distances (m). (a) is a distance of 10 m without obstacle (influenced by interferences) having an RSSI of ~ -20 to ~ -40 dBm, (b) a distance of 20 m is blocked by 1 glass door having an RSSI of ~ -60 dBm (c) is a distance of 30 m blocked by 3 glass doors having RSSI of ~ -80 to ~ -100 dBm, and (d) a distance of 40 m blocked by 2 doors having RSSI of ~ -70 dBm and (e) returning to initial position ie there is no barrier as in position (a). therefore, from this experiment, it can be seen that RSSI (-dBm) has an attenuation signal by Obstacle which is Glass material, according to Fig. 26.

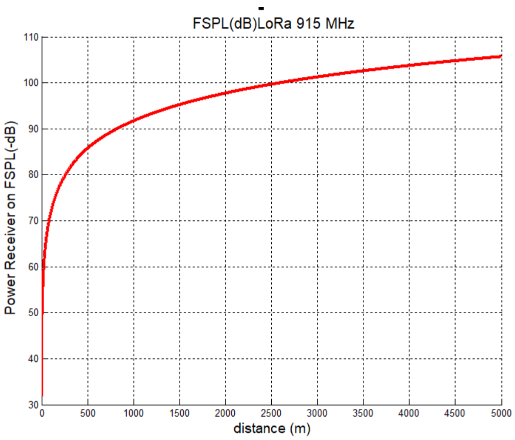


Fig. 24. Power Receiver of LoRa Module 915 MHz.

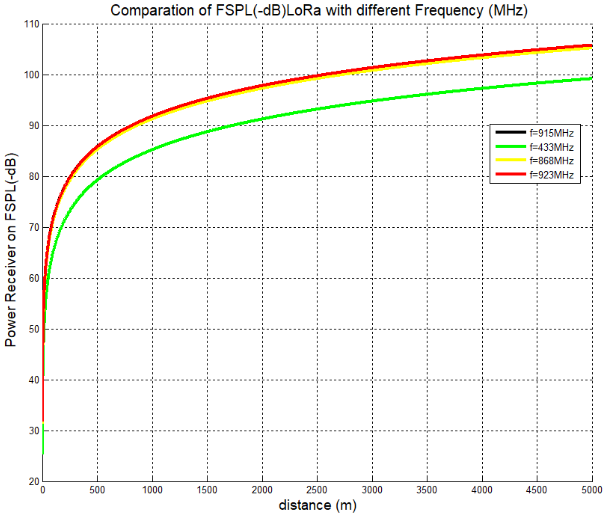


Fig. 25. Power Receiver of LoRa (-dB) Module with different Frequency.

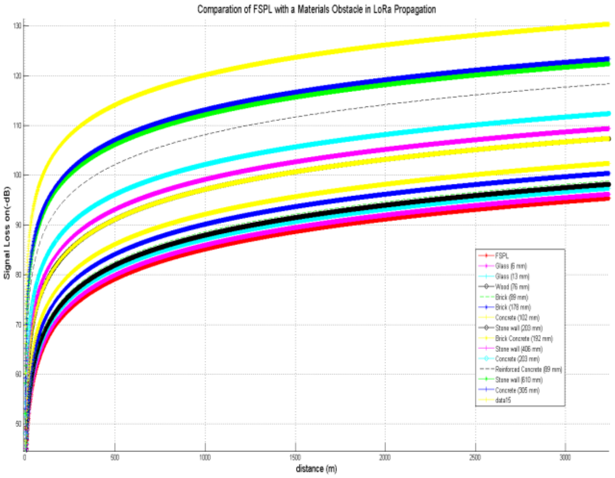
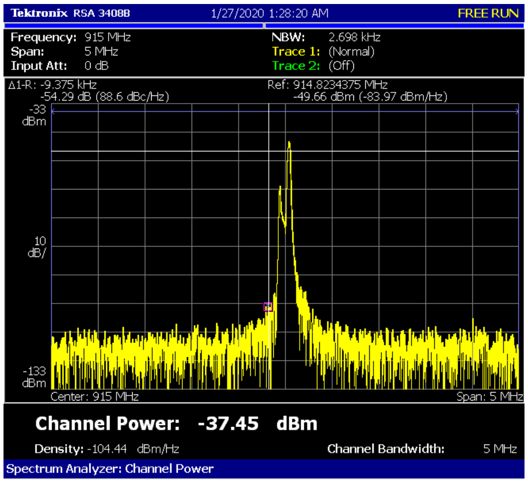


Fig. 26. Comparation of FSPL with Materials Obstacle LoRa Propagation.



Tektronix RSA 3408B 1/27/2020 1:28:20 AM FREE RUN

Frequency: 915 MHz
Span: 5 MHz
Input Att: 0 dB
NBW: 2.698 kHz
Trace 1: (Normal)
Trace 2: (Off)

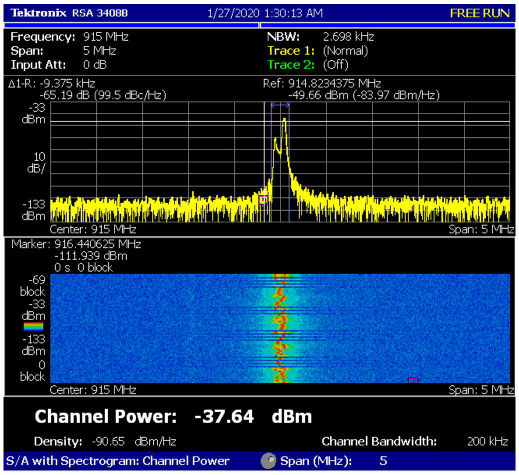
Δ1-R: -9.375 kHz
-54.29 dB (88.6 dBc/Hz)
Ref: 914.8234375 MHz
-49.66 dBm (-83.97 dBm/Hz)

Channel Power: -37.45 dBm

Density: -104.44 dBm/Hz
Channel Bandwidth: 5 MHz

Spectrum Analyzer: Channel Power

Fig. 27. LoRa Signal use a Signal Analyzer.



Tektronix RSA 3408B 1/27/2020 1:30:13 AM FREE RUN

Frequency: 915 MHz
Span: 5 MHz
Input Att: 0 dB
NBW: 2.698 kHz
Trace 1: (Normal)
Trace 2: (Off)

Δ1-R: -9.375 kHz
-65.19 dB (99.5 dBc/Hz)
Ref: 914.8234375 MHz
-49.66 dBm (-83.97 dBm/Hz)

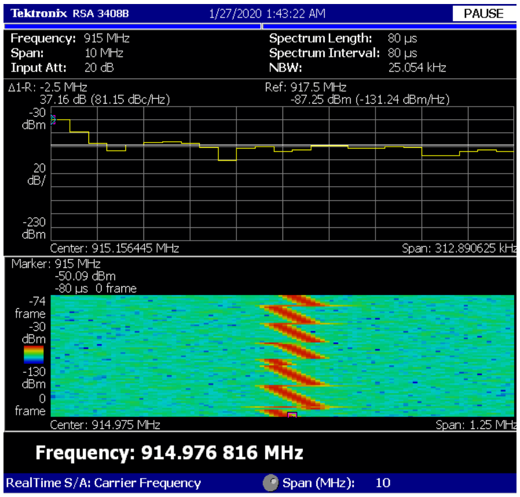
Marker: 916.440625 MHz
-111.939 dBm
0 s 0 block

Channel Power: -37.64 dBm

Density: -90.65 dBm/Hz
Channel Bandwidth: 200 kHz

S/A with Spectrogram: Channel Power
Span (MHz): 5

Fig. 28. LoRa Signal use a Signal Analyzer with Modulation Signal.



Tektronix RSA 3408B 1/27/2020 1:43:22 AM PAUSE

Frequency: 915 MHz
Span: 10 MHz
Input Att: 20 dB
Spectrum Length: 80 µs
Spectrum Interval: 80 µs
NBW: 25.054 kHz

Δ1-R: -2.5 MHz
37.16 dB (81.15 dBc/Hz)
Ref: 917.5 MHz
-87.25 dBm (-131.24 dBm/Hz)

Marker: 915 MHz
-50.09 dBm
-80 µs 0 frame

Frequency: 914.976 816 MHz

RealTime S/A: Carrier Frequency
Span (MHz): 10

Fig. 29. LoRa Chirp Signal.

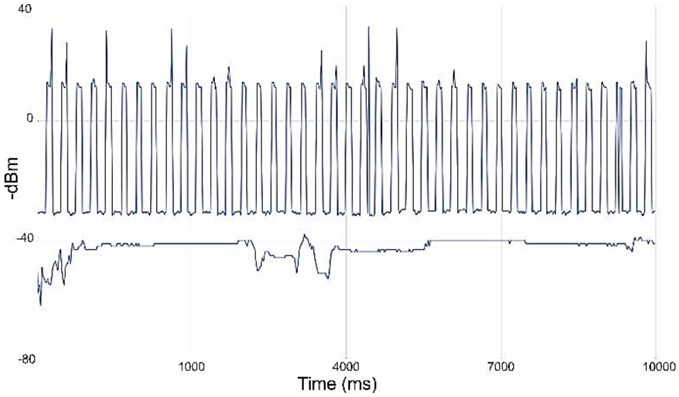


Fig. 30. RSSI and SNR Output on the LoRa Rx.

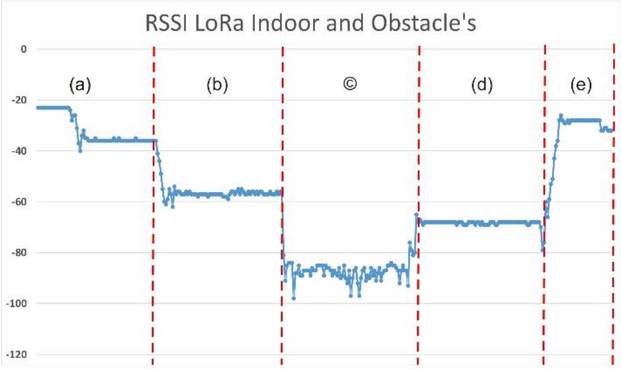


Fig. 31. RSSI (-dBm) of RF96 LoRa on Indoor Obstacles.

E. Consumption Node and ToA Analysis

In this chapter, an analysis of the Consumption sensor node and Time on Air (ToA), with parameters 1. In other research, Power Consumption requires a method or model for efficiency, such as the use of the ARSy Framework Model to protect resources on CPU, Battery, and memory [27].

```
Parameters_1 :
LoRa Modem Setting
Spreading Factor (SF) = 7 - 12
Bandwidth (BW)= 125,250 and 500 kHz
Coding Rate (CR) = 1,2,3 and 4 (4/CR+4)
Low DataRate = ON
Packet Configuration
Payload Length = 8 bytes
Programmed Preamble = 6 symbols
CRC Enabled
RF Settings
Centre Frequency = 915 MHz
Transmit Power (Tx) = 20 dBm
Battery Voltage 3.7 volt, 1000 mAH
Duty Cycle 2000 ms
ACK length 2 byte
Interrogation 4 per day
```

In Fig. 32, 33 and 34, it can be seen that the change in Time on Air (ms) is based on changes in Spreading Factor (SF) and the difference in Bandwidth (BW) gives different values on Time on Air (ms), in all three figures, it can be concluded that with a large bandwidth (kHz), it will have fast (ms) data transferring time. e.g, 500 kHz with SF 12 requires a ToA of ~ 37 ms. Whereas 125 kHz on SF 12 requires a longer ToA of ~ 62 ms. Or on a 250 kHz bandwidth of around ~ 45 ms.

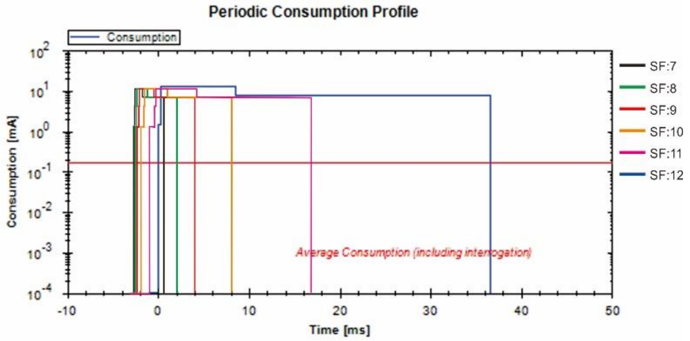


Fig. 32. Periodic Consumption with 500 kHz, CR=1,2,3,4 and SF =7 to 12.

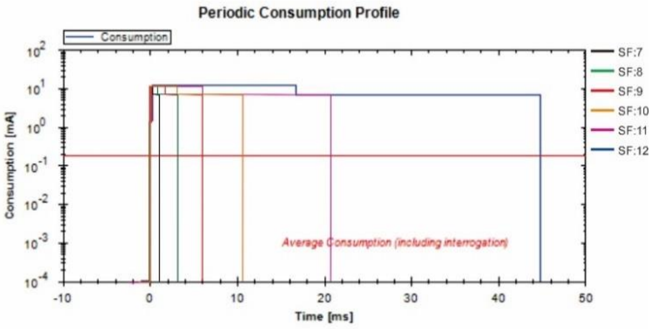


Fig. 33. Periodic Consumption with 250 kHz, CR=1,2,3,4 and SF =7 to 12.

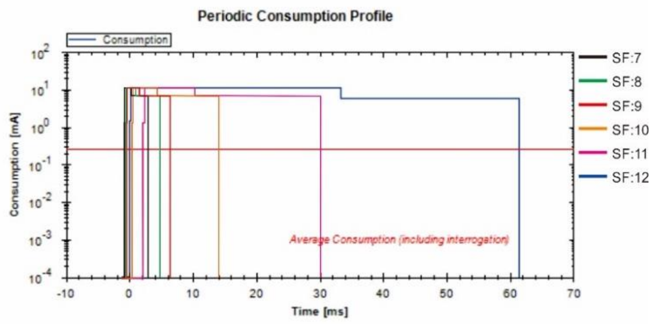


Fig. 34. Periodic Consumption with 125 kHz, CR=1,2,3,4 and SF =7 to 12.

V. Conclusions

The MH-ET Live sensor can produce several outputs, eg, IR Value, SPO_2, and Heart Rate (*BPM*), this sensor is compatible with ATmega 328p MCU with various types e.g., Arduino mini, therefore, right to create light-sized sensor nodes making it easier for users in SPO_2 and Heart Rate (*BPM*) testing. Spreading Factor affects the amount of bitrate (*bps*) and Time on Air (*ms*), the greater the bandwidth, the faster the process of transmitting data (*ms*). Attenuation when the process of transmitting sensor data (Signal propagation) is influenced by the type of material that becomes the obstacle signal T_x to Rx at a certain distance. The farther the distance, the greater attenuation (*dB*), e.g. FSPL LoRa at 3 km without obstacle is ~ 80 dB. Whereas with an obstacle at the same distance, attenuation occurs to ~ 130 dB. furthermore, The Two Ray Ground Reflection model is only used if the position of the Antenna LoRa transmitter (T_x) is at a certain height (H_T) which affects the reflection signal which causes attenuation which affects the signal strength at the receiver (R_x).

Acknowledgment

Thanks for Micro Electronics Research Laboratory (MeRL) Kanazawa University, JAPAN for supporting through a materials e.g., devices, analyzer. and to Professor Akio Kitagawa who has guided the writer and has provided support and role models.

References

[1] Prasetya, D.A, Yasuno, T. “Cooperative control of multiple mobile robot using particle swarm optimization for tracking two passive target”, Proceedings of the SICE Annual Conference, 2012.

[2] Prasetya, D.A, Yasuno, T, Zhang, “Cooperative tracking control of multiple mobile robot for moving target using particle swarm optimization”, Proceedings of the SICE Annual Conference, 2013.

[3] Puput Dani Prasetyo Adi, Rahman Arifuddin ,"Design Of Tsunami Detector Based Sort Message Service Using Arduino and SIM900A to GSM/GPRS Module", JEEMECS (Journal of Electrical Engineering, Mechatronic and Computer Science) Volume 1, No.2, 2019, ISSN : 2614-4859, DOI:10.26905/jeemecs.v1i1.1982.

[4] Prasetya, D.A, Nguyen, P.T, Faizullin, R, Iswanto, I, Armay, E.F "Resolving the shortest path problem using the haversine algorithm", Journal of Critical Reviews, Volume 7, Issue 1, 2020.

[5] Puput Dani Prasetyo Adi and Akio Kitagawa, "Performance Evaluation WPAN of RN-42 Bluetooth based (802.15.1) for Sending the Multi-Sensor LM35 Data Temperature and RaspBerry Pi 3 Model B for the Database and Internet Gateway" International Journal of Advanced Computer Science and Applications(IJACSA), 9(12), 2018. DOI: 10.14569/IJACSA.2018.091285.

[6] Mehrdad Babazadeh, "Edge Analytics for anomaly detection in water network by an Arduino 101-LoRa based WSN", ISA Transactions Article, February, 2019, DOI. 10.1016/j.isatra.2019.01.015.

[7] Puput Dani Prasetyo Adi and Akio Kitagawa, "Performance Evaluation of E32 Long Range Radio Frequency 915 MHz based on Internet of Things and Micro Sensors Data" International Journal of Advanced Computer Science and Applications(IJACSA), 10(11), 2019. doi/10.14569/IJACSA.2019.0101106.

[8] Jetmir Haxhibeqiri, Floris Van den Abeele, Ingrid Moerman and Jeroen Hoebeke, "LoRa Scalability : A Simulation Model Based on Interference Measurements", Sensor MDPI, DOI. 10.3390/s17061193.

[9] Amir Muaz Abdul Rahman, Fadhlan Hafizhelmi Kamaru Zaman, Syahrul Afzal Che Abdullah, "Performance Analysis of LPWAN Using LoRa Technology for IoT Application", International Journal of Engineering & Technology, 7(4.11)(2018) 212-216.

[10] Borja Martinez, Marius Monton, Ignasi Vilajosana, Joan Daniel Prades, "The Power of Models: Modeling Power Consumption for IoT devices", IEEE Sensor Journal, DOI. 10.1109/JSEN.2015.2445094.

[11] Muhammad Niswar, Amil Ahmad Ilham, Elyas Palantei, Rhiza S.Sadjad, Andani Ahmad, Ansar Suyuti, Indrabayu, Zaenab Muslimin,Tadjuddin Waris, Puput Dani Prasetyo Adi, "Performance evaluation of ZigBee-based wireless sensor network for monitoring patients' pulse status", 2013 International Conference on Information Technology and Electrical Engineering (ICITEE) DOI: doi/10.1109/ICITEED.2013.6676255.

[12] Puput Dani Prasetyo Adi and Akio Kitagawa, "ZigBee Radio Frequency (RF) Performance on Raspberry Pi 3 for Internet of Things (IoT) based Blood Pressure Sensors Monitoring" International Journal of Advanced Computer Science and Applications(IJACSA), 10(5), 2019. DOI: 10.14569/IJACSA.2019.0100504.

[13] Puput Dani Prasetyo Adi and Akio Kitagawa, "Quality of Service and Power Consumption Optimization on the IEEE 802.15.4 Pulse Sensor Node based on Internet of Things" International Journal of Advanced Computer Science and Applications (IJACSA), 10(5), 2019. DOI: 10.14569/IJACSA.2019.0100518.

[14] Mehmet Ali Erturk, Muhammad Ali Aydin, Muhammet Talha Buyukakkaslar, and Hayrettin Evirgen, "A Survey on LoRaWAN Architecture, Protocol and Technologies", Future Internet MDPI, 2019, 11, 216, DOI: 10.3390/fi11100216.

[15] Nurhayati, Muhammad Suryanegara, "The IoT LoRa system design for tracking and monitoring patient with mental disorder", DOI: 10.1109/COMNETSAT.2017.8263587.

[16] Dong-Hoon Kim, Eun-kyu Lee, Jibum Kim, "Experiencing LoRa Network Establishment on a Smart Energy Campus Testbed", Sustainability MDPI, 2019, 11, 1917, DOI. 10.3390/su11071917.

[17] Jansen Christiano Liando, Amalinda Gamage, "Known and Unknown Fact of LoRa : Experiences from a Large scale Measurement Study", ACM Transactions on Sensor Networks, February 2019, DOI.10.1145/3293534.

[18] Jetmir Haxhibeqiri, Eli De Poorter, Ingrid Moerman, and Jeroen Hoebeke, "A Survey of LoRaWAN for IoT : From Technology to Application", Sensor MDPI, Sensor 2018, 18, 2995; DOI:10.3390/s18113995.

[19] Philip A.Catherwood, David Steele, Mike Little, Stephen Mccomb, and James Mclaughlin, "A Community-Based IoT Personalized Wireless Heathcare Solution Trial", Point-Of-Care Technologies, IEEE Journal of Translational Engineering in Health and Medicine, 2018.

[20] Dmitry Bankov, A.Lyakhov, Evgeny Khorov, "Mathematical model of LoRaWAN channel access", 2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), DOI. 10.1109/WoWMoM.2017.7974300.

[21] Irawan Sukma, Akio Kitagawa, "Comparison Topologies of Resonant Tank from Class-C Wireless Power Transfer", 2018 IEEE International Workshop on Electromagnetics: Applications and Student Innovation Competition (iWEM), DOI: 10.1109/iWEM.2018.8536684.

[22] MD Hossinuzzaman, Dahlila Putri Dahnil, "Enhancement of Packet Delivery Ratio during Rain Attenuation for Long Range Technology", International Journal of Advanced Computer Science and Applications (IJACSA), Vol.10. No. 10, 2019.

[23] Orion Afisiadis, Matthieu Cotting, Andreas Burg, Alexios Balatsoukas-Stimming,"On the Error Rate of the LoRa Modulation with Interference", IEEE Transactions on Wireless Communications, DOI: 10.1109/TWC.2019.2952584.

[24] Muhammad Fauzan Edy Purnomo, Akio Kitagawa, "Triangular Microstrip Antenna for Circularly-Polarized Synthetic Aperture Radar Sensor Application", DOI. 10.11591/ijeecs.v12.i1.pp310-318.

[25] Taoufik Bouguera, Jean-Francois Diouris, Jean-Jacques Chaillout, Randa Jaouadi, and Guillaume Andrieux, "Energy Consumption Model for Sensor Nodes Based on LoRa and LoRaWAN", Sensor MDPI, 2018, 18, 2104, DOI. 10.3390/s18072104.

[26] Mariusz Slabicki, Gopika Premsankar, Mario Di Francesco, "Adaptive Configuration of LoRa Networks for dense IoT deployment", NOMS2018 IEEE/IFIP Network Operations and Management Symposium, DOI. 10.1109/NOMS.2018.8406255.

[27] Jumadi Mabe Parengreng, Akio Kitagawa, Dyah Darma Andayani, "A Study of Limited Resources and Security Adaptation for Extreme Area in Wireless Sensor Networks", Journal of Physics Conference Series 1244:012013, DOI. 10.1088/1742-6596/1244/1/012013.

Comparison of Anomaly Detection Accuracy of Host-based Intrusion Detection Systems based on Different Machine Learning Algorithms

Yukyung Shin[1], Kangseok Kim[1, 2*]
Department of Data Science, Graduate School of Ajou University, Suwon, Korea[1]
Department of Cyber Security, Ajou University, Suwon, Korea[2]

Abstract—**Among the different host-based intrusion detection systems, an anomaly-based intrusion detection system detects attacks based on deviations from normal behavior; however, such a system has a low detection rate. Therefore, several studies have been conducted to increase the accurate detection rate of anomaly-based intrusion detection systems; recently, some of these studies involved the development of intrusion detection models using machine learning algorithms to overcome the limitations of existing anomaly-based intrusion detection methodologies as well as signature-based intrusion detection methodologies. In a similar vein, in this study, we propose a method for improving the intrusion detection accuracy of anomaly-based intrusion detection systems by applying various machine learning algorithms for classification of normal and attack data. To verify the effectiveness of the proposed intrusion detection models, we use the ADFA Linux Dataset which consists of system call traces for attacks on the latest operating systems. Further, for verification, we develop models and perform simulations for host-based intrusion detection systems based on machine learning algorithms to detect and classify anomalies using the Arena simulation tool.**

Keywords*—*Anomaly detection; host based intrusion detection system; system calls; cyber security; machine learning; simulation

I. Introduction

Owing to the recent developments in the fields of software, hardware, and mobile networks, as well as the proliferation of information services, such as social network services (SNS), people are now more closely connected to the Internet than ever before. However, this extensive use of information systems over the Internet has exposed us to many threats, including hacking and malicious software (malware), such as ransomware. To mitigate such threats, a firewall, which forms an essential part of any Internet and network security system, prevents intrusions from external networks to internal networks or devices on those networks; nevertheless, these networks are still considerably vulnerable to other attacks, such as Denial of Services (DoS) attacks that cannot be prevented by a firewall [1]. Furthermore, another disadvantage of firewalls is that they block only some of the hacking attacks that are made against a system or network. Considering this drawback and owing to the emergence of intelligent cyberattacks, the importance of attack detection and security on systems and networks has significantly increased in recent times. Thus, intrusion detection systems (IDS) [2], which have been studied for a considerable time, have been developed as next-generation security systems against hacking methods.

In general, network packets pass through the IDS after passing through the firewall, and the IDS generates an alarm if it detects malicious activities or determines anomalies in the incoming data [3]. Therefore, an IDS has a role similar to that of a firewall, but it also detects internal hacking and malicious codes that the firewall cannot detect and defend against. In addition, the IDS detects and responds to unauthorized activities against target systems that are not certified [4]. Thus, an IDS is an important tool for detecting security violations in real time. An IDS can be classified into two types: a host-based IDS (HIDS) and network IDS (NIDS) based on the position and purpose of the detection area according to datasource-based classification [2]. In order to detect malicious behaviors such as DoS attacks and port scans, an HIDS analyzes information collected from specific host systems, while an NIDS monitors network traffic [5]. Unlike the NIDS which detects attack vectors based on network traffic, the HIDS focuses on monitoring and analyzing the internal system, instead of the external network.

A HIDS can further be classified according to the type of model used for intrusion detection, namely misuse detection method and anomaly (or behavior) detection method. Both use information extracted from the analysis target to determine if an intrusion has occurred [6, 7, 8]. The misuse detection method, which is used in a signature-based (or knowledge-based) HIDS, is effective in detecting known attack vectors; nevertheless, it is vulnerable to attacks from unknown attack vectors. Therefore, there is a need for the anomaly detection methods [8, 9]. In particular, anomaly detection methods define and detect any anomalies that deviate from normal behavior patterns based on existing network usage scenarios, internal system calls, and so on.

In order to define normal behavior patterns, it is, therefore, necessary to extract normal behavior and anomaly patterns in HIDS. Then, machine learning algorithms based on iterative learning or data mining can be used to develop intrusion detection models using mathematical and statistical methods on these extracted patterns. Extensive research has been performed on applying data mining techniques on the new dataset to develop models for HIDS [8] as well as on network traffic data to develop models for NIDS [7]. Furthermore, the existing HIDS design suffers from the problem of a high false alarm

[*] Corresponding Author.

rate, thereby increasing the detection rate [8]. Since the approach suggested by [10], the works to reduce the false alarm rate based on system calls (which are interactions between programs and kernel) patterns in HIDS have resulted in a lot of researches [8]. However, it should be noted that the accuracy of the methods is not sufficiently high still.

Therefore, the objective of this study focuses on improving the accuracy of attack detection by applying the three different machine learning approaches to data preprocessed from system call sequence dataset released by [11]. Then the N-gram [12] method, which is one of data representation techniques, is used to preprocess the system call sequence dataset. In addition, after applying and comparing the results of various machine learning algorithms with the preprocessed data, we propose the most suitable machine learning algorithm model that improves the intrusion detection accuracy of HIDS. Furthermore, we verify the anomaly detection accuracy of the proposed HIDS models by performing simulations using the Arena simulation tool [13].

The remainder of the paper is organized as follows. Section II discusses the experimental datasets which are the system call sequence data released by [11] as well as previous studies that integrate machine learning algorithms and intrusion detection systems. The data preprocessing using N-gram method and the machine learning algorithms to be applied are explained in Section III. Section IV describes the experiments conducted in our study using the preprocessed datasets with various machine learning algorithms. Section V presents information on the verification tasks performed via simulations as well as the experimental results. Finally, Section VI provides our conclusions and directions for future research.

II. Datasets and Related works

A. Data Collection

Among the various experimental datasets used for research on HIDS, the publicly-available knowledge discovery and data mining (KDD) cup 99 datasets [14] has provided a systematic approach to forming intrusion detection system data. However, over time, this dataset has become outdated, and thus, many have criticized its use or are skeptical about applying it to current Internet environments [11]. Since computer and network systems have evolved, new attack vectors and vulnerabilities have emerged. Therefore, HIDS developed on the basis of existing datasets does not properly take into account the features of current attack vectors, so these existing datasets are not suitable for HIDS evaluation and validation [15].

Thus, alternative datasets reflecting current attack vectors have been proposed in [11]; an example of such a dataset is the research dataset provided by the Australian Defense Force Academy (ADFA) [11]. In many recent works, the ADFA dataset along with the latest attack vector features have been used for research on intrusion detection verification. In particular, the ADFA dataset was developed to evaluate a system call based HIDS as well as anomaly detection in signature-based HIDS.

The ADFA dataset is divided into the ADFA Linux dataset (ADFA-LD) and ADFA Windows dataset (ADFA-WD). The ADFA-LD reflects the features of current Linux-based operating systems, compared to many existing datasets used to evaluate the HIDS, and consists of thousands of system call traces collected from Linux local servers for the most recent attacks and vulnerabilities that occur in various applications. Considering this, the ADFA-LD is expected to become a new benchmark for evaluating and verifying HIDS.

Thus, in this study, using ADFA-LD, we extracted the attack patterns against the current HIDS and applied the machine learning and data mining techniques to the patterns to improve the accuracy of the attack and anomaly detection of the HIDS.

As previously mentioned, the ADFA-LD has thousands of normal traces collected from hosts on Linux servers, including abnormal trace files for six new types of cyberattacks, general user behavior and cyberattack path, and audit daemon setup, among others. In particular, during sampling periods for the ADFA-LD, a host captures system call traces that are generated by normally functioning legitimate programs and stores the corresponding data in a file. Among them, 8-20 abnormal call traces are stored as attack data files using call traces generated after a cyberattack is initiated against the test host. As listed in Table I, the ADFA-LD consists of three different data groups, each of which contains their own system call trace files. These data groups include training data master (TDM) and validation data master (VDM) groups, which represent normal data, whereas attack data master (ADM) group consists of call traces representing attack data. Furthermore, the ADM consists of six types of attack data: “Adduser”, “Hydra-FTP”, “Hydra-SSH”, “JavaMeterpreter”, “Meterpreter”, and “Web-Shell”.

B. Related Works

A number of system calls-based anomaly detection models have been designed to increase the accurate detection rate and to reduce the false alarm rate in HIDS. The paper [16] provides a survey of the host-based intrusion detection system with system calls, from the viewpoint of algorithms, techniques, datasets, application areas, and future research trends to inspire researchers about system-call-based HIDS in the big data and cloud environment. Also the paper [17] reviewed the research regarding intrusion detection techniques based on the HMM and provided challenges in this field. In this section we discuss existing works which integrate machine learning models and host-based anomaly detection systems.

Many studies have assessed the use of Hidden Markov Model (HMM), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN) algorithms, and so on for attack pattern recognition to improve the HIDS by reducing its false alarm rate and increasing the detection rate. As mentioned in the paper [18], the use of the Sequence Time-Delay Embedding (STIDE) algorithm in [8, 10] was problematic because the STIDE algorithm requires re-training for each particular process, and still has a high false alarm rate to any system call sequence data which do not appear in the training data. In [19], a scalable anomaly detection model was developed for servers in a cloud environment. The research proposed a nested-arc hidden semi-Markov model (NAHSMM) based on HHMM [20] which is a method for detecting anomalies in cloud servers. The anomaly detection algorithm is derived by integrating state summarization and NAHSMM and was evaluated with NGIDS-DS [21] dataset. The dataset is composed of labeled host logs and network packets. The model can work effectively with a smaller number of training samples and less processing time than recurrent neural networks (RNNs). In future work, with the NGIDS-DS data set and other data representations, we will consider comparing the study with various machine learning algorithms and deep learning models such as LSTM (Long Short-Term Memory) model and GRU (Gated Recurrent Unit) model.

In another study [15], considering the advancements in computer systems, as a preliminary work, researchers used the ADFA-LD dataset to evaluate a new host-based anomaly detection system (HADS) instead of the older datasets that were previously used. The common patterns and frequency of attacks were evaluated by the KNN-based HADS with the AFDA-LD dataset. Although acceptable detection results were obtained for some attacks by their proposed HADS, it still had a weakness in that it could not identify the behaviors of some attacks from normal behavior through the KNN algorithm.

In [22], researchers developed a frequency-based misuse detection method using ensemble classification. After preprocessing the raw ADFA-LD system call traces using the N-gram method, patterns were generated by extracting features; in addition, the number of patterns were balanced based on class through the synthetic minority over-sampling technique (SMOTE) [23]. The classification of temporal sequences with data-driven method do not need parameter estimation [12]. Therefore, in future work we will have to consider configuring a N-gram matrix that reflects well the data structure. Furthermore, their classification design was based on a majority voting ensemble technique of Naive Bayes, SVM [24], PART [25], Decision Tree, and Random Forest algorithms as well as Principal Component Analysis (PCA); their proposed misuse intrusion detection method showed good performance in terms of attack detection. Also host-based anomaly intrusion detection by Radial Basis Function neural network and Random Forest was conducted. The simulation study showed good performance in terms of detecting anomalies and normal activities.

The researchers in [26] used various machine learning classification algorithms to extract patterns from labelled new generation system call traces for modern exploits and attacks because they considered anomaly detection in ongoing processes using system call traces as a typical pattern recognition problem for machine learning. They evaluated the performance of the enhanced vector space representation technique for the ADFA-LD and their results showed good performance in distinguishing process behavior from exploits and attacks by using system calls.

TABLE. I. DATA GROUPS IN THE ADFA-LD DATASET

Data Groups	Type of Traces	Number of Traces
TDM	Normal	833
VDM	Normal	4372
ADM	Adduser	91
	Hydra-FTP	162
	Hydra-SSH	176
	JavaMeterpreter	124
	Meterpreter	75
	Web-Shell	118

III. Proposed HIDS Detection Method

Although the performance of an intrusion detection method based on misuse detection has been verified in a previous work [19, 24], to the best of our knowledge, there is a lack of machine learning approach on anomaly intrusion detection methods. Therefore, we study the classification of extracted system call sequence data into normal or malicious behaviour using machine learning algorithms. Section A describes data preprocessing using the N-gram method, and Section B presents the various machine learning algorithms used in the study. Fig. 1 shows a methodology of anomaly detection system using various machine learning approaches conducted in our study with the ADFA-LD dataset.

A. Data Preprocessing

The extracted system call trace data [11] consists of a series of numbers corresponding to system calls made on the Linux operating system. We apply machine learning algorithms to the system call trace data and then classify the process operation into normal behaviour or six specific attack types. First, we used the N-gram technique to extract attribute vectors from the system call trace dataset. The N-gram method involves cutting a sample text into a contiguous sequence of N characters or words. For an N-gram of size 1, i.e., N = 1, the N-gram is referred to as Uni-gram (1-gram), while for an N-gram of size 2, i.e., N=2, the N-gram is referred to as Bi-gram (2-gram). In this study for N-gram, a word units consist of system call numbers, and the number of system call sequence attributes is derived by creating an array of N words according to the given word order. By doing this step repetitively, the call attributes of the system call traces can be obtained. Fig. 2 shows an example of applying the N-gram technique on system call trace data.

In particular, N-gram data is expressed as a two-dimensional matrix; the columns of this matrix consist of the attribute values by matching the entire word belonging to each gram according N, while the rows represent instances that belong to each trace. The value corresponding to the row and column of the data represents the number of occurrences of N-gram in each trace as shown in Table II. As the value of N increases, the model becomes more complicated and requires considerably more storage space, thereby increasing processing time. Therefore, in this study, we limited N to 1 to 5. Furthermore, in order to extract those instances that occur most frequently in an entire trace, we extracted and used only instances that they were used more than once in the entire trace and had more than 30% (0.3) of all the instances of in the entire trace because the used instances were small.

B. Applied Machine Learning Algorithms

After pre-processing the data using the N-gram technique, we detected anomalies in the system call trace data using machine learning algorithms, based on which, classified and predicted normal data and six types of attack data from new system call trace files. In order to apply machine learning algorithms, we divided the dataset into training and test data. Then, the training data was used to model the algorithm, and the test data was used to validate the algorithm to ensure accuracy. In our study, the ADM dataset, which is the attack data to be detected, was used in a 7: 3 ratio for the training to test data.

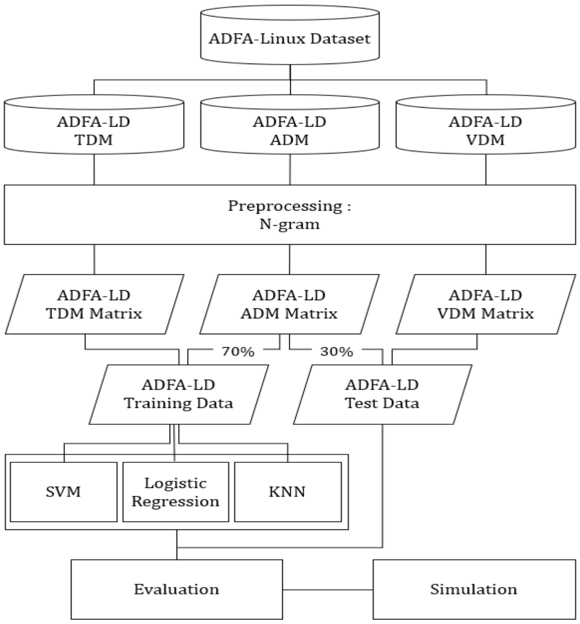


Fig. 1. Methodology Flow of Evaluation (Training/Testing) and Simulation Performed for the Proposed Anomaly Detection.

1-gram: 6/ 11/ 45/ 33/ 192/ 33/ 5/ ⋯

6 11 45 33 192 33 5 ⋯ ⟹ 2-gram: 6 11/ 11 45/ 45 33/ 33 192/ ⋯

3-gram: 6 11 45/ 11 45 33/ 45 33 192/ ⋯

Fig. 2. An Example of N-Gram units.

TABLE. II. THE NUMBER OF OCCURRENCES OF N-GRAM IN EACH TRACE

N-gram	**1-gram**				⋯	**5-gram**			
System Call Trace	x_1	x_2	⋯	x_α	⋯	$x_{1'}$	$x_{2'}$	⋯	$x_{\alpha'}$
d_1	$N_{1,1}$	$N_{1,2}$	⋯	$N_{1,\alpha}$	⋯	$N_{1,1'}$	$N_{1,2'}$	⋯	$N_{1,\alpha'}$
d_2	$N_{2,1}$	$N_{1,2}$	⋯	$N_{1,\alpha}$	⋯	$N_{2,1'}$	$N_{1,2'}$	⋯	$N_{1,\alpha'}$
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
d_m	$N_{m,1}$	$N_{m,2}$	⋯	$N_{m,\alpha}$	⋯	$N_{m,1'}$	$N_{m,2'}$	⋯	$N_{m,\alpha'}$

The SVM algorithm proposed by Boser in 1992 is one of the most successful classification algorithms in the field of data mining. In particular, the SVM algorithm is a method of finding a hyperplane that maximizes the margins that are farthest from data among the hyperplanes that dichotomically divide data based on training data [27]. The SVM algorithm is based on a kernel function, which can solve large dimension issues, i.e., SVM does not suffer from problems associated with high dimensionality; in addition, the generalization ability of the SVM method can be enhanced by increasing margins during the training process [28]. Considering these features of the SVM algorithm, we considered it suitable to classify the data which was represented using a large-sized matrix preprocessed by the N-gram technique with the SVM algorithm for experiments.

Furthermore, the logistic regression algorithm uses the predictive model by representing the relationship between dependent and independent variables as a function. It is similar to the linear regression algorithm; however, unlike the linear regression algorithm, which uses continuous data, the dependent variables in the logistic regression algorithm are categorical data [29]. Thus, the logistic regression algorithm is considered useful for classifying categorical data and labeling the normal data and the six types of attack data considered in this study. Therefore, the logistic regression algorithm was used as another machine learning algorithm for experiments.

The third machine learning algorithm used in the study is the KNN algorithm which is a pattern recognition algorithm widely used for classification and regression. In this method, for classification, the input consists of the nearest K training data within a feature space and provides the class membership as output [30]. The KNN algorithm was used as a machine learning algorithm to detect anomalies in the system call trace data because it is effective in classifying data by labeling it as normal and attack data and specifying K.

In next section we will show the experimental results obtained from using the preprocessed datasets with various machine learning algorithms mentioned in this section.

IV. EXPERIMENTS WITH THREE DIFFERENT MACHINE LEARNING APPROACHES

In this study, we conducted experiments using the three machine learning algorithms: SVM, Logistic Regression, and KNN. The TDM and VDM data groups are the normal data, while the ADM group consists of the six types of attack data, which are listed along with their labels in Table III. After modeling the training data and the label of the training data using the machine learning algorithms, the label of the test data can be predicted with the test data.

For the SVM algorithm, we conducted experiments using the Linear function, Polynomial function, Sigmoid function, and Radial Basis Function (RBF) as kernel functions. The label prediction accuracy for the six types of attack and normal data as well as the time taken after applying each kernel of the SVM algorithm are listed in Table IV. By setting hyperparameter C, we can confirm the label prediction accuracy of the SVM algorithm based on C.

Furthermore, the prediction accuracy of labeling and the time taken after applying each kernel for the logistic regression algorithm are listed in Table V. In a manner similar to the SVM approach, by setting parameter C, we can confirm the label prediction accuracy of logistic regression based on C.

TABLE. III. ADFA-LD DATA LABELING

Data Groups	Type of Traces		Labeling
TDM	Normal		0
VDM	Normal		0
ADM	Attack	Adduser	1
		Hydra-FTP	2
		Hydra-SSH	3
		JavaMeterpreter	4
		Meterpreter	5
		Web-Shell	6

TABLE. IV. ACCURACY AND TIME TAKEN FOR LABELING DATA USING SVM

Kernel	Parameter C	Accuracy (%)	Time (sec)
Linear	0.1	79.5648	27.08
	1	79.0917	33.33
Polynomial	0.1	78.8079	24.79
	1	79.1864	26.15
Sigmoid	0.1	78.8079	22.48
	1	78.5241	22.78
RBF	100	80.9839	24.72
	1,000	82.6869	26.21
	10,000	80.5109	35.19

TABLE. V. ACCURACY AND TIME TAKEN FOR DATA LABELING USING LOGISTIC REGRESSION

Parameter C	Accuracy (%)	Time (sec)
0.1	76.4427	1.44
1	78.9025	1.49
10	78.1457	1.84
100	78.7133	2.98

TABLE. VI. ACCURACY AND TIME TAKEN FOR DATA LABELING USING KNN (K=7)

KNN type	Accuracy (%)	Time (sec)
BallTree	83.4437	3.35
KDTree	82.5922	2.44
Brute-force	84.7682	0.55

We also conducted experiments by applying the KNN algorithm for labeling data using three different KNN approaches, namely BallTree, KDTree, and Brute-force Search. The prediction accuracy and time taken for labeling each of these KNN algorithm types are listed in Table VI.

We evaluated the performance of each model using the AUROC curve (Area Under the ROC curve) in order to compare the predictions of each machine learning algorithm for applying to simulation described in Section V. Fig. 3 shows the AUROC curves for the SVM, Logistic Regression, and KNN machine learning algorithms. Table VII shows the summary of the highest prediction accuracies of the different models. In particular, the SVM with an RBF kernel and C= 1,000 has the highest AUC of 0.95 among the SVM kernels (Table IV); similarly, the Logistic Regression model with C=1 (Table V) and KNN Brute-force model (Table VI) have the corresponding highest accuracy based on the AUC model performance.

Fig. 3 depicts ROC Curve of Class N (where N = 0, 1, 2, 3, 4, 5, 6) expressed according to labels listed in Table III. The first figure in Fig. 3 representing the AUROC curve of the RBF model among the kernels of the SVM algorithm shows that the overall model performance is 95%. The second figure in Fig. 3 representing the AUROC curve when the parameter C of the Logistic Regression algorithm is set to 1 shows that the model performance is 96%. The last figure in Fig. 3 representing the KNN AUROC Curve using the Brute-force approach shows that the model performance is 96%. Overall, the AUC performance of the modeled machine learning algorithms is over 95% in all cases, which indicates that they are suitable for the machine-learning-algorithm-based HIDS model after appropriate data preprocessing and pattern extraction.

TABLE. VII. HIGHEST ACCURACY AND AUC PERFORMANCE OF THE APPLIED MACHINE LEARNING ALGORITHMS

KNN type	Accuracy (%)	AUC
SVM_RBF (C=1,000)	82.6869	0.95
Logistic Regression (C=1)	78.9025	0.96
KNN_Brute-force	84.7682	0.93

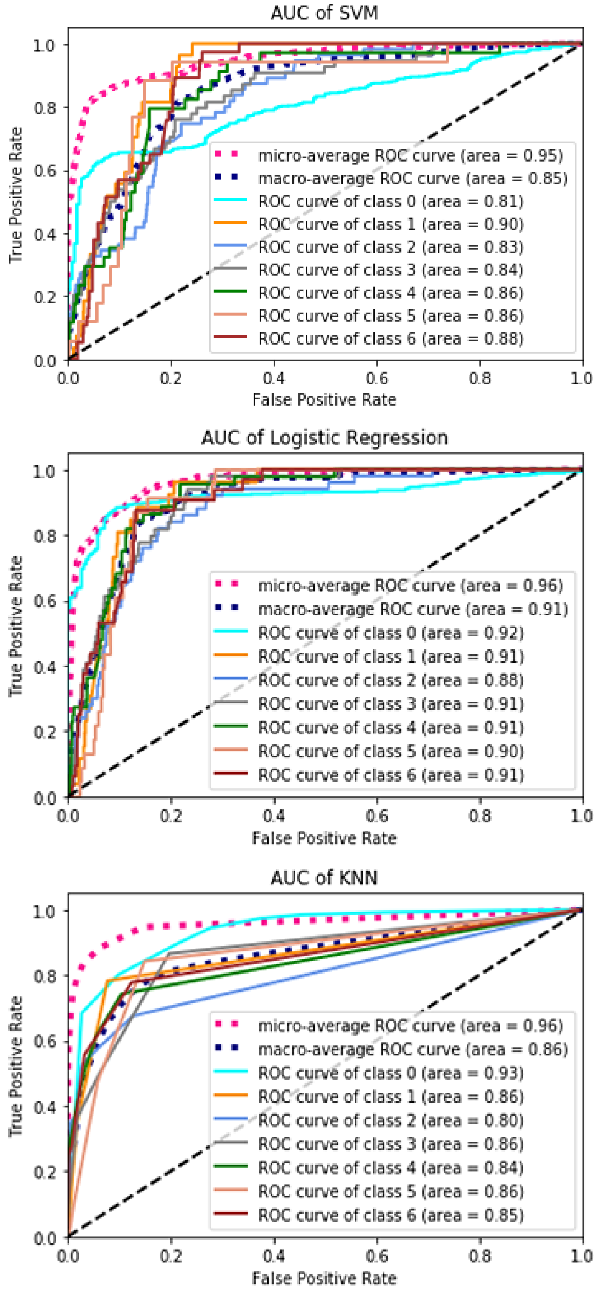


Fig. 3. AUROC of Applied Machine Learning Algorithms.

Our comparison experiments on the three machine learning algorithms - SVM, Logistic Regression, and KNN - indicate that there is significant difference in model performance when the Logistic Regression and KNN algorithms are employed, which can be attributed to their similarity. However, the Logistic Regression algorithm shows the best model performance with the performance values of most labeling (class) models reaching over 90%. Thus, the logistic regression algorithm is the most suitable one in terms of model performance; however, we did not confirm the reason for the model using Brute-force KNN algorithm having a high prediction accuracy.

V. Verification Simulation and Result Analysis

In practice, a HIDS can be installed and operated on different operating systems including on servers as well as clients. However, if simulation experiments for intrusion detection are conducted by directly installing the HIDS on personal computers, several problems need to be considered, including cost incurred because of performance, virus infections, or host malfunctions arising from IDS errors. Considering these possible issues, in this study, we verify the performance of the machine-learning-algorithm-based HIDS via simulations, which are quite similar to performing verification on actual systems. In particular, we constructed the HIDS model using the Arena simulation software, which is a proprietary software [13].

Arena simulation provides a simulation and animation environment designed to model discrete / continuous event system. The simulation system is easy to configure because the proprietary code is used to create models consisting of blocks and elements without the need for any additional code. In addition, these blocks are organized in a flow chart format; therefore, it facilitates easy progress monitoring [31]. The manner in which system calls are used in the HIDS is depicted in Fig. 4. At the user application level, the system call, read(), initiates a system call, such as sys_read(), at the kernel level through the HIDS. Furthermore, as indicated in Fig. 4, we can develop simulations by assigning the installation location of the HIDS to the system call interface that is placed from the user application level to the kernel level. The schematic design of the HIDS simulation model is shown in Fig. 5. The model reads the system call patterns of the normal data and attack data using the read-write module “System Call” after being initiated by the user module “user”. Then, the sub-model “HIDS” classifies the data based on the accuracy results of SVM, Logistic Regression, and KNN, which are the three machine learning algorithms used in the experiments that are described in Section IV. After classifying in the sub-model, with the six types of attack data and normal data, they are classified as follows: ‘Attack’ which is classified as attack, ‘Normal’ that is classified as normal, and ‘MissAttack’ in which the six types of attack are misclassified as normal.

The schematic design of the sub-model “HIDS” is shown in Fig. 6. The data read through the “System Call” module is classified by the “Classify” module into the six types of attack data and normal data with the names (“IsAdduser,” “IsHydraFTP,” “IsHydraSSH,” “IsJavaMeterpreter,” “IsWebshell,” “IsMeterpreter”) expressed using the “n-way by condition” module. Then, we classify the accuracy results for each algorithm, which are obtained via our experiments, by applying the modules of “IsAdduser,” “IsHydraFTP,” “IsHydraSSH,” “IsJavaMeterpreter,” “IsWebshell,” “IsMeterpreter,” and “IsNormal.” In the case of the six types of attack, if the result of the “decide” module is true, the “count” module corresponding to the respective attack increases the count by one. However, if the result is false, the “CountMissAttack” module adds one to the number of misclassified attacks. Furthermore, in the case of normal data, the “CountNormal” and “CountMissNormal” modules store the counts for true and false results similar to the previous case. The classification results obtained through simulation are shown in Fig. 7. In Fig. 7, the “CountMissNormal” and “CountMissAttack” values are important. First, the simulation result of the SVM algorithm shows that the number of misclassifications as indicated by “MissNormal”, which is the count for the number of instances when normal data is misclassified, was zero. Therefore, it can be seen that normal data are classified considerably well in simulations compared with the experimental results of SVM_RBF that had an accuracy of only 82%. Furthermore, in the case of the KNN model simulation, the number of “MissNormal” instances is larger than that for the SVM simulation; however, the number of misclassifying the attack data is less than that of “MissAttack”. Finally, from the logistic regression simulation results, it is clear that the model based on logistic regression performs worse than the other two machine learning algorithm models considered in this study. The most important feature in an IDS is that the false negative count “MissAttack” should be the least. Consequently, from the results shown in Fig. 7, we can confirm that the KNN-based model provides good results for detecting and classifying attack data, while the SVM-based model performs well in detecting normal data.

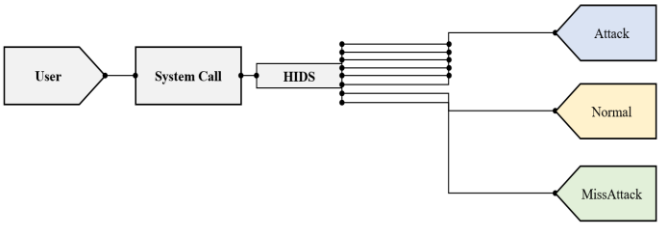


Fig. 4. System Call Process.

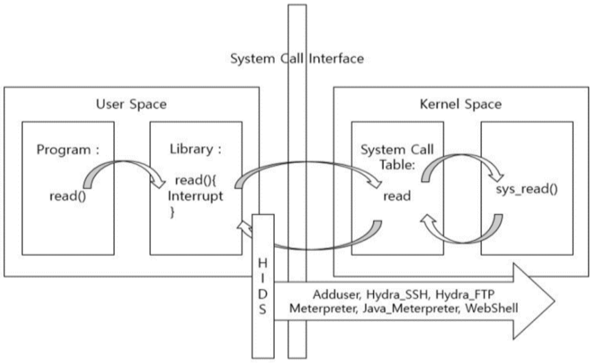


Fig. 5. HIDS Model based on Simulation.

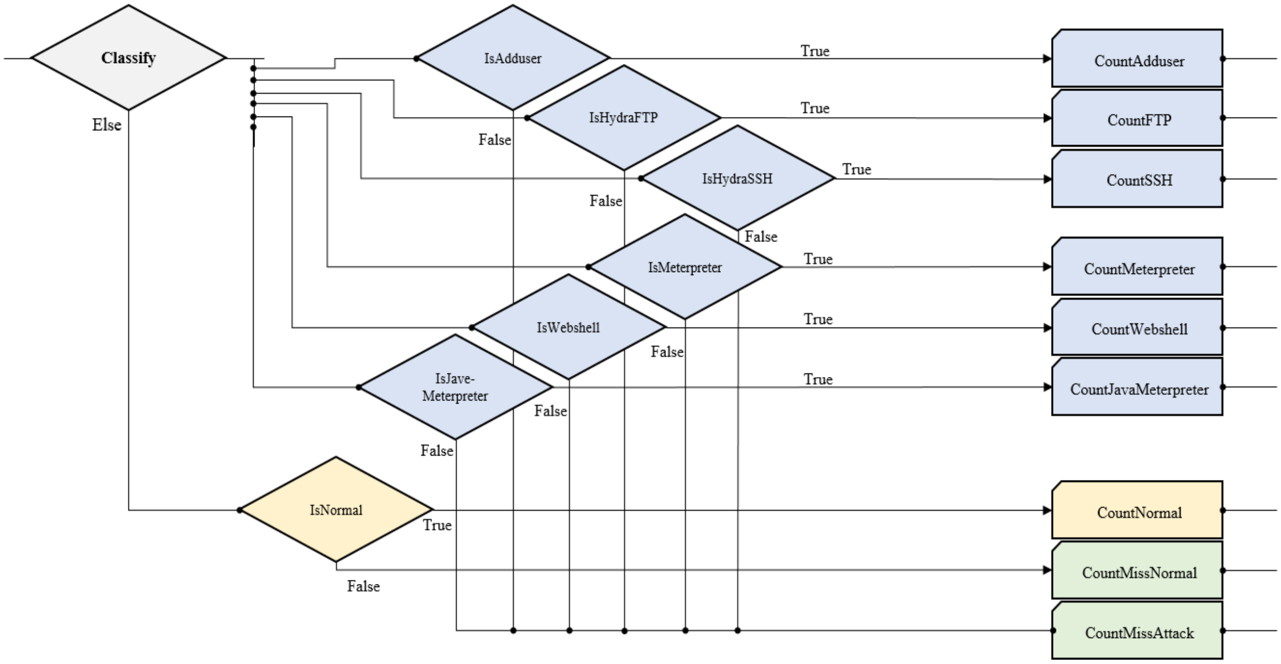
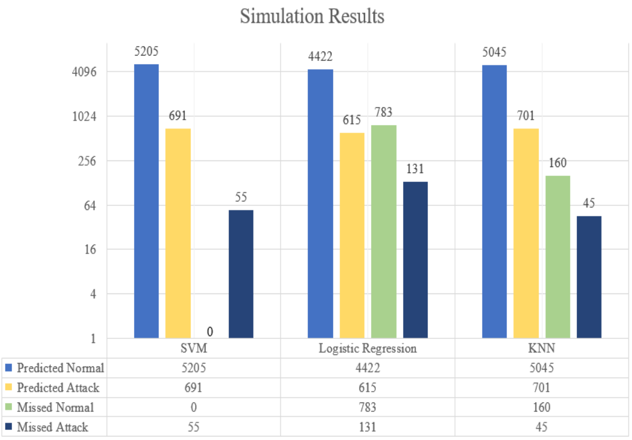


Fig. 6. Schematic Diagram of the HIDS Sub-Model based on Simulation.



	SVM	Logistic Regression	KNN
Predicted Normal	5205	4422	5045
Predicted Attack	691	615	701
Missed Normal	0	783	160
Missed Attack	55	131	45

Fig. 7. Simulation Results.

VI. Conclusions

In this study, we propose a method to increase intrusion detection accuracy by applying and comparing various machine learning algorithms that are suitable for intrusion detection models in order to overcome the disadvantages of an anomaly-based intrusion detection method. Using the ADFA-LD, which consists of various system call traces for attacks on the latest operating systems, we preprocessed the data using the N-gram technique and proposed a methodology to overcome the limitations of the STIDE algorithm.

For verification of our proposed methods, we simulated models using the Arena simulation tool to detect and classify anomalies in HIDS based on the machine learning algorithms considered in our study and verified the accuracy of these models. Based on our simulation results, we confirmed that changes in methodology, compared to previous studies, have made progress in improving the accuracy of anomaly detection in HIDS.

In conclusion, the methodology proposed in this study enables the detection of normal data and attack data as well as the classification of each attack data by extracting the patterns and features of anomalies using machine learning algorithms and applying them to anomaly detection in the HIDS, thereby significantly improving the HIDS, and thus, accurate detection rate.

In future work, we will consider to increase the accurate detection rate of anomaly-based intrusion detection systems using a variety of machine learning and deep learning models with a variety of dataset such as the NGIDS-DS dataset as well as ADFA-LD system call sequence dataset. In addition, we will conduct research on the adjustment of parameters and the development of improved machine learning algorithms to overcome the disadvantages of each machine learning algorithm.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT: Ministry of Science and ICT) (No. NRF-2019R1F1A1059036).

References

[1] A. D. Keromytis, V. Misra, and D. Rubenstein, “SOS: Secure Overlay Services," Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM '02), pp. 61-72, Pittsburgh, PA, USA. Aug. 19-23, 2002.

[2] D. Wagner and P. Soto, "Mimicry Attacks on Host-Based Intrusion Detection Systems," Proceedings of the 9th ACM Conference on Computer and Communications Security (CCS '02), pp. 255-264,

Washington DC, 18-22 Nov. 2002, http://dx.doi.org/10.1145/586110.586145

[3] H. Cavusoglu, B. Mishra, and S. Raghunathan, "A Model for Evaluating IT Security Investments," Communications of the ACM, vol. 47, no. 7, pp. 87-92, July 2004.

[4] K. Richards, "Network based Intrusion Detection: A Review of Technologies," Computers & Security, vol. 18, no. 8, pp. 671-682, 1999.

[5] C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, and M. Rajarajan, "A Survey of Intrusion Detection Techniques in Cloud," Journal of Network and Computer Applications, vol. 36, no. 1, pp. 42-57. Jan. 2013.

[6] O. Depren, M. Topallar, E. Anarim, and M. K. Ciliz, "An Intelligent Intrusion Detection System (IDS) for Anomaly and Misuse Detection in Computer Networks," Expert Systems with Applications, vol. 29, no. 4, pp. 713-722, Nov. 2005.

[7] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez, and E. Vazquez, "Anomaly-based Network Intrusion Detection: Techniques, Systems and Challenges," Computers & Security, vol. 28, no. 1-2, pp. 18-28, 2009.

[8] G. Creech and J. Hu, "A Semantic Approach to Host-based Intrusion Detection Systems using Contiguous and Discontiguous System Call Patterns," IEEE Transactions on Computers, vol. 63, no. 4, pp. 807-819, Apr. 2014.

[9] A. Torkaman, G. Javadzadeh, and M. Bahrololum, "A Hybrid Intelligent HIDS Model using Two-layer Genetic Algorithm and Neural Network," 5th Conference on Information and Knowledge Technology (IKT), pp. 92-96, 28-30 May 2013.

[10] S. Forrest, S. A. Hofmeyr, A. SoMayaji, and T. A. Longstaff, "A Sense of Self for Unix Processes," Proceedings of IEEE Symposium on Security and Privacy, pp. 120-128, May 1996.

[11] G. Creech and J. Hu, "Generation of a New IDS Test Dataset: Time to Retire the KDD Collection," Wireless Communications and Networking Conference (WCNC 2013), Shanghai, 7-10 April 2013 http://dx.doi.org/10.1109/WCNC.2013.6555301

[12] X. Zhang, Y, Wang, M. Gou, M. Sznaier, and O. Camps, "Efficient Temporal Sequence Comparison and Classification using Gram Matrix Embeddings on a Riemannian Manifold," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4498-4507, June 2016. https://doi.org/10.1109/CVPR.2016.487

[13] Arena Simulation Software, http://www.arenasimulation.com/

[14] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases /kddcup99 /kddcup99.html, Feb. 2019.

[15] M. Xie and J. Hu, "Evaluating Host-Based Anomaly Detection Systems: A Preliminary Analysis of ADFA-LD," 6th IEEE International Congress on Image and Signal Processing (CISP '03), pp. 1711-1716, Dec. 2013.

[16] M Liu, Z Xue, X Xu, C Zhong, J Chen, "Host-based Intrusion Detection System with System Calls: Review and Future Trends," ACM Computing Surveys (CSUR), vol. 51, no. 5, pp. 1-36, Nov. 2018. https://doi.org/10.1145/3214304

[17] A. Ahmadian Ramaki, A. Rasoolzadegan and A. Javan Jafari, "A Systematic Review on Intrusion Detection based on the Hidden Markov Model," Statistical Analysis and Data Mining - Wiley Online Library: The ASA Data Science Journal, vol. 11, no. 3, pp. 111-134, Apr. 2018. https://doi.org/10.1002/sam.11377

[18] G. Creech, "Developing a High-accuracy Cross Platform Host-Based Intrusion Detection System Capable of Reliably Detecting Zero-day Attacks," Ph.D. Dissertation, University of New South Wales, Canberra, Australia, 2014.

[19] W. Haider, J. Hu, Y. Xie, X. Yu, and Q. Wu, "Detecting Anomalous Behavior in Cloud Servers by Nested Arc Hidden SEMI-Markov Model with State Summarization," IEEE Transactions on Big Data, vol. 5, no. 3, pp. 305-316, sept. 2019. https://doi.org/10.1109/TBDATA.2017.2736555

[20] S. Fine, Y. Singer, and N. Tishby, "The Hierarchical Hidden Markov Model: Analysis and Applications," Machine learning, vol. 32, no. 1, pp. 41–62, 1998. https://doi.org/10.1023/A:1007469218079

[21] W. Haider, J. Hu, J. Slay, B. Turnbull, and Y. Xie, "Generating Realistic Intrusion Detection System Dataset based on Fuzzy Qualitative Modeling," Journal of Network and Computer Applications, vol. 87, pp. 185–192, June 2017. https://doi.org/10.1016/j.jnca.2017.03.018

[22] E. Aghaei, "Machine Learning for Host-based Misuse and Anomaly Detection in UNIX Environment," M.S. Thesis, Computer Science in University of Toledo, May 2017.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, June 2002.

[24] W. Haider, J. Hu, and M. Xie, "Towards Reliable Data Feature Retrieval and Decision Engine in Host-based Anomaly Detection Systems," IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), pp. 513-517, June 2015. https://doi.org/10.1109/ICIEA.2015.7334166

[25] H. Berger, D. Merkl, and M. Dittenbach, "Exploiting Partial Decision Trees for Feature Subset Selection in e-Mail Categorization," Proceedings of the ACM Symposium on Applied Computing, Dijon, France, 2006.

[26] B. Borisaniya and D. Patel, "Evaluation of Modified Vector Space Representation Using ADFA-LD and ADFA-WD Datasets," Journal of Information Security, vol. 6, no. 3, pp. 250-264, 2015.

[27] Y. B. Bhavsar and K. C. Waghmare, "Intrusion Detection System Using Data Mining Technique: Support Vector Machine," International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 3, pp. 581-586, March 2013.

[28] W. S. Noble, "What is a Support Vector Machine?," Nature Biotechnology, vol. 24, no. 12, pp. 1565-1567, Dec. 2006.

[29] C. J. Peng, K. L. Lee, and G. M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting," Journal of Educational Research, vol. 96, no. 1, pp. 3-14, 2002. https://doi.org/10.1080/00220670209598786

[30] P. Laskov, P. Düssel, C. Schäfer, and K. Rieck, "Learning Intrusion Detection: Supervised or Unsupervised?," International Conference on Image Analysis and Processing, Lecture Notes in Computer Science, vol. 3617, Springer, Berlin, Heidelberg, 2005.

[31] A. Vieira, L. Dias, G. Pereira, and J. Oliveira, "Comparison of SIMIO and ARENA Simulation Tools," 12th Annual Industrial Simulation Conference (ISC2014), University of Skövde, Skövde, Sweden, pp. 5-13, June 2014.

A Smart Home System based on Internet of Things

Rihab Fahd Al-Mutawa[1], Fathy Albouraey Eassa[2]
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—The Internet of Things (IoT) describes a network infrastructure of identifiable things that share data through the Internet. A smart home is one of the applications for the Internet of Things. In a smart home, household appliances could be monitored and controlled remotely. This raises a demand for reliable security solutions for IoT systems. Authorization and authentication are challenging IoT security operations that need to be considered. For instance, unauthorized access, such as cyber-attacks, to a smart home system could cause danger by controlling sensors and actuators, opening the doors for a thief. This paper applies an extra layer of security of multi-factor authentication to act as a prevention method for mitigating unauthorized access. One of those factors is face recognition, as it has recently become popular due to its non-invasive biometric techniques, which is easy to use with cameras attached to most trending computers and smartphones. In this paper, the gaps in existing IoT smart home systems have been analyzed, and we have suggested improvements for overcoming them by including necessary system modules and enhancing user registration and log-in authentication. We propose software architecture for implementing such a system. To the best of our knowledge, the existing IoT smart home management research does not support face recognition and liveness detection within the authentication operation of their suggested software architectures.**

Keywords*—*Internet of Things (IoT); smart home; system; architecture; security; management

I. Introduction

The Internet of Things (IoT) is a system of sensors and actuators embedded in physical objects equipped with unique identifiers and the ability to transfer data over both wired and wireless networks [1]. The substantial development activity in IoT includes many categories, such as smart grid, smart logistics, environment and safety testing, intelligent transportation, industrial control and automation, finance and service, military defense, health care, fine agriculture, and smart homes [2]. Smart homes are homes that incorporate a communication network that connects the key sensors and actuators, and allows them to be accessed, monitored or controlled remotely [3]. In a smart home, there are certain characteristics; the network size is small, the number of users is very few (as it is restricted to the family members), and different network connectivity can be used, such as 3G, 4G, and Wi-Fi. The data management occurs through a local server; IoT Devices are using RFID or WSN wireless technologies, and the bandwidth requirement is small [2]. The smart home is also known as house automation, in which domestic activities are made more comfortable, convenient, secure and economical. As a result, home automation became popular due to its numerous benefits.

A home automation system consists of four main components. The first is the user interface, such as a computer or phone used to give orders to the control system. The second component is the transmission mode, which is the Ethernet (wired), or Bluetooth (wireless). The third is the central controller, which is the hardware interface that communicates with the user interface by controlling electronic devices. The final component is various electronic devices, such as an air-conditioner, a lamp, or a heater that are compatible with the mode of transmission, and connected to the central controlling system [4].

There are many challenges present in IoT systems, such as management, performance, privacy, and security. The security challenges include authorization, authentication, and access control [5]. Therefore, registration and log-in are important security operations in a smart home system, as unauthorized access to the system (such as a cyber-attack) could cause danger by opening the door for a thief, threatening the safety of residents and their belongings. In this paper, we suggest a user-friendly multi-factor authentication for the proposed smart home system. One of those factors is the password, and the other factor is face recognition. The integration scenario for face recognition and liveness detection with the log-in operation to smart home is a novelty for this paper. Multi-factor authentication is a secure authentication process combined of more than one authentication technique chosen from various independent categories of credentials to provide better way of validating legitimate users [6]. Multi-factor authentication creates a layered hindrance, thus making it more difficult for an unauthorized individual to reach the system. In this case, even if attackers break one factor, they still have one more impediment to break before they can access the system. The two-factor authentication solution is cost effective to customers, and has the means of providing flexible and strong authentication. It reduces the fraud rate when compared to one-factor authentication [7]. The use of multi-factor authentication is growing in order to help verify the identities of users requesting to access the system for information that could be sensitive or could control the system. The four most common types of authentication factors are the cognitive information (such as passwords), items that a user possesses (such as smart cards), a biometric trait of the user (such as face recognition and fingerprints), and a user's location information (such as IP address and GPS) [6].

This research presents a novel contribution in comparison to the previous research by suggesting a log-in module for managing the operations of user registration and log-in more securely. This module is integrated within the suggested software architecture of the IoT smart home, and then it is explained in detail. In this research, the added features to the smart home system are compared to the smart home systems of related work in the discussion section of this paper. The integration of face recognition and liveness modules within the log-in module is first presented by this research.

The paper consists of six sections and is organized as follows: Section 1 introduces the research problem. Section 2 presents background information. Section 3 reviews the related work. Section 4 proposes the system architecture. Section 5 discusses the suggested solution. Finally, Section 6 concludes the work and presents prospects for future research.

II. Background

This section presents a literature review of the concepts behind the used modules by the proposed system architecture.

A. Hashing

Hashing is recommended for securing passwords [8]. It is used during the registration and log-in operations [9]. A hashing password is better than encrypted passwords, as hashing is a one-way function; we cannot regenerate the plain text value of the password from its hash value. Thus, to secure passwords, SHA256 and SHA512 are recommended cryptographic hash functions [8]. The secure hash algorithms consist of cryptographic hash functions published by the National Institute of Standards and Technology as a United States Federal Information Processing Standard, including:

1) SHA-1: A 160-bit hash function that resembles the earlier MD5 algorithm. It was designed by the National Security Agency to become part of the Digital Signature Algorithm. Since 2010, the standard is no longer approved for most cryptographic uses due to cryptographic weaknesses that were discovered in SHA-1.

2) SHA-2: A family of two hash functions that are similar to each other, with different block sizes, named as SHA-256 and SHA-512. Their difference is in the word size; SHA-256 uses 32-bit words and SHA-512 uses 64-bit words. The National Security Agency also designed truncated versions of each standard, named as SHA-224, SHA-384.

3) SHA-3: This hashing function was formerly called Keccak, chosen in 2012 after a public competition between non-National Security Agency designers. Its internal structure is different from the rest of the SHA family, and it supports the same hash lengths as SHA-2.

4) SHA-4: Hash functions that have different block sizes are known as SHA-512 [10].

B. Liveness Detection

Liveness detection is an active research area. Face recognition systems can be spoofed by photographs, video recordings, and dummy faces made of materials like silica gel or rubber. Face recognition algorithms do not have a mechanism for differentiating a live face from a fake face; therefore, liveness detection must be integrated with the system in order to verify whether the facial image is alive or reproduced synthetically [11]. A spoofing attack is a direct attack occurring outside the system or at the sensor level, while indirect attacks occur inside the system if an intruder manipulates the templates in the database or evades the feature matcher or extractor [12]. Liveness detection differentiates between a live feature set and non-live feature set [11]. Liveness detection techniques are classified into four main categories, namely motion-analysis-based, texture-analysis-based, image-quality-analysis-based, and hybrid [12].

C. Face Recognition

Face recognition is a popular biometric authentication method [13]. Biometric characteristics provide accurate evidence of personal identity; thus, biometric authentication provides an advantage when compared to other non-biometric identifiers [14]. Moreover, biometric authentication does not need to be memorized, thus it is preferred over other traditional techniques [12]. The main steps of a face recognition system are face detection, face alignment, feature extraction, and feature matching [15]. The goal of the facial detection phase is to determine whether there are any faces in the image. If it discovers a face it returns the image location and the extent for each face. Therefore, face detection is more challenging than face localization where the assumption that the image contains only one face [16]. The face alignment is the first phase to transform the detected faces into a standard pose. The use of face alignment methods can significantly improve face recognition accuracy [17]. Feature extraction identifies facial feature components, such as eyebrows, eyes, nose, and mouth [18].

D. Voice Recognition

Voice recognition software in general encompasses four technologies: spoken recognition of human speech (which is also known as speech recognition or speech-to-text), synthesis of human readable characters into speech (which is also known as speech synthesis or text-to-speech), speaker identification and verification, and natural language understanding [19].

1) Speech recognition: Speech recognition (also referred to as speech-to-text) is the process that transforms the computer's acoustic signal (i.e., speech) into set of typed words [20]. The first type of speech recognition is command and control (or spoken command recognition), and the second is dictation. Command and control recognizes single words or short phrases spoken continuously, such as "Count the Lines" or "Accept and Save." Dictation technology has two divisions: discrete and continuous. The discrete dictation requires lower processing power requirements, as the end-user needs to place a short pause between each spoken word. Continuous dictation, has overcame this limitation, and it is used in the radiology implementation [19].

2) Speech synthesis: Synthesizers, also referred to as text-to-speech, are computer systems that have the ability to read any text aloud when an operator introduces the text in the computer [21]. In a speech synthesis system, the computer produces the same phonemes that humans would make when they read text aloud [19].

3) Speaker identification and verification: Speaker identification and verification are two related processes. They deal with the identity of the human speaker, unlike in speech recognition and speech synthesis where these processes deal with what was spoken by a human or with synthesizing a particular human voice. However, the speaker verification process is applied in order to authenticate a given human speaker against a database pool of enrolled users [19].

4) Natural language understanding: Understanding natural language refers to inputting spoken or typed sentences into a computer and then processing them to extract their meanings; thus, the computer can understand human language [22].

E. Chatbot

A chatbot engine is a natural language engine, as it has the responsibility to translate natural language into instruction understandable by machines. There is a complexity to chatbot engines, as they use various natural language processing models and machine learning techniques to provide an acceptable level of accuracy [23]. There are two key functional components for the chatbot engine - that is intent and entities [24].

1) Intent: The user’s utterance is first analyzed for intent. Intents refers to what the user is looking to accomplish, such as getting sensor data or turning devices on or off. To understand the user’s intent, it maps the natural language phrases to canonical phrases in order to conclude the specific action that should be taken by the smart home system [23].

2) Entity: This refers to the information specific to certain domain that is extracted from the textual utterance of the user. They are used for identifying the required parameters in order to take a specific action. To train the chatbot engine, entities are typically grouped together according the expectation that they will give the same actions. For example, for the utterance “thermostat”, the chatbot can recognize the words "thermostat", “A/C”, “air conditioner”, or "air conditioning" and convert it into the keyword ‘$thermostat’ [23]. Using keywords from the trained chatbot raises the efficiency of the system, so that the smart home system can receive the appropriate request from this component and then direct the request to the target device through the ThingsManager. After that, the system responds to the end user. The user requests are training phrases that are imported into the chatbot engine along with the trained responses. However, the system should also be trained for temperature commands, humidity questions/commands, light state questions/commands, the rest of sensors questions/commands, microcontroller questions, general questions, and navigation commands [25].

F. Message Queuing Telemetry Transport Broker (MQTT)

The most popular communication protocol for IoT is Message Queuing Telemetry Transport (MQTT) [26]. MQTT is an ISO standard (ISO/IEC PRF 20922) lightweight protocol that depends on the principle of publishing messages and subscribing to topics (pub/sub). It is useful for sensor devices with limited resources. The client can connect to a broker and subscribe to its selected topics. Clients also publish messages to topics after connecting to the broker [27]. The Internet of Things system integrates information from heterogeneous sensors, allowing these devices to deliver different sensed information through networks. Therefore, the IoT broker acts as an information exchange center in the system, relaying periodic messages from heterogeneous appliances to IoT clients [28]. The MQTT broker is an open source code at the heart of all MQTT arrangements. It provides a connecting link between physical devices and smart home systems. The features and limitations for the five most commonly used brokers can be found in [26]. The MQTT broker can provide services, such as monitoring and controlling room temperatures, suppressing fires, and controlling alarms, all of which require microcontrollers to be used as the IoT end devices that connect sensors and actuators to the smart home system through a Wi-Fi channel [29]. If the user requests data from sensors or actuators, then the smart home system sends a message to the appropriate microcontroller through the MQTT broker to perform the user request [25]. The MQTT and broker acts as a simple, common interface to which everything can connect. Topics are arranged in a hierarchy, using a slash (/) separator. The client receives messages by creating subscriptions. A subscription can be for an explicit topic, in which case-only messages to that topic will be received, or for more than one topic by including wildcards that is either + or # [27].

III. Related Work

The IoT has many characteristics [30]. They require new management (including security management methods) or an entirely new approach to the prominent management systems, as there is a need for managing the growing number of things connected to the Internet, which generates a large amount of network traffic for devices with low power capabilities. To address this concern, the authors introduced an Internet of Things management system for operations, such as sensing and actuating mobile software agents. Their proposal is amongst the first studies that address managing these things as part of the IoT. The system supports fundamental management functions, including operation, such as requesting sensor data or actuating, monitoring, and communicating, whether local or remote. The proposed model architecture is a simple two-tier model that can vary to a more complex model based on a hierarchal and distributed structure consisting of many managers and agents in order to provide the needed functionalities that are a part of traditional network management systems. The proposed system is then expanded in [31], their work demonstrating some of the monitoring and control capabilities for management provided in the proposed system. They conducted an experiment to offer an example of how the IoT management system can be used to support management over services, such as remotely controlling and monitoring things remotely over the Internet using a mobile application. The results collected from the experiment validated the management capabilities of the proposed Internet of Things management system, and demonstrated its successful deployment. Moreover, the smart home system has more added features in [32], as in addition to offering a management solution for things that suffer from limited computation and power resources, a middleware solution is proposed for the system to enable the management of things based on seven components. One of those components is the security module, which enforces the software agent to be registered by the manager, and then grants the permission control for the agent whether read or read/write. After that, the agent must be approved by the system administrator, as the remote agent is a third-party application that can manage things if granted permission. The manager keeps a database of authorized agents registered by their Agent IDs in order to ensure that only authorized agents are granted permission to connect to the manager. Agents have an access type to the thing that is either read or read-write. In [23], the smart home system had an additional chatbot feature. The research integrated a chatbot, which is an intelligent conversational software agent in the IoT scenario with text-based inputs to the smart home system. They presented a novel paradigm, combining the chatbot concept with the IoT concept in a single solution. The suggested software architecture requires using a chatbot channel, such as Facebook Messenger, to interact with the bot. The integration between the smart home system and the third-party software is over the application layer through an HTTP RESTful API (which is a part of the IoT Cloud smart home system) to allow the user to interact with the smart home using the chatbot. However, in [25], the integrated chatbot within the smart home system has the added feature of delivering the command to the system through the user's voice, and accordingly hearing back the response through speakers. The chatbot can understand text or voice commands using natural language processing, as with the use of natural language processing, home devices become more user-friendly for end-users. The solution is integrating third-party APIs and open source technologies into one mash-up, using multi-tier architecture for the rapid development of the IoT smart home system. The ready to use services are the Dialogflow API for the efficient integration of the chatbot to the smart home system, the Web Speech API for the voice recognition and synthesis features, MQTT for controlling sensors and actuators, and Firebase as a database for dynamic data storage. This integration that is based on third party APIs and open source technologies is first presented by their work.

This paper expands the current IoT smart home systems by incorporating face recognition and liveness detection with the enhanced operations of user registration and log-in authentication. These operations are essential part of the system, and we propose software architecture for a smart home system.

IV. System Architecture

The proposed system architecture for the smart home system consists of six modules, as shown in Fig. 1, and based on web services so the user can control his or her home from outside the house. A web service is a collection of open protocols and standards used for data exchange between systems or applications. They are XML-based information exchange systems. It supports interoperability, as web services use open standards, thus software systems written in different programming languages and running on different platforms can use web services for exchanging data over the Internet in a similar manner, as if the communication is for inter-process on a single computer [33]. The description for each web service is provided below, and an explanation for the concept behind the internal components is presented in the background in Section 2. The proposed system can be installed on computers, tablets, and mobile phones.

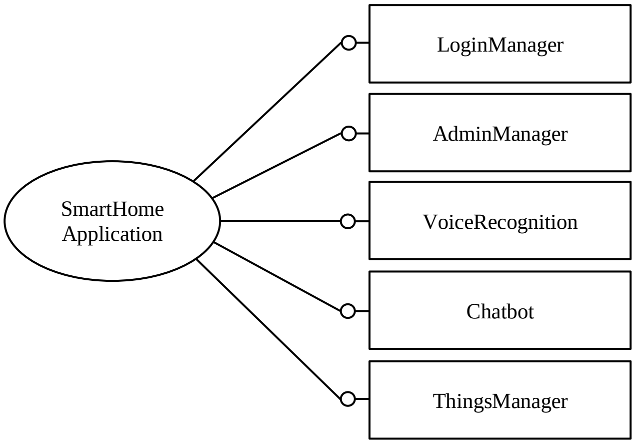


Fig. 1. System Architecture based on Web Service Technology.

A. Login Manager

This module is the main significant contribution of this paper. The log-in manager is responsible for managing the user authorization and authentication operations, which handles user registration and log-in operations. It consists of four software components: hashing, liveness detection, face recognition, and notification, as shown in Fig. 2. The main functions of this module are register and log-in. It applies the multi-factor authentication method as an extra layer of security for logging into the system. The first factor is the password and the second factor is face recognition as according to [34], face recognition recently has become popular, as it is a non-invasive biometric technique. Also, it is easy to use with the available cameras embedded in most computers and smartphones [35].

1) Hashing: To secure the user password during the registration and log-in operations, this component hashes the password using a cryptographic hashing function.

2) Liveness detection: This component is needed to check face liveness, as the face recognition component can be spoofed easily if it is used without this module.

3) Face recognition: The face recognition component is used for extracting the facial features during log-in and registration, in addition to verifying the user's face while log-in to the system.

4) Notification: Sending notification messages with an instant picture for the current user to system admin during suspicious registration and log-in attempts are increasing the security level in the smart home system. During a user registration operation, the system sends two notification messages to the system admin. The first message is an approval request for the newly registered user sent by the end of the user registration operation. The second message is sent when there is an attempted spoofing attack after checking the face liveness when the user tries to register with a picture, recorded video, or a dummy face. Approving registration requests is another security method utilised to delay any suspicious attempts for controlling the home. Fig. 3 represents the sequence diagrams for user registration operations, including sent notifications. Fig. 4 represents the algorithm of the user registration operation.

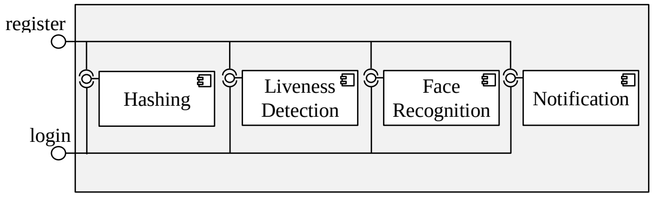


Fig. 2. Internal Components of Log-in Web Service.

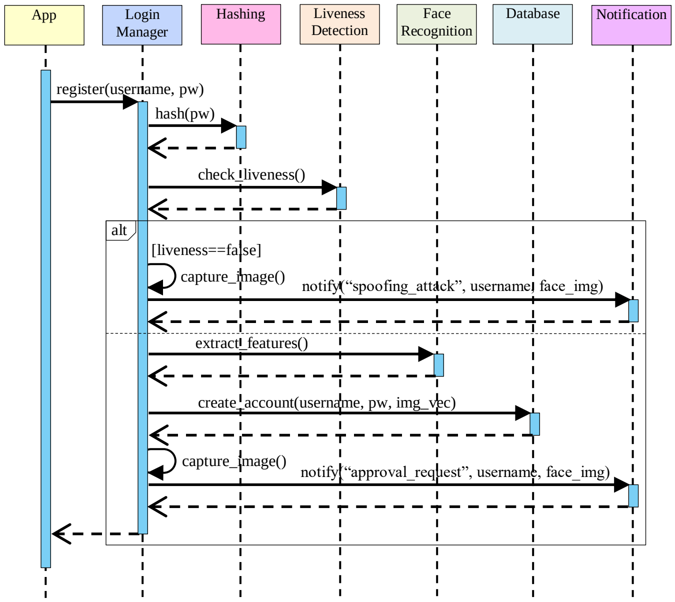


Fig. 3. User Registration Sequence Diagram.

Name: UserRegistration

Input: Username, password, and live streaming video for the face.
Output: A confirmation for registration and a notification message sent to the system admin.

```
if (username=="") then
   return 'empty username'
else if (check_exist(username)) then
   return 'username already exist'
else if (password=="") then
   return 'empty password'
else if (password!="") then
   hash(password)
   liveness=check_liveness()
   if (liveness==false) then
      /*take an image for the current user*/
      face_img=capture_img();
      /*send the captured image in a spoofing attack notification
      message to the system admin*/
      notify("spoofing_attack",face_img)
   else
      /*extract the facial features*/
      img_vector=extract_features()
      /*create an account with the username, password, and image
      vector data*/
      create_account(username,pw,img_vector)
      ask the current user for a picture
      /*take an image for the user*/
      face_img=capture_img();
      /*send the captured image in an approval and activation
      request notification message to the system admin*/
      notify("approval_request",username,face_img)
      return a registration confirmation
return false
```

Fig. 4. User Registration Algorithm.

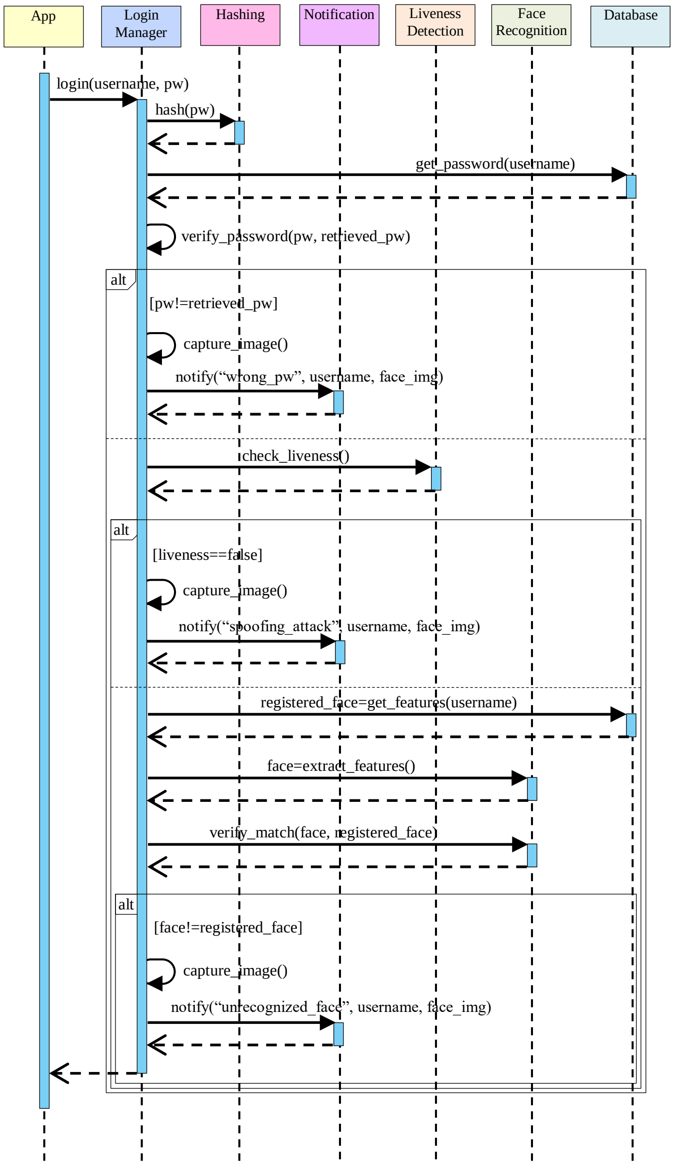


Fig. 5. User Log-in Sequence Diagram.

During the user log-in operation, the system sends three notification messages to the system admin. The first message is sent when the user tries to access the system by entering an incorrect password. The second message is sent when there is a spoofing attack attempt after checking the face liveness. The third message is sent when the system detects an unrecognized face, as it is unregistered in the system database. Fig. 5 represents the sequence diagrams for user log-in operation, including the sent notifications. Fig. 6 represents the algorithm of user log-in operations.

B. Admin Manager

This module is another security component used by the system admin for managing the user's access to the system. The specified permissions (i.e. the access type) for controlling the house's appliances can be read, write, or both. The read permission allows for getting the sensor data, while the write permission allows the user to give a request for an actuator. The newly registered user cannot use the smart home system before being granting permissions and activating his or her account. The system admin will need at least three functions provided by this component: one for retrieving pending accounts, another for retrieving user approval, and one for setting user approval.

C. Voice Recognition

The voice recognition feature is optional, as the user might prefer to deal with the system by text. In this web service, there are two included components, as shown in Fig. 7.

1) Speech recognition: This component is provides a function responsible for converting the user's speech into typed text.

2) Speech synthesis: This component provides a function responsible for reading the textual system respond aloud.

```
Name: UserLogin

Input: Username, password, and live streaming video for the face.
Output: Log-in confirmation

if (username==””) then
  return 'empty username'
else if (!check_exist(username)) then
  return 'username does not exist'
else if (password==””) then
  return 'empty password'
else if (password!=””) then
  hash(password)
  /*get the registered password for the given username*/
  retrieved_password=get_password(username)
  /*verify the password match*/
  if (password!=retrieved_password) then
    /*take an image for the current user*/
    capture_img()
    /*send the captured image in a wrong password's notification
    message to the system admin*/
    notify(“wrong_password”,username,face_img)
  else
    /*check the face liveness*/
    liveness=check_liveness()
    if (liveness==false) then
      /*take an image for the current user*/
      face_img=capture_img()
      /*send the captured image in a spoofing attack
      notification message to the system admin*/
      notify(“spoofing_attack”,username,face_img)
    else
      /*get the registered face for the username*/
      retrieved_img_vector=get_features(username)
      /*extract the facial features of the current user*/
      img_vector=extract_features()
      /*verify the match between the current face and
      registered face*/
      matchness=verify_match(img_vector,retrieved_img_vector)
      if (matchness==false)
        /*take an image for the current user*/
        face_img=capture_img()
        /*send the captured image in an unrecognized face
        notification message to the system admin*/
        notify(“unrecognized_face”,username,face_img)
        return false
      else
        return true /*log-in confirmation*/
return false
```

Fig. 6. User Log-in Algorithm.

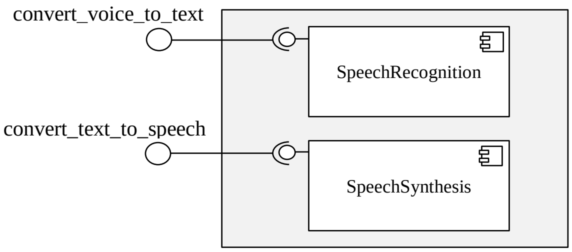


Fig. 7. Internal Components of VoiceRecognition Web Service.

D. Chatbot

There are two components included within the chatbot web service that are intents and entities as shown in Fig. 8.

1) Intent classification: This component can act as a preprocessing step by providing a function that converts the input text into its canonical phrase.

2) Entity recognition: Understanding the user request and getting the chatbot response is accomplished by the end of executing this component. An example [25] of trained user requests and system responses related to temperature are shown in Table I.

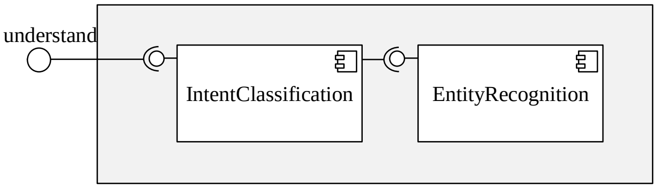


Fig. 8. Internal Components of Chatbot Web Service.

TABLE. I. TEMPERATURE REQUESTS AND RESPONSES

User Requests	System Responses
I want the temperature of lounge.	The temperature of $Room1 is $temperature$%.
What is the temperature of living room?	
Give me the temperature of sitting room.	
What is the temperature of all rooms?	As temperature differs from one room to the other, you need to specify which room.

E. Things Manager

This web service is responsible for passing the instruction to the MQTT broker, which monitors and controls the smart home appliances.

V. DISCUSSION

A summary of the supported smart home system features by this research compared to the smart home systems of related work are presented in Table II.

The proposed system architecture has collected some features from different smart home systems to enhance its functionality, in addition to other added features, namely multi-factor authentication, face recognition, liveness detection, and access notification. However, three features that were considered advantages on other smart systems have not been included in this proposal, so they were not added to Table I. The first feature is the multi-tier architecture suggested by [25] as it is not practical to increase the number of the system servers when all the web services are developed in-house, and as one of the characteristics of the smart home system, as mentioned in the introduction section, is that data management occurs through a local server, thus the use of third-party online services provided through an API was avoided in the suggested architecture. The second feature is the hierarchal mobile software agent distributed system supported by [5], [31], and [32], because creating an agent for each sensor or actuator device is not practical economically, takes up considerable space in the home, and presents a low level of security, as a mobile agent third party application can be untrusted and it may harm the system. Moreover, there is a possibility for increased network traffic when there are a large number of mobile software agents concurrently interacting with the system. Therefore, we used web service technology for the suggested smart home systems. Web service technology has many advantages for this application. For instance, it supports remote procedure calls so that the system can be controlled remotely over the Internet, which also solves the interoperability issues, and it is scalable and easily maintainable. The third unsupported feature present in this research is in [23], which is hosting the smart home system on a cloud. The cloud is not suitable for any sensitive data due to its many security vulnerabilities, such as loss of data, data breaches, account information theft, replacement, or modifications (such as editing the face template by an employee working in the cloud company or by a hacker) and malware. Additional security precautions taken by this system include multi-factor authentication, checking for face liveness, and alerting the system admin with a picture for the current user when there is any suspicious registration or log-in activities. Therefore, this work is expected to increase the security level compared to the current smart home systems. Moreover, third party APIs working on the internet were avoided by the suggested software architecture, as they may threaten the information security and privacy, and this threat increases if the source is untrusted. Additionally, creating in-house web services for the main services in the system facilitates the future functional enhancement process through the module's replacement or extendibility. However, two explained concepts: speaker identification and verification, and natural language understanding were included in the background section but they were excluded from the components of the proposed solution because the speaker identification and verification require using the registered user's voice, but using the voice is an optional feature in the smart home system as it might be more convenient for the user to use text. On the other hand, the natural language understanding was not included, as it is part of the chatbot solution, and there is no need for redundancy.

It is worth mentioning that if security is not the main concern in some smart home systems as when the system admin do not connect sensitive sensors and actuators, such as smart door lock or camera, to the system, and the multi-factor authentication can be replaced with face recognition authentication to make accessing the system easier, especially for older people or people with disabilities. It can also be switched to a password entry mode if the system cannot recognize the user; for example, when it is dark or when the user is wearing a mask.

TABLE. II. SMART HOME SYSTEMS COMPARISON

System **Feature**	**This paper**	**[5]**	**[31]**	**[32]**	**[23]**	**[25]**
Multi-factor authentication	✓					
Face recognition	✓					
Liveness detection	✓					
Voice recognition	✓					✓
Chatbot	✓					
Permission control					✓	✓
Access notification	✓			✓		

VI. CONCLUSION

The paper presents a software architecture for managing the smart home in the IoT context. The proposed solution provides the required modules for managing sensors and actuators remotely. Moreover, the most significant contribution of this work on the improvement of smart home security is that it proposes a log-in module for managing the operations of user registration. This log-in is based on multi-authentication factors, and is supported by a suggested notification module that alerts the system admin for any suspicious registration or log-in attempts to the smart home system. This log-in module is integrated into the proposed smart home software architecture with a detailed explanation for its functionality. The face recognition and liveness modules are also integrated into the smart home system through the log-in module as a preceding suggested method in authorizing and authenticating the users into the smart home system.

For future work, we suggest the following: extending the system functionality by adding a scheduled web service for appliances to the system architecture, implementing the proposed solution as the integrations in this work are theoretical, highlighting the limitations of chatbots that support the Arabic language, improving the current implementations of Arabic chatbots, and integrating an Arabic chatbot to the IoT smart home system. The Arabic language is the first language of more than 340 million speakers [36], yet there are no presented solutions in literature for a smart home that is controlled using the Arabic language.

REFERENCES

[1] Sultan, M., & Ahmed, K. N. SLASH: Self-learning and adaptive smart home framework by integrating IoT with big data analytics. In 2017 Computing Conference (pp. 530-538). IEEE. July, 2017.

[2] Patgiri, R., & Nayak, S. Data of Things: The Best Things Since Sliced Bread. In 2018 International Conference on Communication and Signal Processing (ICCSP) (pp. 0341-0348). IEEE. April, 2018.

[3] Dzogovic, B., Santos, B., Noll, J., Feng, B., & van Do, T. Enabling smart home with 5G network slicing. In 2019 IEEE 4th International

Conference on Computer and Communication Systems (ICCCS) (pp. 543-548). IEEE. February, 2019.

[4] Gunge, V. S., & Yalagi, P. S. Smart home automation: a literature review. International Journal of Computer Applications, 975, 8887. 2016.

[5] Elkhodr, M., Shahrestani, S., & Cheung, H. Managing the internet of things. In 2015 IEEE International Conference on Data Science and Data Intensive Systems (pp. 579-585). IEEE. December, 2015.

[6] Dasgupta, D., Roy, A., & Nag, A. Multi-factor authentication. In Advances in User Authentication (pp. 185-233). Springer, Cham. 2017.

[7] Vaithyasubramanian, A. C. D. S. S., Christy, A., & Saravanan, D. Two factor authentications for secured log-in in support of effective information preservation and network security. India: ARPN Journal of Engineering and Applied Sciences, 10(5). 2015.

[8] Sriramya, P., & Karthika, R. A. Providing password security by salted password hashing using bcrypt algorithm. ARPN journal of engineering and applied sciences, 10(13), 5551-5556. 2015.

[9] Dhivya, B., Thenmozhi, S., Parameswari, V., Archana, G., Kiruthika, V. Implementation of Security in Log-in Page Using Salt and Pepper Algorithm. 2019.

[10] Simon, b. k., & Nair, a. p. Securing the Transfer of Confidential Data in Fiscal Devices using Blockchain. International Research Journal of Engineering and Technology. 6(5). 2019.

[11] Sain, M., & Kant, C. Liveness Detection for Face Recognition in Biometrics: A Review. IOSR Journal of Computer Engineering, Special Issue-AETM'16, 31-36. 2016.

[12] Arora, G., Tiwari, K., & Gupta, P. Liveness and Threat Aware Selfie Face Recognition. In Selfie Biometrics (pp. 197-210). Springer, Cham. 2019.

[13] Huang, F. L., Liao, Z. Z., Wang, T. H., Chen, Q. M., Wu, T. H., & Chang, C. H. Intelligent and Disaster Prevention Hard Hat Based on AIOT and Speeches Recognition. In 2019 International Conference on Machine Learning and Cybernetics (ICMLC) (pp. 1-5). IEEE. July, 2019.

[14] Zhang, Z., Aziz, E. S., Esche, S., & Chassapis, C. A virtual proctor with biometric authentication for facilitating distance education. In Online Engineering & Internet of Things (pp. 110-124). Springer, Cham. 2018.

[15] Monteiro, C. E., & Trevelin, L. C. Studies of computing techniques for performing face recognition with a focus in the crowds: A distributed architecture based on cloud computing. In 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA) (pp. 1-5). IEEE. November, 2015.

[16] Nanni, L., Lumini, A., Dominio, F., & Zanuttigh, P. Effective and precise face detection based on color and depth data. Applied Computing and Informatics, 10(1-2), 1-13. 2014.

[17] Ni, W., Vu, N. S., & Caplier, A. Unsupervised joint face alignment with gradient correlation coefficient. Pattern Analysis and Applications, 19(2), 447-462. 2016.

[18] Zhao, X., & Zhang, S. A review on facial expression recognition: feature extraction and classification. IETE Technical Review, 33(5), 505-517. 2016.

[19] Mehta, A., Dreyer, K. J., Schweitzer, A., Couris, J., & Rosenthal, D. Voice recognition—an emerging necessity within radiology: experiences of the Massachusetts General Hospital. Journal of Digital Imaging, 11(2), 20-23. 1998.

[20] Sunardi, A., Mahardika, R., Alfarisie, M., Retnoasih, S. S., Sunarko, Suherkiman, H., & Mustofa, K. Study of personnel access system using voice and gait recognition in experimental power reactor (RDE). In AIP Conference Proceedings (Vol. 2180, No. 1, p. 020047). AIP Publishing LLC. December, 2019.

[21] Shanmugam, K., & Vanathi, B. Hardcopy Text Recognition and Vocalization for Visually Impaired and Illiterates in Bilingual Language. In Computational Intelligence and Sustainable Systems (pp. 151-163). Springer, Cham. 2019.

[22] Gope, R., Pasricha, Y., & Ekka, A. From Concept to Reality: Intervention of Information Technology in Healthcare. Humanizing work and work Environment (HWWE 2016): English, 313. 2018.

[23] Kar, R., & Haldar, R. Applying chatbots to the internet of things: Opportunities and architectural elements. arXiv preprint arXiv:1611.03799. 2016.

[24] Keyner, S., Savenkov, V., & Vakulenko, S. Open data chatbot. In European Semantic Web Conference (pp. 111-115). Springer, Cham. June, 2019.

[25] Alexakis, G., Panagiotakis, S., Fragkakis, A., Markakis, E., & Vassilakis, K. Control of Smart Home Operations Using Natural Language Processing, Voice Recognition and IoT Technologies in a Multi-Tier Architecture. Designs, 3(3), 32. 2019.

[26] Soni, D., & Makwana, A. A survey on mqtt: a protocol of internet of things (iot). In International Conference on Telecommunication, Power Analysis and Computing Techniques (ICTPACT-2017). April, 2017.

[27] Tantitharanukul, N., Osathanunkul, K., Hantrakul, K., Pramokchon, P., & Khoenkaw, P. Mqtt-topic naming criteria of open data for smart cities. In 2016 International Computer Science and Engineering Conference (ICSEC) (pp. 1-6). IEEE. December, 2016.

[28] Ray, P. P. A survey on Internet of Things architectures. Journal of King Saud University-Computer and Information Sciences, 30(3), 291-319. 2018.

[29] Kang, D. H., Park, M. S., Kim, H. S., Kim, D. Y., Kim, S. H., Son, H. J., & Lee, S. G. Room temperature control and fire alarm/suppression IoT service using MQTT on AWS. In 2017 International Conference on Platform Technology and Service (PlatCon) (pp. 1-5). IEEE. February, 2017.

[30] Khan, M., Din, S., Jabbar, S., Gohar, M., Ghayvat, H., & Mukhopadhyay, S. C. Context-aware low power intelligent SmartHome based on the Internet of things. Computers & Electrical Engineering, 52, 208-222. 2016.

[31] Elkhodr, M., Shahrestani, S., & Cheung, H. A smart home application based on the Internet of Things management platform. In 2015 IEEE International Conference on Data Science and Data Intensive Systems (pp. 491-496). IEEE. December, 2015.

[32] Elkhodr, M., Shahrestani, S., & Cheung, H. A middleware for the internet of things. arXiv preprint arXiv:1604.04823. 2016.

[33] Kataria, B. XML Enabling Homogeneous and Platform Independent Data Exchange in Agricultural Information Systems. International Journal of Scientific Research in Science. 1(2). 2015.

[34] Maw, H. M., Thu, S. M., & Mon, M. T. Face Recognition based on Illumination Invariant Techniques Model. In 2019 International Conference on Advanced Information Technologies (ICAIT) (pp. 120-125). IEEE. November, 2019.

[35] Hossain, M. S., & Muhammad, G. (2017). An emotion recognition system for mobile applications. IEEE Access, 5, 2281-2287.

[36] Droua-Hamdani, G., Selouani, S. A., Alotaibi, Y. A., & Boudraa, M. Speech Rhythm in L1 and L2 Arabic. Arabian Journal for Science and Engineering, 41(3), 1173-1181. 2016.

Integrated Fuzzy based Decision Support System for the Management of Human Disease

Blessing Ekong[1], Idara Ifiok[2], Ifreke Udoeka[3], James Anamfiok[4]
Dept. of Computer Science, Akwa Ibom State University
Ikot Akpaden, Nigeria

***Abstract*—To eliminate some of the inaccuracies in the diagnosis of human diseases, decision support systems based on algorithms and technologies such as Artificial Neural Network, Fuzzy Logic etc. have been used. The results of such diagnosis are used for treatment and management purposes. Inaccurate and imprecise diagnosis may lead to wrong treatment methods which in turn may result in death or complications. Although treatments are widely carried out using drugs, there exist other treatment methods such as alternative medicine, complimentary medicine which could be used for treatment. We propose an Integrated Fuzzy Based Decision Support System which focuses on the integration of both alternative and pure medicine for the management of malaria. The results obtained showed that integrating these two treatment and management methods will eliminate the limitations of the individual methods therefore bridging the gap between alternative and pure medicine in the treatment and management of human diseases. The system is implemented in C#.**

Keywords—Human diseases; fuzzy based decision support system; human disease; fuzzy logic; C#

I. Introduction

Inaccuracies and imprecision in the conventional method of diagnosis, treatment and management of human diseases (e.g. malaria), recent trends in technology and the demand to stay relevant in today’s competitive world of Information Technology have compelled many in the medical field to adopt the computer- based method of diagnosis, treatment and management of diseases. Such systems operate on the principles of Artificial Intelligence (AI) [1] and are developed for both diagnosis and treatment of diseases. The intelligence of such system is based on the result of an interactive decision process that uses facts and rules to solve real life problems [2], based on knowledge obtained from one or more human expert in specific areas. This could be achieved using fuzzy logic which is a branch of AI.

Fuzzy Logic is a form of multi-valued logic derived from fuzzy set theory to deal with approximate reasoning [3]. It offers a suitable way of representing and processing linguistic information and subjective attributes of the real world, this is presented in the work of [4]. According to the work Fuzzy logic systems adopt a specific lifecycle model which could be divided into four different components or stages [5]; fuzzifier component, fuzzy inference engine, and defuzzifier component. Fuzzy based systems are built to handle complex and sophisticated tasks which may be ambiguous to handle otherwise.

Decision Support System (DSS) refers to a group of components (machine and/or human) working together to provide information for efficient and strategic decision making. Efficiency in this context includes the cost in terms of time and resources involved in implementing the option specified by the decision support system. Conventionally, DSS is seen as a software or collection of computer programs which process raw data to provide information for efficient and strategic decision making. In the medical field, the use of DSS reduces or eliminates inaccuracies in diagnosis and even treatment of diseases.

Diagnosis is not very useful if there is no associated treatment for the diagnosed illness. As such, most diagnosis systems are developed for both diagnosis and treatment. The result of the diagnosis phase determines the treatment recommended. Unfortunately, most of the treatments recommended by such systems are based purely on drugs i.e. pure medicine known as orthodox method. As good as this may be, it is not a one size fit all approach that works always at all times. There exists many non-orthodox treatment approaches such as herbalism which could be used for the treatment of diseases. Herbalism involves the use of herbs for the treatment of diseases. This could be used where drugs are not available or appropriate or complimentarily with other drugs.

In this work, an integrative fuzzy based approach for the diagnosis, treatment and management of diseases is proposed. Diagnosis is carried out using fuzzy logic, based on the outcome derived from the diagnosis phase, treatment or management could be recommended. Treatments are administered using drugs and/or herbs. The medical history of the patient is strictly considered in the treatment of diagnosed illnesses. Medical history contains information such as last treatment period, and details of other health conditions such as pregnancy, ulcer etc. which may interfere with some drugs or herbs. Management in the context of this work involves preventive measures such as anti-malaria drugs, changes in lifestyle such as the use of mosquito nets for the prevention or control of diseases.

II. Fuzzy Logic

Fuzzy Logic (FL) incorporates a simple, rule-based such as IF (condition) AND (condition) THEN (action) method to handle problems instead of the use of mathematical model. It provides a precise approach which could be used to make conclusion using imprecise assumptions. FL is a form of many-valued logic or probabilistic logic; it deals with reasoning that is approximate instead of fixed and exact [6]. It is similar in many ways to Fuzzy sets. A fuzzy set B in Y is expressed as a set of ordered pairs:

$$B = \{(y, \mu B(y)) \mid y \epsilon Y\} \quad (1)$$

$$\mu B: Y \rightarrow N \quad (2)$$

Where: μB is the membership function.

N is the membership space and each element of Y is mapped to N such that if $N = \{0, 1\}$, B is a crisp set. Notwithstanding, if $\{0 \leq N \geq 1\}$, B is a fuzzy set.

III. Management of Malaria

Disease management refers to the concept of reducing health care costs and improving quality of life for individuals with chronic conditions by preventing or minimizing the effects of the disease through integrated care [7].Some of the activities involved in disease management include; prevention, diagnosis, treatment, education.

Prevention: According to the old saying that prevention is better than cure, the cost of preventing malaria is usually lower than the cost of treatment. Although some of the drugs used for the treatment of malaria may also be taken for prevention, Malaria could be prevented in the following ways;

Medication: This involves the use of drugs such as doxycycline, chloroquine, etc. for the prevention of malaria. This method of prevention is also known as chemoprophylaxis. The limitations in this method include the fact that some of these drugs have not been tested for long-term use, are not without side effect and may also be costly. Notwithstanding, anti-malarial drugs can be administered in a particular way to people at high-risk for malaria, such as pregnant women and infants [8][1]. Here, such people are given a dose(s) of anti-malaria drugs for prevention. This is referred to as Intermittent Preventive Therapy (IPT).

Vector Prevention: This implies prevention without the use of medications such as; the use of insecticide treated mosquito net, adequate environmental hygiene, the use of long-clothing and insect repellent in the evenings and at night [8], when mosquitoes are very active. The number of mosquitoes can also be reduced by the use of chemical substance known as insecticide and spraying of repellents indoors.

A. *Orthodox Medicine Diagnosis and Management Approach*

Conventionally, blood tests can show the presence of the parasite and help tailor treatment by determining the present of malaria infection, type of malaria causative parasite, and the organs affected [9][2]. According to the 2017 world malaria report by World Health Organization, antimalarial drugs should only be administered to individuals with suspected malaria case after they have been subjected to parasitological confirmation of diagnosis with either microscopy or Rapid Diagnostic Test (RDT). Treatment based on clinical grounds should only be given if diagnostic testing is not immediately accessible within 2 hours of patients presenting for treatment. Drugs for treatment are prescribed based on the type of malaria parasite discovered, the severity of the infection, age and pregnancy statue. The antimalarial medicines recommended in the WHO model list of essential medicines [10][3] for the curative treatment of malaria include; amodiaquine, artemether, artemether+lumefantrine, artesunate, artesunate+amodiaquine, chloroquine, mefloquine, sulfadiazine, pentamidine, etc.

B. *Alternative Medicine Diagnosis and Management Approach*

The use of traditional medicines for the treatment of malaria has been in existence for thousands of years and can be described as the origin of the two major classes (artemisinin and quinine derivatives) [11] of antimalarial drugs available today.

The major motivations to the use of traditional medicines today include increase in the number of drug resistance and the challenges associated with accessing of effective antimalarial services or drugs by people in the rural areas. Presently, more than one thousand plant species from over 100 families have been very useful in the treatment of malaria [12]. Some plants recommended for this purpose include: the use of Apple Cider Vinegar, Ginger, Cinnamon, Fever Nut, Orange Juice, Grapefruit, Citrus Limetta Fruit, Holy Basil, Alum, Herbal Teas, Chirayta, Datura, Fenugreek Seeds, Mustard Seed Oil and Turmeric.

IV. Related Work

In the study conducted by [13], an appraisal of ten potent African medicinal plants is presented. An up to date overview of ten potential medicinal plants from the African biodiversity which can be grouped under future phyto-pharmaceuticals for the treatment or management of several infectious and chronic diseases is carried out. Prominent scientific databases were explored to examine trends in the number of publications on the medicinal values of most African plants. A Decision Support System model for diagnosing tropical diseases using Fuzzy Logic is proposed in [14]. Results from the experiments carried out indicate that FL could be used to solve problems associated with data ambiguity, imprecision and even uncertainty. In [15], an algorithm for malaria diagnosis using fuzzy logic for treatment in Ghana is presented. A case study was conducted in Juaso District Government Hospital. MATLAB 7.8.0 was used for the design and simulation of the algorithm. A Fuzzy-Based System for the diagnosis and treatment of tuberculosis is presented in [3]. Mamdani inference method was used. The system was designed with Java, Microsoft Visio (2013), MySQL workbench, MySQL database, JSP and XHML. In [6] a survey of the medicinal plants used by traditional healers for the treatment of malaria in the Chipinge district in Zimbabwe is presented. The survey was undertaken to document how malaria is conceptualized

[1] Malaria diagnosis and treatment–Mayo Clinic.Available at https://www.mayoclinic.org/diseases conditions/malaria/diagnosis-treatment/drc-20351190

[2] Disease Management. Final Academy of Managed Care PharmacyAvailable at http://www.amcp.org/WorkArea/DownloadAsset.aspx?id=9295Medical plants used in traditional treatment of malaria. Also available at academicjournals.org/article/ar.

[3] WHO model list of essential medicines 20th list, 2017http://www.who.int/medicines/publications/essentialmedicines/en

and diagnosed by traditional healers, and to record the medicinal plants used in the prevention and treatment of malaria, their mode of preparation and administration. In the work of [16] a documentation of herbal Medicines used for the treatment and management of human diseases by some communities in Southern Ghana is presented. The results of the study showed that herbal medicines are used for treatment and management of both common and specialized human diseases and that, factors of place and time are considered important during harvesting of plants for treatments.

A review of the existing related work revealed that there is a wide gap between alternative and pure medicine in the treatment of human diseases. Also, since both pure medicine and alternative medicine treatment methods are not without weaknesses or limitations, it becomes necessary to have a management approach housing these two major treatment methods.

V. The Proposed System

The proposed system is a fuzzy based system based on the integration of both alternative and pure medicine for the management of malaria. Management in this context involves diagnosis, treatment and prevention of malaria. The intelligent of the system is derived from the result of an interactive decision process based on knowledge obtained from experts (medical doctors and alternative medicine practitioners) in the field.

The result of the diagnosis determines the next step in the management process to be taken. That is, prevention if malaria has not been diagnosed and treatment otherwise. Treatment here could be purely medical (drugs) or herbal or a combination of both when appropriate. Information in the patients' medical profile such as pregnancy status, age, presence of terminal diseases such as diabetes, ulcer etc. is also considered in the treatment process.

A. The Proposed System Architecture

System design shows the components that make up a system and also the relationship between them. Fuzzy based systems are made up of four major components; knowledge base, fuzzifier, inference engine and the defuzzifier. Some of these components are discussed in details. The architecture of the proposed system is presented in Fig. 1. It shows the connections existing between the system's components. The proposed integrated fuzzy based system consists of two major layers; the fuzzy layer and the management layer. Some of the sub-components of these layers are briefly described.

The fuzzy layer houses the fundamental components of a fuzzy based system. This is where fuzzy logic or rule is applied. As such, greater portion of the intelligent possessed by the system resides here. The components in this layer are discussed.

Fuzzifier: This component takes care of fuzzification. Fuzzifiication is the use of fuzzy membership function to change a real scalar value to a fuzzy value. Firstly, values within specified range are assigned variables or terms known as linguistic variables. The linguistic variables used in this work are; minor, moderate severe and very severe. A collection of the set of values that make up a linguistic variable is known as fuzzy set. The equation of the fuzzy set used in this work is expressed in (3), (4), (5) and (6).

$$(X) = \begin{cases} 0 & \text{if } x \le 0.1 \\ \frac{x-0.1}{0.2} & \text{if } 0.1 \le x \le 0.3 \\ \frac{0.2-x}{0.1} & \text{if } 0.2 \le x \le 0.3 \\ 0 & \text{if } x \ge 0.2 \end{cases} \tag{3}$$

$$\mu\text{moderate}(x) = \begin{cases} 0 & \text{if } x \le 0.3 \\ \frac{x-0.3}{0.3} & \text{if } 0.3 \le x \le 0.6 \\ \frac{0.45-x}{0.15} & \text{if } 0.45 \le x \le 0.6 \\ 0 & \text{if } x \ge 0.45 \end{cases} \tag{4}$$

$$\mu\text{severe}(x) = \begin{cases} 0 & \text{if } x \le 0.5 \\ \frac{x-0.6}{0.2} & \text{if } 0.6 \le x \le 0.8 \\ \frac{0.7-x}{0.1} & \text{if } 0.7 \le x \le 0.8 \\ 0 & \text{if } x \ge 0.7 \end{cases} \tag{5}$$

The development of rules which are used to determine the degree of membership is also part of the fuzzification process. A rule is fired whenever any of the precedence parameter evaluates to true. The first five rules used in this work are given in Table I.

TABLE. I. Fuzzy Rule base for Diagnosing Malaria

No	Fever	Headache	Nausea	Vomiting	Jaundice	Enlarge lever	Joint pain	Body weakness	Dizziness	Loss of appetite	MP	Conclusion
1.	m	m	m	m	M	m	m	m	s	M	m	m
2.	mo	m	m	m	M	m	mo	mo	s	S	mo	mo
3.	s	mo	m	m	M	m	m	s	s	S	mo	s
4.	vs	m	m	m	M	m	s	s	m	M	s	vs
5.	mo	m	m	mo	M	m	mo	mo	mo	S	mo	mo

[a] Legend: m – mile, mo - moderate, s - severe, vs - very severe

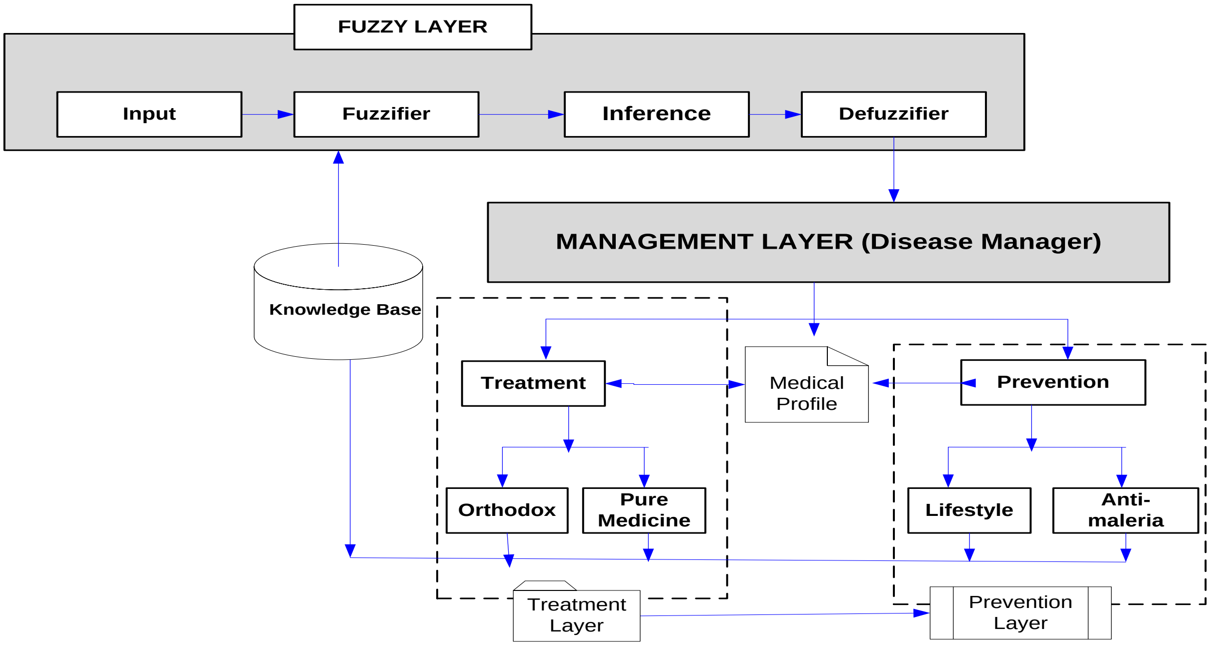


Fig. 1. Proposed System Architecture.

Deffuzzifier: It generates real values from fuzzy sets and corresponding degrees of membership. The output produced by the defuzzifier is known as crisp output. The center of gravity defuzzification method is used in this work. This is expressed in (6).

$$\mathrm{C_o\,G} = \frac{\sum \mu y(x_i) x_i}{\sum \mu y(x_i)} \qquad (6)$$

The Inference Engine: It processes the facts in the knowledge base to produce output for decision making. The intelligent of a fuzzy based system resides in the inference engine. It is modeled after the reasoning of experts in the field. The Root Sum Square (RSS) inference technique expressed in (7) is used in this work.

$$\sqrt{\sum R^2} = \sqrt{(R_1^2 + R_2^2 + R_3^2 +, \dots, R_n^2} \qquad (7)$$

Where: $R_1^2 + R_2^2 + R_3^2 + \cdots + R_n^2$ are the strength values (truth values) of different rules which share the same conclusion. The Management Layer: Management at this phase excludes diagnosis because diagnosis is carried out at the fuzzy layer. The results derived from the fuzzy layer are channeled to the management layer for further action. The major components of this layer are; the treatment layer and the prevention lager.

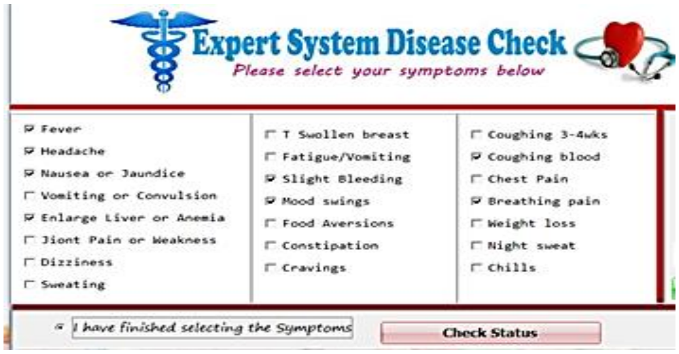
This is as also presented in the proposed system architecture in Fig. 1.

The Treatment Layer: The treatment layer is further divided into two; pure medicine which administers or recommends pure medicine (drugs) for treatment and orthodox which recommends herbs for treatment. These two treatment methods may also be used complementarily when necessary. The choice of a particular treatment method depends on users' preferences, availability and suitability. The information in the patients' medical profile which include pregnancy statue, age, present of other illness, certain drugs and herbs is considered to ensure accuracy and safety in the administration of treatments.

The Prevention Layer: This takes care of the preventive measures which could be used to control the stop malaria infection. Preventive measures in the form of anti-malarial drugs and/or change in lifestyle are prescribed or recommended for people who have not yet been infected by malaria or people that have been offered treatment for malaria. Lifestyle changes include; the use of mosquito nets, personal hygiene etc.

B. *Experimental Setup and Results*

The proposed system is implemented in C# in visual studio Integrated Development Environment. MySQL database is used for data storage. In the proposed system, adding a patient to the system activates the symptoms module shown in Fig. 2. The symptoms module shown in Fig. 3 determines the malaria statue of the patient based on symptoms such as fever, headache, etc. diagnosis based on symptoms alone could be misleading because the same set of symptoms may denote different medical conditions. To deal with this imprecision, a fuzzy analysis of the symptoms is performed by the fuzzification module shown in Fig. 4 Fuzzification is performed using the rules in the rule base given in Table I. This module determines the degree or the intensity of the identified symptoms as shown in Fig. 5. For the purpose of accuracy, the system is built to work with raw inputs i.e. symptoms and also medical laboratory test results. This figure shows the degree of each symptom based on medical laboratory test result supplied and a defuzzified or crisp value of the fuzzy value. The result derived from the fuzzy analysis and other conditions such as age and pregnancy statue is used to present appropriate management (prescription) and treatment method. Treatments which could be herbal as shown in Fig. 6 or medical presented in Fig. 7 or a combination of both are prescribed. Preventive measures as shown in Fig. 8 are also specified for all infected and non-infected patients. The result of this experiment shows that the proposed system can manage malaria. Management in this context includes diagnosis, treatment which is made up of pure medical treatment, alternative (herbal) treatment, complimentary treatment which is a combination of alternative (herbal) treatment method and pure medical treatment method.


Expert System Disease Check

Please select your symptoms below

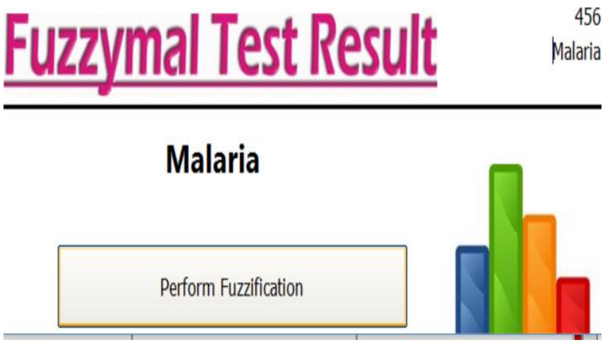
☑ Fever
☑ Headache
☑ Nausea or Jaundice
☐ Vomiting or Convulsion
☑ Enlarge Liver or Anemia
☐ Jiont Pain or Weakness
☐ Dizziness
☐ Sweating

☐ T Swollen breast
☐ Fatigue/Vomiting
☑ Slight Bleeding
☑ Mood swings
☐ Food Aversions
☐ Constipation
☐ Cravings

☐ Coughing 3-4wks
☑ Coughing blood
☐ Chest Pain
☑ Breathing pain
☐ Weight loss
☐ Night sweat
☐ Chills

● I have finished selecting the Symptoms

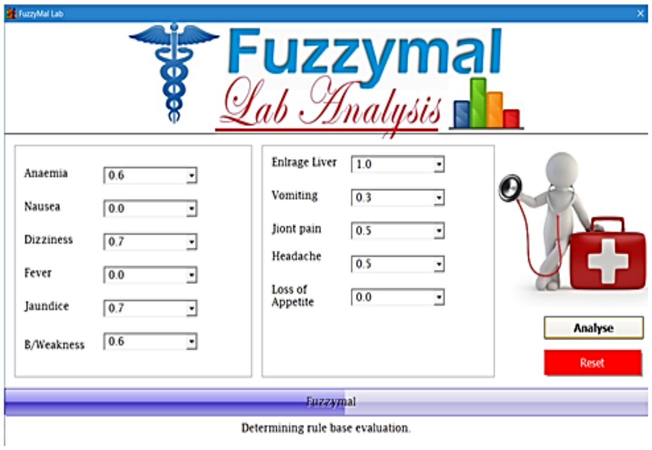
Fig. 2. Symptom Module.


Fuzzymal Test Result

456
Malaria

Malaria

Fig. 3. Fuzzification Platform.


FuzzyMal Lab

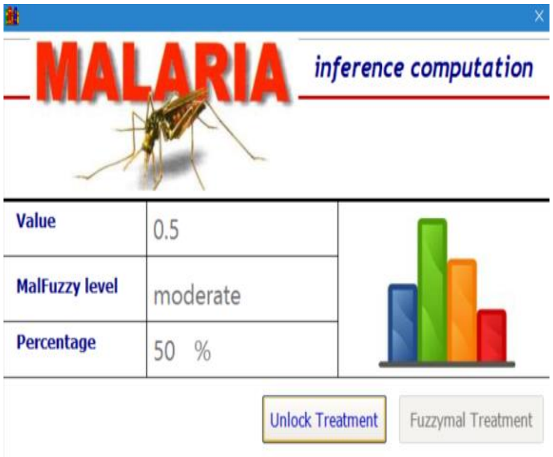
Fuzzymal
Lab Analysis

Anaemia: 0.6
Nausea: 0.0
Dizziness: 0.7
Fever: 0.0
Jaundice: 0.7
B/Weakness: 0.6

Enlrage Liver: 1.0
Vomiting: 0.3
Jiont pain: 0.5
Headache: 0.5
Loss of Appetite: 0.0

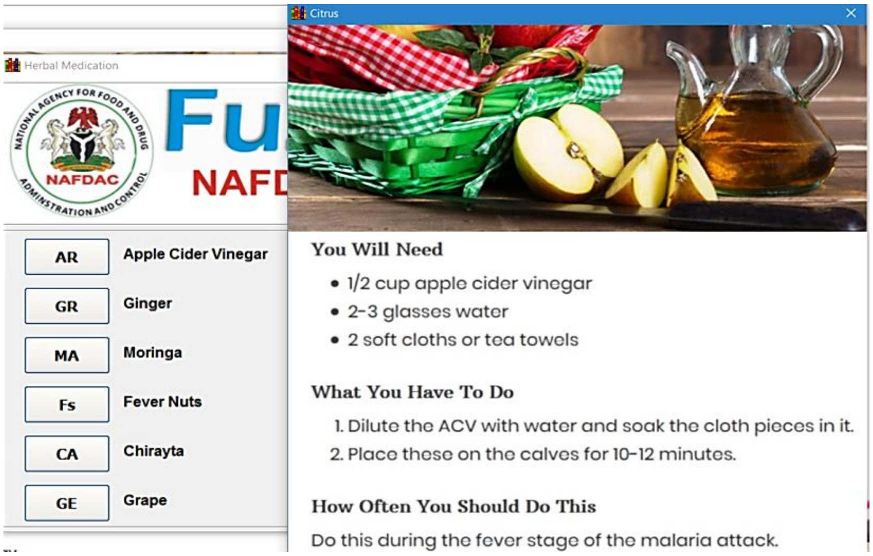
Determining rule base evaluation.

Fig. 4. Fuzzification Process.


MALARIA inference computation

Value: 0.5
MalFuzzy level: moderate
Percentage: 50 %

Fig. 5. Inference Computation.


Herbal Medication
NATIONAL AGENCY FOR FOOD AND DRUG ADMINISTRATION AND CONTROL
NAFDAC
Fu
NAFD
Citrus

AR Apple Cider Vinegar
GR Ginger
MA Moringa
Fs Fever Nuts
CA Chirayta
GE Grape

You Will Need

- 1/2 cup apple cider vinegar
- 2-3 glasses water
- 2 soft cloths or tea towels

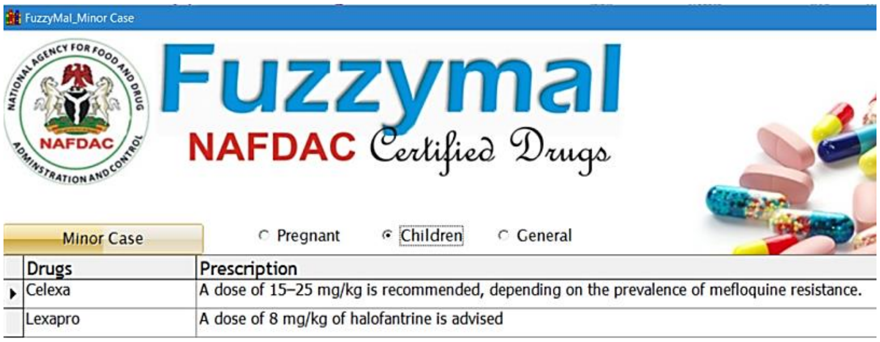
What You Have To Do

1. Dilute the ACV with water and soak the cloth pieces in it.
2. Place these on the calves for 10-12 minutes.

How Often You Should Do This

Do this during the fever stage of the malaria attack.

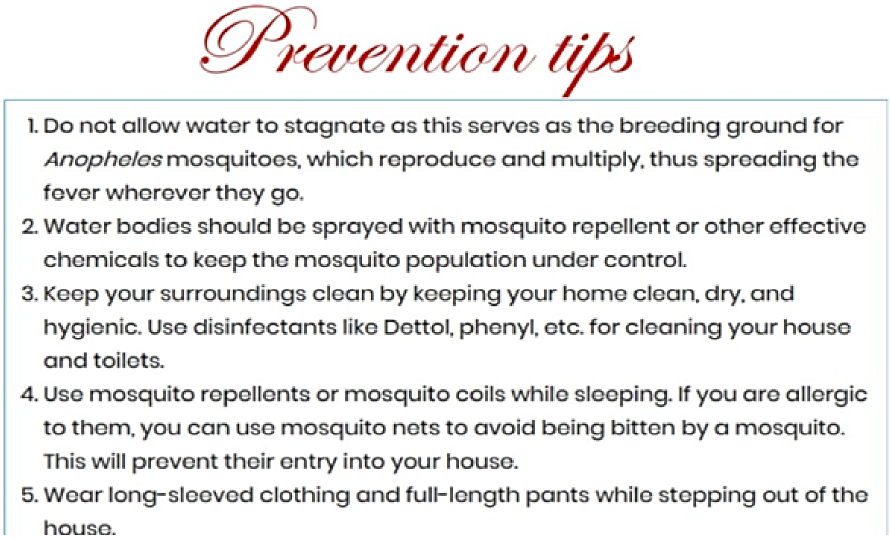
Fig. 6. Herbal (Prescribed) Treatment.



○ Pregnant ● Children ○ General

Drugs	Prescription
Celexa	A dose of 15–25 mg/kg is recommended, depending on the prevalence of mefloquine resistance.
Lexapro	A dose of 8 mg/kg of halofantrine is advised

Fig. 7. Medical (Prescribed) Treatment.



Prevention tips

1. Do not allow water to stagnate as this serves as the breeding ground for *Anopheles* mosquitoes, which reproduce and multiply, thus spreading the fever wherever they go.
2. Water bodies should be sprayed with mosquito repellent or other effective chemicals to keep the mosquito population under control.
3. Keep your surroundings clean by keeping your home clean, dry, and hygienic. Use disinfectants like Dettol, phenyl, etc. for cleaning your house and toilets.
4. Use mosquito repellents or mosquito coils while sleeping. If you are allergic to them, you can use mosquito nets to avoid being bitten by a mosquito. This will prevent their entry into your house.
5. Wear long-sleeved clothing and full-length pants while stepping out of the house.

Fig. 8. Malaria Prevention Tips.

VI. Conclusion and Recommendations

In conclusion, an integrated fuzzy based method which combines both orthodox (herbal) and pure medical treatment methods for diagnosis and management of human diseases (malaria) is proposed. With the result obtained it is clear that the proposed system can be used to manage malaria both medically and alternatively. Combining these two methods eliminates the weaknesses in the individual methods of management. However, the addition of complimentary treatment approach will be a reasonable improvement to the proposed method.

VII. Contributions to Knowledge

This work presents the following contributions to knowledge:

- Development of an integrated fuzzy based system for the management of human diseases.
- This approach improves on the use of medical drugs or herbs for the treatment of diseases.
- The study establishes the possibility of using two treatment methods for the treatment of diseases.

Acknowledgments

Our thanks go to the experts and reviewers who have contributed towards the development of this work.

References

[1] J. Awotunde, O. Matiluko, and O. Fatai, “Medical diagnosis system using fuzzy logic”, African Journal of Computing & ICT. Vol. 7 No. 2, 99-106. 2014.

[2] Y. Djam, M. Gregory, H. Kimbi, and V. Nachamada, V. “A fuzzy expert system for the management of malaria. International Journal of Pure and Applied Sciences and Technology” Vol. 5 No.2, 84-108.2011.

[3] A. Angbera,M. Esiefarienrh, and I. Agaji, “Efficient fuzzy-based system for the diagnosis and treatment of tuberculosis”, International Journal of Computer Applications, Technology and Research, Vol. 5, Issue 2, 2016.

[4] C. Chuen, “Fuzzy logic control system: Fuzzy logic controller –part I. IEEE Transaction on systems, man, and cybernetics”, Vol.20, No.2, 404 -418. 2014.

[5] G. William, “An optimization approach to employee scheduling using fuzzy logic”. MSc. Thesis, California Polytechnic State University, San Luis Obispo). 2011.

[6] N. Talkmore, E. Charlott, A.Klooster, M. Jong,., and V. Jan (2015) “Medicinal plants used by traditional healers for the treatment of malaria in the Chipinge district in Zimbabwe”. Journal of Ethno-pharmacology, Vol. 159, 224-237.

[7] Schrijvers G. (2009). Disease management: a proposal for a new definition. International journal of integrated care, 9, e06.

[8] Malaria diagnosis and treatment–Mayo Clinic. Available at https://www.mayoclinic.org/diseases conditions/malaria/diagnosis-treatment/drc-20351190.

[9] Disease Management. Final Academy of Managed Care Pharmacy. Available at http://www.amcp.org/WorkArea/DownloadAsset.aspx?id= 9295Medical plants used in traditional treatment of malaria. Also available at academicjournals.org/article/ar.

[10] WHO model list of essential medicines 20th list, 2017. http://www.who.int/medicines/publications/essentialmedicines/en

[11] Willcox, M. L., & Bodeker, G. (2004). Traditional herbal medicines for malaria. *BMJ (Clinical research ed.)*, *329*(7475), 1156–1159

[12] Treatment and Management of Malaria Parasite. Available at http://www.malaria.com/questions/malaria-treatment-management

[13] F. Mahomoodly, "Traditional medicines in Africa: An appraisal of ten potent medicinal plants".Evidence-based complementary and alternative medicine. Vol. 2013, Article ID 617459. 2013.

[14] S. Olabiyisi, E. Omidiora, M. Olaniyan, and O. Derikoma, "Decision Support Model for Diagnosing Tropical Diseases Using Fuzzy Logic", Afr. J comp & ICT Vol. 4. No. 2, 1-6, 2011.

[15] D. Quashie,K. Joseph, and B. James, "Designing algorithm for malaria diagnosis using fuzzy logic for treatment", International Journal of Computer Applications Vol. 91, No.17. 2014.

[16] A. Boadu and A. Asase, "Documentation of herbal medicine used for the treatment and management of human diseases by some communities in southern Ghana. Evidence Based Complementary and Alternative Medicines" Vol. 2017. 2017.

A Hybrid Intrusion Detection System for SDWSN using Random Forest (RF) Machine Learning Approach

Indira K[1]
School of Computing
Sathyabama Institute of Science and Technology
Chennai, India

Sakthi U[2]
Department of Computer Science and Engineering
St.Joseph's Institute of Technology
Chennai, India

***Abstract*—It is indeed an established fact which network security systems had certain technical problems that mostly tends to lead to security risks. Nowadays, Attackers could still continue to abuse the security vulnerabilities as well as shatter the systems and networks, and is quite pricey and even sometimes extremely difficult to resolve all layout and computing faults. The above appears to suggest that methodologies relying on preventive measures seem to be no longer secure and perhaps tracking of intrusion is necessary as a last line of defense. A Hybrid in Software Defined Wireless Sensor Network (SDWSN) the Intrusion Detection System is designed for this paper which really incorporates the benefits of Salp Swarm Optimization (SSO) algorithm as well as the classification of Machine Learning method it is based upon Random Forest (RF). We propose SSO optimization procedures to guarantee that the ideal features for the intrusion detector are chosen and in addition for improving the Random Forest (RF) classifier detection efficiency. To assess / calculate the reliability of the proposed approach here we make use of the generic NSL KDD dataset. Therefore, our proposed hybrid IDS-SSO-RF classifier further analyzes these detected abnormal activities. The known and unknown attacks are also identified. Hybrid framework also shown by the experimental results can reliably detect anomaly behavior and obtains better results in terms in terms of delay, delivery ratio, drop overhead, energy consumption and throughput.**

Keywords*—*SDWSN; IDS; Salp Swarm Optimization; Random Forest Classifier

I. Introduction

Increasing computer data size has made the protection of information more critical [1]. Protection of information means protecting from unauthorized access to information and information systems [2]. When data is accessed in a network environment and transmitted via an unreliable medium, information security becomes more important [3]. Network security approaches can be typically divided in two main categories. They are (1) prevention-based techniques (2) detection based techniques [4]. The Detection-based technique strategies seek to detect intrusions that impact data centers after prevention-based techniques have failed [5]. Hence a detection system for intrusion is a detection-based strategy that detects malicious or anomalous activity by either networks or other devices [6]. By admiring defensive technologies like those of firewalls, fast encryption [30] and user access IDSs has become a key component of corporate IT security management and they are defined as systems based on misuse or anomaly [7]. In [27], only IDS are deployed in cluster heads not in every node. This method saves energy for remaining nodes and minimizes computational cost.

Technology is evolving rapidly every day and so many advancements and software developments are always being designed to protect Computer Systems from every network intrusion assault that involves various machine learning, deep learning [33] and heuristic approaches [8]. Some of the sophisticated methods in this sense, such as those focused on machine-learning techniques, e.g. Support Vector Machines (SVMs), Artificial Neural Networks, Fuzzy Logic, Bayesian Networks, Decision Trees, Random Forests, Clustering and Methods Ensemble [9][10][11][31][32]. Likewise heuristic methods like those of Genetic Algorithm (GA), Particular Swarm Optimization Algorithm (PSO), Ant Colony Optimization Algorithm (ACO), and Cuttlefish Optimization (GWO) [12][13]. Each of these approach-based IDSs, though, must produce low false-positive levels as well as comply with a large database for learning and estimation of imbalanced datasets, categorical and continuous features and a large number of features [14]. In [29], classification is done by combining SVM and KNN. In addition, some researchers are working on Data Mining (DM) technique. In a secure network environment increasingly used to detect these attacks, anomalies or intrusions [15][16].

In this work, we intend to introduce a Knowledge and Behavior-based Hybrid Intrusion Detection System (IDS) in a Software Defined Wireless Sensor Network in which Salp Swarm Optimization (SSO) and Random Forest (RF) classifier based on Decision Tree approaches plays a major role in ensuring the ideal features for intrusion detector selection and improved detection efficiency. The manuscript remaining segment is structured in the following section. In section II using machine learning algorithm and other similar works on IDS a comprehensive literature survey has convey the various intrusion detection systems. Section III addresses about the Preliminaries i.e. in proposed approach has various modules. Section IV discusses about our proposed framework. Section V by assess the performance on several metrics discuss about the simulation results of the proposed framework. Section VI provides the ending remarks.

II. Related Work

Some of the researchers past research works have been briefly described in this section among the various research works on intrusion detection method. Table I indicates the writers ' literature works with its merits and demerits.

Saurabh Dey (2019) et.al[17] presented to involve heterogeneous client networks an intrusion detection scheme for mobile clouds based on machine learning. The suggested strategy doesn't really probably require regular updates to the rules, and its level of complexity can be tailored to meet client network requirements. Technically, there are two steps in the presented scheme: multi-layer traffic screening and Virtual Machine (VM) selection based on decisions. Their experimental results show however presented scheme was really pretty constructive at intrusion detection.

Huijun Peng (2018) et.al[18] provides flow detection method based by SDN, develops anomaly SDN flow detection structures and performs flow classification detection for K-nearest neighboring algorithms using transductive confidence machines double P-value. The experiment demonstrate results that perhaps the presented algorithm reaches higher accuracy, a lower false positive rate and better adaptation SDN setting than other related algorithms.

D. Jianjian (2018) et.al[19] has primarily introduced an intrusion detection algorithm depend upon enhanced AdaBoost-RBFSVM, developed a WSN denial of service (DoS) intrusion detection system (IDS) on based the presented method. The learning result was achieved to render the RBF-SVM algorithm as the soft classifier of AdaBoost. The IABRBFSVM algorithm was presented using the impact of parameter π on RBF-SVM on the smoothness of AdaBoost weights and the template training error effect. But the other hand, the eigen space for all the attack were proposed after evaluating the DoS attack, as well as the corresponding framework for intrusion detection were developed. The presented IDS can be modeled.

Due to the inherent transparency of the communication channel, wireless networking is vulnerable to specific amount of cyber-attacks and intrusion attempts. Among other threats, electronic jamming attack stands out. As the sophistication of the attacks continues to increase, it is necessary to develop new and more reliable detection mechanisms. To address the issue of IEEE 802.11 networks electronic jamming attacks, D. Santoro (2017) et.al[20] present a novel Hybrid-NIDS (HNIDS) on based the proof theory from Dempster-Shafer (DS). The suggested approach is intended to combine the advantages of NIDSs based on signature and anomaly.

Kai Lin (2016) et.al [21] introduced a new globoid model to assess for the quality of all-directional detection during efficiently network saving the energy, dividing the sensing field into the outermost shell and interior region. Initially, they present an outermost shell coverage algorithm to ensure intruding events ' recognition performance. Then a model of Markov prediction was designed to predict the probability of movement in the adjacent area based on intruders ' historical trajectories. Using SDR technology various working frequencies will allocated to the protected nodes, according to the expected performance. In addition, a path correction plan was suggested during the operation to retrieve the missing intruders. The quality evaluations demonstrate the efficacy. Our scheme is in terms of network life, path estimation exactly and correction strategy success rate.

TABLE. I. Literature Survey

Year	Author	Contribution	Merits	Demerits
2019	Saurabh Dey, et.al	A Machine Learning Based Intrusion Detection Scheme for Data Fusion in Mobile Clouds involving Heterogeneous Client Networks	rule updates does not need. in terms of intrusion detection highly effective	Feature extraction based on interpacket delayis used here in which, the sender application times out and resends the packet sometimes when the queuing delay is through.
2018	H. Peng,. et.al	A Detection Method for Anomaly Flow in Software Defined Network	reaches a lower false positive rate. higher precision. better adaptation to the SDN environment	KNN does not work well with large dataset. KNN is sensitive to noise in the dataset
2018	D. Jianjian, T. Yang and Y. Feiyue	For Wireless Sensor Networks a Novel Intrusion Detection System based on IABRBFSVM	improved performance of network by detecting and removing malicious nodes in the network	Less Classification accuracy.
2017	D. Santoro, et.al	For Virtual Jamming Attacks on Wireless Networks a Hybrid Intrusion Detection System	to generate the hybrid IDS 100% DR, 3.8% of FPR and 0% FNR.	Dempster Shafer theory of eveidence has a problem of Potential computational complexity. It lacks a well-established decision theory
2016	Kai Lin, et.al	In SDR-based 3D WSNs node Scheduling for All-directional Intrusion Detection	improved lifetime and trajectory prediction accuracy	Complex Algorithm
2014	S. Shamshir, et al.,	In wireless sensor networks for detecting intrusion Cooperative fuzzy artificial immune system utilized	improves detection accuracy, successful defense rate	Random and uneven distribution of cluster heads by LEACH

S. A bio-inspired process, the cooperative-based fuzzy artificial immune system (Co-FAIS) has been implemented in the paper by Shamshir (2014) et al [22]. It is a modular defense strategy. It is based on the human immune system's risk concept. In terms of background antigen value (CAV) or attackers the agents accompany and collaborate with each other to measure the abnormality of sensor activity, to change the protection response threshold for fuzzy activation. By evaluating the packet components and sending the log file to the next layer sniffer module adapts to the sink node to inspect information in such a multi-node situation. To identify hazardous signal sources the fuzzy detector module combines with a hazard detector system. The contaminated origins were passed to the Fuzzy Q-learning vaccination modules (FQVM) to improve device capabilities in general. To order to produce optimal security techniques, the Cooperative Decision Making Modules (Co-DMM) combines risk detector module with the fuzzy Q-learning vaccine module. Using a network simulator the Low Energy Adaptive Clustering Hierarchy (LEACH) was tested to determine as the efficiency of the presented model.

III. Premilinaries

A. Random Forest (RF): A Machine Learning Approach [23]

Random Forest [26][28] is a learning model for an ensemble that takes tree choice as a fundamental classifier. As the name suggests, with an amount of trees, this algorithm produces the forest. If more trees in the forest, it appears the more robust forest. Similarly, in forest greater amount of trees provides the outcomes of elevated precision in the random forest classification. When entering a sample to be categorized, the final outcome of classification is determined by a single decision tree's output vote. Random forest overcomes decision trees ' over-fitting issue, has excellent noise and anomaly values tolerance, and has excellent scalability and parallelism to the issue of high-dimensional data classification. In contrast, random forest is a data-driven, non-parametric method of classification. It trains rules of classification through sample learning and does not involve previous classification understanding.

The model of random forests is based on forests of K choice. Each tree votes on which class a specified independent variable X belongs to, and the class it deems most suitable is given only one vote. The K decision trees description is as follows:

$$\{h(X,\theta_k), k = 1,2,.........,k\} \tag{1}$$

K represents amount of decision trees in random forests, among them. θ_k reflects random vectors that are autonomous and identically distributed. The technique of random repeated sampling is implemented to randomly extract K samples as a self-service sample set from the initial training set, and then generate regression trees for classification K. Assuming the initial training set has n characteristics, m characteristics are chosen randomly at each tree node (mn). By computing the quantity of data in each feature, a feature with the most classification capacity is chosen for node splitting among the m characteristics. Every tree develops without cutting to its peak. The trees produced are made up of random forest, and random forest classifies the fresh information. The outcomes of the classification are determined by the amount of tree classifiers votes. The resemblance and correlation of decision trees are significant characteristics of random forest to represent efficiency of generalization, while generalization error represents the system's capacity to generalize. Generalization capacity is the system's capacity to make right decisions outside the training sample set on fresh information with the same distribution. Smaller mistake in generalization can cause the scheme.

The similarity and correlation of decision trees are important features of random forest to reflect generalization performance, while generalization error reflects generalization ability of the system. Generalization ability is the ability of the system. To make correct judgments on new data with the same distribution outside the training sample set. Smaller generalization error can lead to better results of the scheme and increased generalization capability.

The error of generalization is specified as follows:

$$PE^* = P_{x,y}(mr(X,Y) < 0) \tag{2}$$

Where PE* represents a generalization error, datatype X, Y shows the probability definition area and margin function is mr(X, Y).The margin function is defined as follows:

$$mr(X,Y) = avg_k I(h(X,\theta_k) = Y) - \max_{j \neq y} avg_k I(h(X,\theta_k) = J) \tag{3}$$

In which, X is the sample of the input, Y is the right classification and J is the wrong classification. I (g) is an indicative function, avgk(g) is an average function, and h(g) is a classification model series. The margin function reflects the extent to which the numbers of votes corresponding to sample X for the correct classification exceeds the maximum number of votes for other incorrect clauses. The greater the margin function value, the greater the classifier's credibility will be. The generalization error convergence expression is described as follows:

$$\lim_{k\to\infty} PE^* = P_{x,y}(P_\theta(I(h(X,\theta_k) = Y)) - \max_{j \neq y} P_\theta(I(h(X,\theta_k) = J))) \tag{4}$$

The formula above shows that the generalization error will tend to an upper limit, and the model will not over-fit with the rise in the amount of decision trees. Depending on the classification intensity of the single tree and the correlation between the trees, the upper limit of the generalization error is accessible. The random forest model aims to establish a random forest with low correlation and high classification intensity. Classification intensity S is the mathematical expectation of mr(X, Y) in the whole sample space:

$$S = E_{x,y} mr(X,Y) \tag{5}$$

Both θ and θ′ vectors are autonomous and identically distributed and the correlation coefficients of mr(θ,X,Y) and mr(θ′,X,Y) are described as follows:

$$\bar{\rho} = \frac{\text{cov}_{x,y}(mr(\theta, X, Y), mr(\theta', X, Y))}{sd(\theta)sd(\theta')} \quad (6)$$

Hence, Sd(a) can be articulated as follows, among them:

$$sd(\theta) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(mr(x_i,\theta) - \frac{1}{N}\sum_{i=1}^{N}mr(x_i,\theta))^2} \quad (7)$$

In Equation (6), it is possible to measure the correlation between the trees of h(X,θ) and h(X,θ′) on the X and Y dataset by means of ρ¯. The greater the ρ¯, the greater the coefficient of correlation, the upper limit of generalization error can be obtained from Chebyshev's inequality:

$$P_{x,y}(mr(X,Y) < 0) \le \frac{\bar{\rho}(1 - s^2)}{s^2} \quad (8)$$

To see the random forest boundary generalization error is negatively linked with a single decision tree's classification intensity S and strongly correlated with the decision trees correlation P. Therefore, the higher the intensity of classification S, the lower the correlation P, The lower the generalization error limit, the greater the accuracy of the classification.

B. *Salp Swarm Algorithm (SSA): A Metaheuristic Technique [24]*

Salps have a transparent body in the form of a barrel. Salps belong to the Salpidae family. Their skin cells are very comparable to those of jelly fish they comparable to jelly fish and also migrate, where water is pumped as propulsion through the body to move forward. In profound oceans Salps frequently form a swarm called the Salp chain. The primary reason of this behavior is not yet evident, but few scientists think is accomplished by using fast coordinated adjustments and foraging to achieve better locomotion. To mathematically model the Salp chains population is first split into two groups (1) leader and (2) supporters. The leader is perhaps the Salp in front of chain, while the remaining salps were regarded as followers. The leader guides swarm and supporters pursue each other (and leader straight from indirectly), as the name of these salps suggests.

In an n-dimensional search space, where n is the number of variables of a given problem, salps position is defined like other swarm-based techniques. Hence a two-dimensional matrix called x, is stored the location of all salps also thought the search space there is a food source called F as the goal of the swarm. The following equation is suggested for updating the leader's position.

$$x_j^1 = \begin{Bmatrix} F_j + c_1((ub_j - lb_j)c_2 + lb_j) & c_3 \ge 0 \\ F_j - c_1((ub_j - lb_j)c_2 + lb_j) & c_3 < 0 \end{Bmatrix} \quad (9)$$

Where x_j^1 shows the j-dimensional position of the first Salp (leader), F_j is the j-dimensional position of the food source, ub_j indicates the j-dimensional upper bound, lb_j indicates the j-dimensional lower bound, c1, c2, and c3 are random numbers. The above equation demonstrates that only with regard to the food source, the leader updates his stance. The coefficient c1 is the SSA's most significant parameter because it balances exploration and exploitation as follows:

$$c_1 = 2e^{-(\frac{4l}{L})^2} \quad (10)$$

Here, the present iteration is l and the highest amount of iterations is L. The parameters c2 and c3 are evenly produced random numbers in the [0,1] interval. In reality, they determine whether the next j-th dimension position should be towards positive infinity or negative infinity along with step size. The following equations are used to update the followers ' position.

$$x_j^i = \frac{1}{2}at^2 + v_o t \quad (11)$$

If i≥2, x_j^i indicates the position of the i-th follower salp in the j-th dimension, t is time, v 0 is the original velocity, and a calculation is as follows:

$$a = \frac{v_{final}}{v_o} where v = \frac{x - x_o}{t} \quad (12)$$

Since iteration is the time in optimization, the difference between iterations is equivalent to 1, and considering v 0=0, the following equation can be expressed as;

$$x_j^i = \frac{1}{2}\left(x_j^i + x_j^{i-1}\right) \quad (13)$$

Where i≥2 and x_j^i shows the position of i-th follower salp in j-th dimension. With equation (9) and (13), the salp chains can be simulated.

IV. Proposed Methodology

Internet security has become even more essential for personal computer subscribers, companies and indeed the army. With those of the invention of the internet, privacy has now become a significant issue, as well as the continuity of privacy enables for a better comprehension of the development of safety features. The entire sector of internet security is massive even in a developmental phase. The research range includes a good summary referring back to the origins of the internet and the current network security growth. In order to comprehend the analysis about the significance of safety to be carried out today and varieties of assaults in the networks, this article focuses on the unique hybrid Intrusion Detection System design and its evaluation in the Software Defined Wireless Sensor Network (SDWSN).

Major obstacles for recognizing intrusion from the Software Defined Wireless Sensor Network (SDWSN) were as stated [25].

The type of attack is diverse and the sources and features of attacks in wireless sensor networks differ more widely from traditional computer networks, including most of the attacks in link layer and network layer assaults that are unique to wireless sensor networks.

Standard computer network assets, like those of networks, directories, system records as well as functions, should never be used in software defined wireless sensor networks, and we have to take into account the functionalities of info which could be used to monitor intrusion in a wireless network sensor.

There are still many different attacks on SDWSN that seem to be different from typical networks. The main issue is enhancing the effectiveness for intrusion detection system to pinpoint unidentified threats as well as to choose the applicable methodologies. Some methods seem to be appropriate for only the identification of noted threats, whilst others are ideal for the identification of unidentified threats.

A. *Intrusion Detection System*

Intrusion detection system monitors scheme activities in a specified setting dynamically and chooses whether these activities look like an assault or not. A primitive intrusion detection system is a detector that processes data that is to be protected from the attackers. The detector's mechanism is to remove unnecessary data from the inspection trial and portray a synthesized point of view of the users ' attitude related to safety. An ultimate decision is then taken to assess the likelihood that these actions can be regarded as intrusion symptoms. Reliable IDS is usually created by using information mining methods because they can detect intrusions excellently and execute generalizations appropriately. Indeed, it can be obviously complicated to implement and install such systems. The intrinsic complications of the schemes could classify into separate problem sets. It is based upon skill, precision and usability parameters. In Fig. 1 the basic structure of IDS represents below. Intrusion Detection System works on specific systems in which the network access to the system, i.e. sending and receiving packets, is monitored and controlled and at the same time the device file auditing is done, and the system administrator is notified about the same if there is any discrepancy. This IDS device installed in the system frequently track the computer's operating system and the benefit of this device is that it can track the entire system reliably and does not allow any other equipment to be mounted.

For attackers, valuable data is always appealing and therefore susceptible to focused network assaults. Intrusion relates with phase when intruder joins the system or system server that transfers malicious packets to client scheme for any private or significant data it can be steeled, modified or corrupted i.e. an attack relates for illegitimate network packets transmission such as user misuse, system misconfiguration, or program failures, the intrusion may occur over the server or system because of current system vulnerabilities. By placing together various vulnerabilities, one can also create a smart intrusion. In a worldwide network large numbers of internet services and millions of large servers run in the scheme. Around the same moment, such networks become more appealing to more attackers and therefore need smart intrusion detection models to protect their network system. Nevertheless, IDS built using data mining methods, primarily these methods on based anomaly detection show a greater proportion of false positive occurrences compared to earlier detection methods. Thus, processing information audit and detecting internet intrusions is hard for these methods. In addition, the learning method of the system needs big quantities of training data and excellent complexity compared to the present methodologies available. Building effective intrusion detection is therefore essential in the protection of the network system and helps to detect assaults over the network. Hence, a hybrid model for intrusion detection based on classification and a range of features are suggested here.

B. *Proposed Hybrid IDS*

Well into the network, the typical behavior of individuals is nothing more than an unusual practice, and one that allows the free flow of information imbalanced by usual activities and abnormalities. To enhance the IDS detection efficiency, this article presents a hybrid Intrusion Detection System (IDS). It is based upon optimized machine learning algorithm. An Intrusion Detection (ID) was described as an operational fault that is malicious and externally caused. IDS play a key role in identifying attacks, in order to detect attack, in the paper it is suggested to use a hybrid IDS technique using Salp swarm optimized random forest classifier. Our main goal is to improve the detection rate and decrease the false alarm discovery rate while identifying attacks. Salp Swarm Optimization (SSO) eliminates redundant features, and Random Forest (RF) detects attack and initiates the alert system. We suggest SSO techniques to ensure that ideal features are selected for the intrusion detector, in addition to enhance the detection efficiency of the Random Forest (RF) classifier, we used SSO for optimization.

Our Hybrid IDSs is the combination of knowledge-based approaches as well as behavior-based approaches. They generally include two detection functions; i.e. one is to accountable for identifying well-known attacks using signatures, while the other module is accountable for identifying and discovering ordinary and harmful patterns or monitoring change from ordinary profile. Hybrid IDSs are more precise with fewer false positives in terms of attack detection. However, in Software Defined Wireless sensor network the precision of the knowledge-based system completeness needs regular updating of the information of attacks. Potential for very low false alarm rates is the advantages of knowledge-based methods and that intrusion detection suggests situational analysis. Detection methods for behavior or anomaly by observing a deviation from the system's normal behavior suppose an intrusion can be detected. The normal behavior model is obtained by multiple means from the reference data gathered. This model is then compared to the present activity by the intrusion detection system and if a deviation is detected, an alarm is raised. Thus our proposed strategies of hybrid intrusion detection can identify efforts to exploit fresh and unforeseen vulnerabilities. They also assist to detect kinds of assaults that do not effectively involve exploiting any vulnerability to security previously. The structure of our proposed Intrusion Detection System (IDS) was shown in Fig. 2.

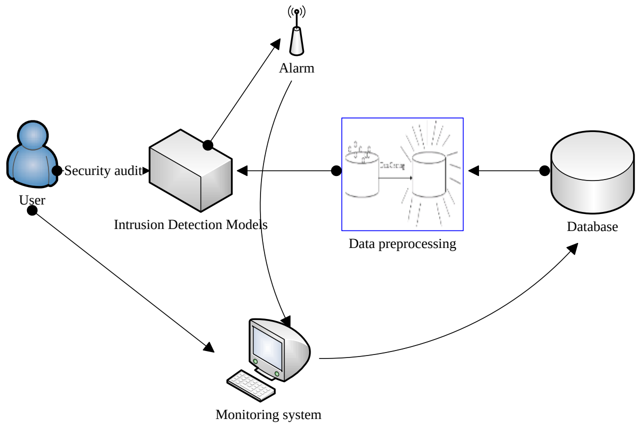


Fig. 1. Structure of IDS.

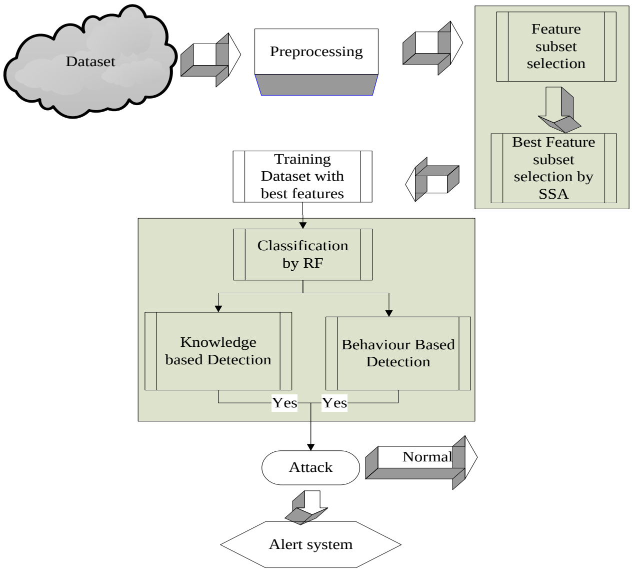

Dataset
Preprocessing
Feature subset selection
Best Feature subset selection by SSA
Training Dataset with best features
Classification by RF
Knowledge based Detection
Behaviour Based Detection
Yes
Yes
Attack
Normal
Alert system

Fig. 2. Proposed Hybrid IDS System.

Data preprocessing step is usually initially utilized information mining. It is efficient for reducing dimensionality and removes irrelevant characteristics which diminish the precision. Here, a Salp Swarm Optimization algorithm is chosen to find an optimal dataset with best features as the input to the classifier. Our proposed Random forest (RF) is a group category. It is used to enhance the precision. Random forest has many decision trees. When compared other traditional classification algorithms Random forest has low classification error. For splitting each node number of trees, minimum node size and number of features are used. In random forest when constructing individual trees, to select the type of attack by split on randomization is applied. But in this work instead of randomly selecting the features we make use of Salp Swarm Optimization algorithm to find an optimal dataset with bestfeatures.

We first choose a set ofdata from the dataset and optimal featuresof the selected data is selectedby using the SSO algorithm. The entropy of each feature is calculated by using equation (14).

$$E = \sum_{\forall c} p(c) \ln \frac{1}{p(c)} \tag{14}$$

Assume that if there is an data i, feature j is used to define the split quality, which is stated as

$$Q(i, j) = \exp\{-(E_i + E_r)\} \tag{15}$$

Based on the amount of information contained in each feature, a feature with the most classification ability is selected among the k features to split the type of data withmost important and unimportant features until the decision tree grows to the maximum.

Which and how many features are important we do not know. Hence, to find the important features we take an SSO algorithm strategy. Initially, from the ranked list, we mark top features as ‘important’ and rest of the features as ‘unimportant’.

Consider A is the set of important features and B is the set of unimportant features. At each iteration these sets of features are updated.

Based upon selection criteria the features which satisfy the condition will be separated as important and unimportant features and new dataset is formed. Selection means selecting the minimum number of features that are essential for the classifier to define the normal and intrusive activity effectively and efficiently.

The generated new set of data with best features θ_k, $\{h(X, \theta_k), k = 1,2,.........,k\}$ are the input to the random forest classifier, random forest is used to classified new optimized dataset which is shown in Fig. 3. To detect the attack the final categorization solution are decided by the number of votes of the tree classifiers.

Random forest technique operates on dividing the rule and conquering system used in the task of classification. It amalgamates a group of vulnerable learners as it is an ensemble method to create well-built leaner that can exactly categorize the information. It unites the bagging system and random feature choice. In random forests, N number or tresses are generated. Each tree reflects malicious classes that are regular and different.

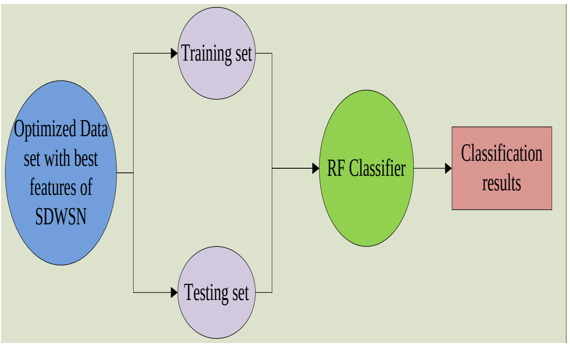


Fig. 3. Random Forest Classifier.

A big amount of datasets are readily managed by an algorithm of random forests. However, the choice of features by our suggested SSO algorithm improves several problems of the Intrusion Detection System. The pseudo code for our proposed approach in Algorithm 1 represents below.

Algorithm 1: Pseudo code for proposed Hybrid IDS

```
Input: NSL KDD dataset of SDWSN
Output: Classified result as attack or not
Initialize the Salp Population by the Dataset of
SDWSN by considering ub and lb
While ( end condition is not satisfied)
Calculate the fitness the data with specific feature
conditions
F= Data with best features
Update the optimal data list one by one
For each salp of xi
 If (i==1)
 Update the data with most important features
  Else
  Update the data with unimportant features
  end
End
Amend the salps based on the upper and lower bound
of the data
Return f
The "N" characteristics of optimal data set are
randomly selected where k << n
Create the root node N;
if (T belongs to same category C)
  {leaf node = N;
Mark N as class C;
 Return N;
  }
For i=1 to n
{Calculate Information gain (Ai);}
: ta= testing attribute;
N.ta = attribute having highest information gain;
if (N.ta == continuous )
create "n" the number of trees that a forest builds.
{ find threshold;}
For (Each T in splitting of T)
if (T is empty)
{child of N is a leaf node;}
else
{child of N= dtree T)}
calculate classification error rate of node N;
return N;
```

For the proposed system can operate in both during the offline and online phase; the training dataset is passed through the classifier while offline. This RF classification module builds the patterns that are useful for detecting intrusion. Feature selection algorithm and parameter construction with random forest algorithm is employed in this module which handles the imbalanced intrusions and after the patterns are mined, they are sent as an input to the hybrid Intrusion detector module. Similarly the intrusions are identified during the internet stage also in which the Network traffic captures the packets. For each link captured from network traffic, the pre-processors produce the feature characteristics. The detector module classifies the ordinary traffic or intrusion relation. It utilizes the models constructed in the stage of offline. Finally, the system raises an alert when attack is identified.

V. Performance Evalution

The simulation framework developed with Network Simulator (NS2) to test the data set as well as the methodology with code and devices to monitor data processing and performance. The performance evaluation using KDDCup99 dataset is the standard which includes a broad range of threats that represent serious-world intrusions in a database server. For training and testing the configuration of the workbench is carried out via the dataset KDD cup 99 among that 20% of the dataset is used here. The metrics examined to determine the feasibility of the solution proposed were listed below.

Delay: The system delay determines that how much time the data is needed to migrate to the destination across the network from the source node. Additionally, time needed to compute and find the malicious data travels on the proposed IDS model will also determine network delay. Fig. 4 demonstrates our suggested solution to traditional methods in accordance with the delay study it display the existing approach in red lines and the blue line indicates the model proposed. The overall delay is defined as time the packet / data take to reach senders through the SDWSN's network recipients. The figure shows that for our proposed strategies, the overall performance delay is minimal and the average delay for the proposed solution wireless network is 75% less than that of other KNN based approach.

Delivery Ratio (DR): The number of packets / data received at the receivers end is calculated with respect to the amount of packets transmitted at the transmission end is shown in Fig. 5. For an effective network performance, the effective SD wireless sensor network must have a significant DR quality. The DR is greater than the KNN based design for the proposed IDS-RF-SSO. For the proposed method maximum value for DR is 0.93, 0.90 and 0.720, 0.721, 0.602 at the same time current values which the existing approach are 0.736 and 0.667 and 0.493 respectively.

If the nodes are delivered to the destined user accurately then the possibility of delivery ratio is maximum and it is shown deliberately in Fig. 5. In which the blue line depicts the proposed approach with maximum delivery ratio of all is about 0.95 and the red line depicts the existing approaches with minimum delivery ratio of all is about 0.49 than that of our proposed approach i.e. our proposed RF-SSO approach delivery ratio is 51 % better than that of KNN based IDS system.

Drop: The drop is evaluated in Fig. 6 using the number of packets / data received at the recipient end to determine the amount of data drops. For successful system performance, effective SD wireless sensor network needs a significant decrease value. The probability of a drop is small if the Identification system works effectively.

In Fig. 6 the drop analysis of our proposed approach with the existing KNN classifier is shown. In whichfor the proposed IDS-RF-SSO the drop is minimum but for KNN it is maximal. Minimum drop value of the proposed approach is 2313, 1120, 3111, 2735, 3277 whereas the existing approaches values are 2920, 22285, 14100, 14500, 20440 respectively.

Energy consumption: That node only devotes the number of resources that are not required for transmission of packets in the output queue to intrusion detection. In fact, once a packet is identified as malicious it will be excluded, as long as its evaluation is not done or labelled as good; it will be redirected according to a proposed algorithm to the destination. The values of energy consumption relating to non-malicious packets are also shown in Fig. 7 when the intrusion detection is carried out at the destination; in addition, the expense of the evaluation itself remains the same and packets are not discarded long before reaching the destination.

As the number of malicious packets increases, the amount of energy saved by early detection and discarding them also increases. In Fig. 7 because of the IDS-RF-SSO algorithm, for attacks the detection rate is better than other algorithms; malicious nodes could be detected and removed faster, slowing down the average node residual power. But energy is exhausted as some nodes, is gradually completed data packet transmission task and gradually reduced the average node's energy consumption.

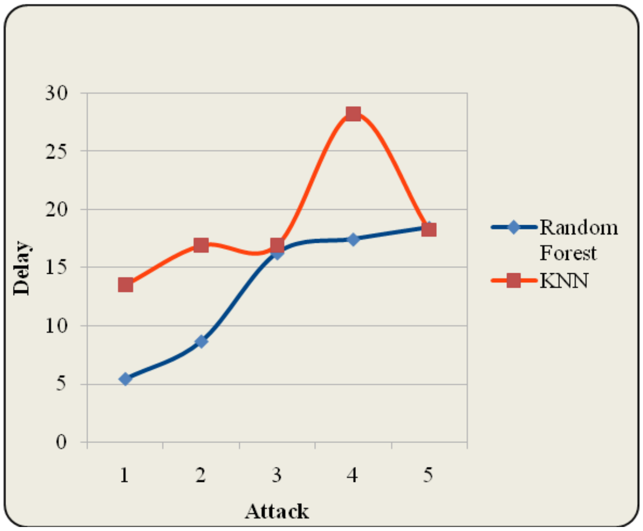


Fig. 4. Comparative Delay Analysis of RFS-SSO and KNN.

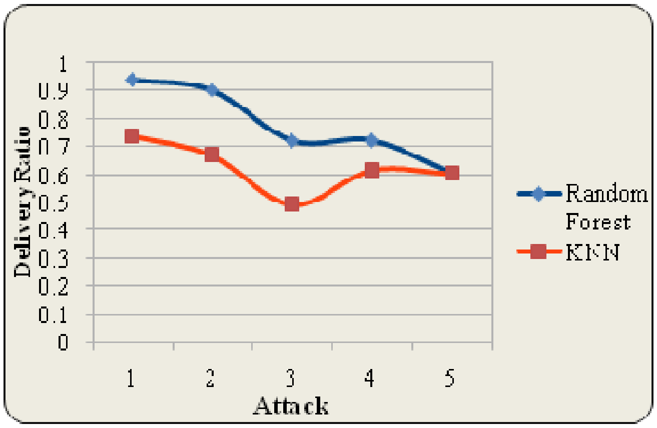


Fig. 5. Comparative Delivery Ratio Analysis of RFS-SSO and KNN.

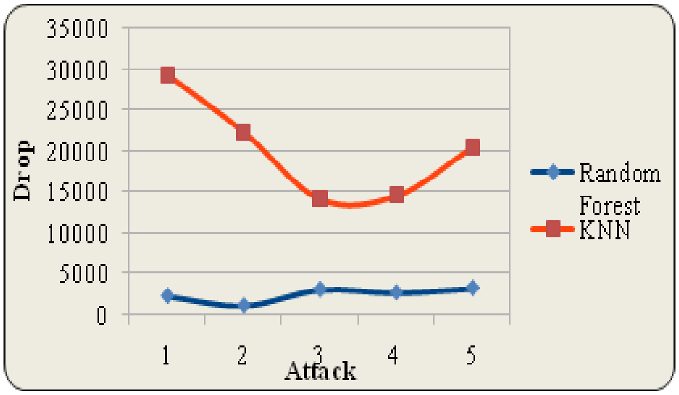


Fig. 6. Comparative Drop Analysis of RFS-SSO and KNN

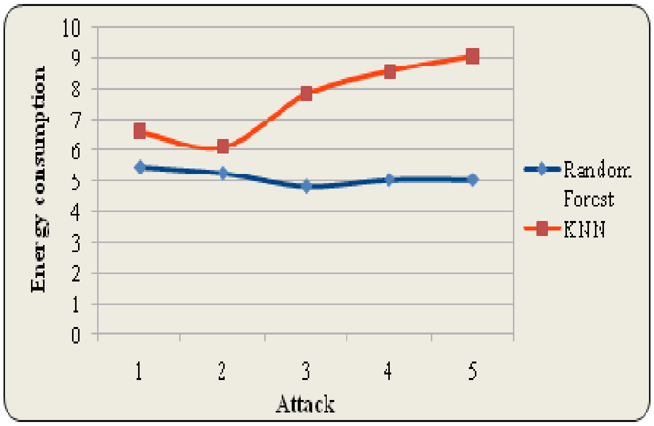


Fig. 7. Comparative Energy Consumption Analysis of RFS-SSO and KNN.

Overhead: The sum of extra data delivered during the communication activities on the network is called the overhead. Fig. 8 demonstrates the overhead during attacks on the existing IDS network and proposed IDS systems. To visualize the performance the proposed method provides using the blue line and the performance of the traditional technique provides using the red line. It depicts that, attack does not affect the overhead output in the suggested approach.

In Fig. 8 the overhead analysis of our proposed approach with the existing KNN classifier is shown. In which for the proposed IDS-RF-SSO the overhead is maximum but for KNN it is minimal. Maximum overhead value of the proposed approach is 8458, 6936, 9696, 6754, 7265 whereas the existing approaches values are 59639, 53514, 58401, 45968, and 25433 respectively.

Throughput: Successful message delivery over a communication medium of usual rate is known as Throughput. In terms of data packets per time slot or data packets per second is calculated by throughput. Compare the performance of the proposed and traditional technique. It represents red line for traditional IDS with KNN Classifier and for proposed green line is used. During IDS attacks according to the observation of results the throughput is increased for the proposed approach significantly and decreased for the existing approach. Therefore from attack the performance of the proposed IDS is not affected.

The analysis of our proposed approach with the existing KNN classifier is shown in Fig. 9. In which the throughput is optimum for the proposed IDS-RF-SSO but minimal for KNN. The proposed solution has a maximum throughput value of 31739, 24984, 24056, 19487 and 17524, while the existing approach values are 21952, 17855, 14423, 14152 and 8038 respectively.

Finally, the paper examined the quality of the full KDD dataset statistically and suggested a new methodology to deal with the challenges. Hybrid IDS with RF-SSO approach can effectively reduce the problem of complexity and multiclass dataset relative to other current algorithms. Pre-processing can easily extract and record as normal or attack the most relevant feature sub-set form network traffic. It is clear that the suggested model eliminates ordinary documentation and reduces the list of attributes, thereby increasing the IDS strain of dealing with a wide set of features. Swarm awareness strategies combined with RF Classifier can therefore effectively increase reliability of identification and deliver optimal solutions. Through finding an optimal solution in the pre-processing system, detection of invasion is rendered more reliable and inaccurate detection of attacks is minimized. The analytical result shows that perhaps the integration of the Hybrid RF-SSO algorithm is quicker and more efficient in solution.

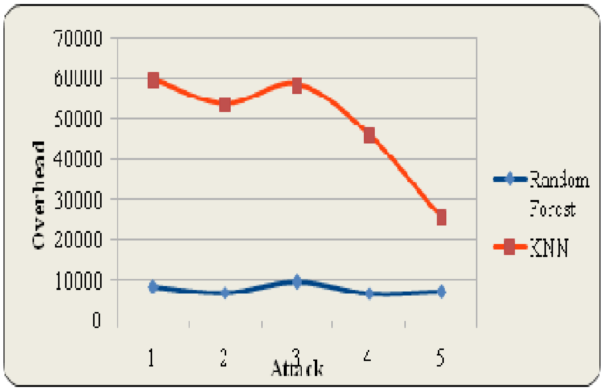


Fig. 8. Comparative Overhead Analysis of RFS-SSO and KNN.

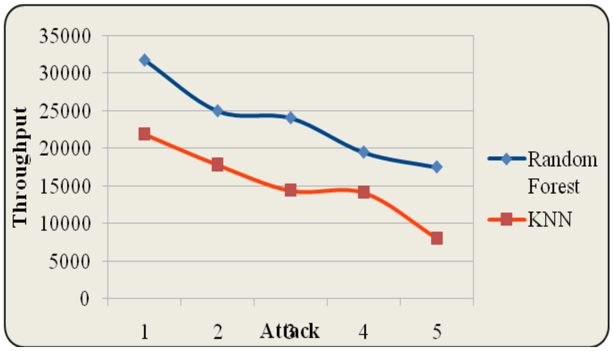


Fig. 9. Comparative Throughput Analysis of RFS-SSO and KNN.

VI. Conclusion

The contributions of this research are indeed the proposal for a hybrid aesthetic system for effective intrusion detection for service provider by utilizing the classification and optimization algorithms to enhance intrusion detection system performance. The reliability of the hybrid invasion detection system was measured in terms of delay, delivery ratio, drop overhead, energy consumption and throughput. To testing an effective hybrid intrusion detection system for SDWSNs in this research work the KDD CUP 1999 Dataset being utilized to test the proposed hybrid IDS. The system was designed on the basis of a combination of Knowledge and Behavior based IDS with RF as classifier and SSO approaches. The experimental study conducted on NSL-KDD dataset found our methodology significantly increased overall system efficiency when relative to the system performance with KNN classifier based IDS system.

References

[1] G. V. Nadiammai, S. Krishnaveni and M. Hemalatha, A Comprehensive Analysis and Study in IDS Using Data Mining Techniques, IJCA, vol. 35, pp. 51–56, November–December (2011).

[2] Arif Jamal Malik, Waseem Shahzad and Farrukh Aslam Khan, Network Intrusion Detection Using Hybrid Binary PSO and Random Forests Algorithm, Security and Communication Networks, (2012).

[3] P. Natesan and P. Balasubramanie, Multi Stage Filter Using Enhanced Adaboost for Network IDS, International Journal of Network Security and its Applications, vol. 4, no. 3, (2012).

[4] Mrutyunjaya Panda, Ajith Abraham and Manas Ranjan Patra, A Hybrid Intelligent Approach for Network Intrusion Detection, UCCTSD, pp. 1–9, (2012).

[5] Md. Al Mehedi Hasan, Mohammed Nasser, Biprodip and Shamim Ahmad, Support Vector Machine and Random Forest Modeling for IDS, JILSA, pp. 45–52, (2014).

[6] Ujwala Ravale, Nilesh Marathe and Puja Padiya, Feature Selection Based Hybrid Anomaly Intrusion Detection System Using K Means and RBF Kernel Function, ICACTA, pp. 428–435, (2015).

[7] Aleksandar Lazarevic, Vipin Kumar and Jaideep Srivastava, Intrusion Detection: An Survey, p. 31.

[8] Araujo, Oliviera, Shinoda and Bhargava, Identifying Important Characteristics in the KDD99 Intrusion Detection Dataset by Feature Selection Using a Hybrid Approach, International Conference on Telecommunications, (2010).

[9] Nanak Chand, Preeti Mishra, C. Rama Krishna, Emmanuel Shubhakar Pilli, and Mahesh Chandra Govil, "A comparativeanalysis of SVM and its stacking with other classification algorithm for intrusion detection", In International Conferenceon Advances in Computing, Communication, & Automation, 1–6,2016.

[10] Gianluigi Folino and Pietro Sabatino. 2016. Ensemble based collaborative and distributed intrusion detection systems: A survey. 66 (May 2016), 1–16.

[11] Nabila Farnaaz and M. A. Jabbar, “Random forest modeling for network intrusion detection system”, Procedia Computer Science 89 (2016), 213–217,2016.

[12] N. Cleetus and K. A. Dhanya, “Multi-objective functions in particle swarm optimization for intrusion detection”, In 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI’14). 387–392,2014.

[13] Adel Sabry Eesa, Zeynep Orman, and Adnan Mohsin Abdulazeez Brifcani,”A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems”,2015.

[14] Susan M. Bridges and Rayford B. Vaughn, “Fuzzy data mining and genetic algorithms applied to intrusion detection”, In National Information Systems Security Conference (NISSC’00), 16–19, 2000.

[15] Anna L. Buczak and Erhan Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection”, IEEE Communications Surveys & Tutorials 18, 2 (2016), 1153–1176, 2016.

[16] I. Butun, S. D. Morgera, and R. Sankar, “A survey of intrusion detection systems in wireless sensor networks”, IEEE Communications Surveys Tutorials 16, 1 (2014), 266–282, 2014.

[17] S. Dey, Q. Ye and S. Sampalli, "A machine learning based intrusion detection scheme for data fusion in mobile clouds involving heterogeneous client networks", Information Fusion, vol. 49, pp. 205-215, 2019.

[18] H. Peng, Z. Sun, X. Zhao, S. Tan and Z. Sun, "A Detection Method for Anomaly Flow in Software Defined Network", IEEE Access, vol. 6, pp. 27809-27817, 2018.

[19] D. Jianjian, T. Yang and Y. Feiyue, "A Novel Intrusion Detection System based on IABRBFSVM for Wireless Sensor Networks", Procedia Computer Science, vol. 131, pp. 1113-1121, 2018.

[20] D. Santoro, G. Escudero-Andreu, K. Kyriakopoulos, F. Aparicio-Navarro, D. Parish and M. Vadursi, "A hybrid intrusion detection system for virtual jamming attacks on wireless networks", Measurement, vol. 109, pp. 79-87, 2017.

[21] Lin, Kai, et al. "Node scheduling for all-directional intrusion detection in SDR-based 3D WSNs." IEEE Sensors Journal 16.20 (2016): 7332-7341.

[22] S. Shamshirband et al., "Co-FAIS: Cooperative fuzzy artificial immune system for detecting intrusion in wireless sensor networks", Journal of Network and Computer Applications, vol. 42, pp. 102-117, 2014.

[23] Paul, Angshuman, et al. "Improved random forest for classification." IEEE Transactions on Image Processing 27.8 (2018): 4012-4024.

[24] Mirjalili, Seyedali, et al. "Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems." Advances in Engineering Software 114 (2017): 163-

[25] Atiku Abubakar and Bernardi Pranggono, "Machine Learning Based Intrusion Detection System for Software Defined Networks",2017 Seventh International Conference on Emerging Security Technologies (EST), pp.138-143,2017.

[26] Bosh, A., Zisserman, A., Munoz, and X.: “Image classification using Random Forests and ferns”. In: IEEE ICCV2007.

[27] Indira K, Christal Joy E, "Energy Efficient IDS for Cluster-Based VANETS", Asian Journal of Information Technology, vol 14(1) ,2015, 37-41

[28] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[29] Indira K, Christal Joy E, "Prevention of Spammers and Promoters in Video Social Networks using SVM-KNN", International Journal of Engineering and Technology, Vol 6, No.5, Oct – Nov 2014, Pg 2024-2030.

[30] Imran Memon, Ibrar Hussain, Rizwan Akthar, Gencai Chen, “Enhanced privacy and authentication: An efficient and secure anonymous communication for location based service using asymmetric cryptography scheme”, wireless personal communication, 2015.

[31] Abirami Devaraj, Karunya Rathan, Sarvepalli Jaahnavi, K Indira, “Identification of Plant Disease using Image Processing Technique”, International Conference on Communication and Signal Processing (ICCSP) 2019.

[32] K .Indira, U.Sakthi, “Security issues, countermeasures and dynamic scheduling for SDWSN”, 2nd International Conference on signal processing and communication (ICSPC), 2019.

[33] K. Indira, P. Ajitha, V.Reshma, A.Tamizhselvi, “An efficient secured routing protocol for Software Defined Internet of Vehicles”, International Conference on Computational Intelligence in Data Science (ICCIDS), 2019.

The Internet of Things for Crowd Panic Detection

Habib Ullah[1], Ahmed B. Altamimi[2], Rabie A. Ramadan[3]
College of Computer Science and Engineering
University of Ha'il, Ha'il
Saudi Arabia

***Abstract*—Crowd behavior detection is important for the smart cities applications such as people gathering for different events. However, it is a challenging problem due to the internal states of the crowd itself and the surrounding environment. This paper proposes a novel crowd behavior detection framework based on a number of parameters. We first exploit a computer vision approach based on scale invariant feature transform (SIFT) to classify the crowd behavior either into panic or normalness. We then consider a number of other parameters from the surroundings namely crowd coherency, social interaction, motion information, randomness in crowd speed, internal chaos level, crowd condition, crowd temporal history, and crowd vibration status along with time stamp. Subsequently, these parameters are fed to deep learning model during training stage and the behavior of the crowd is detected during the testing stage. The experimental results show that proposed method renders significant performance in terms of crowd behavior detection.**

Keywords*—VANET; smart cities; crowd behavior; deep learning; Recurrent Neural Networks (RNN); Convolution Neural Networks (CNN); Invariant Feature Transform (SIFT)

I. Introduction

Smart cities are intended to ease the automated services for people inside their living areas. Recently, smart city services have been the subject of many research articles [1-4]. For example, one of the services reported in [2] is measuring the health of historical building where maintaining such building is challenge since it requires a frequent visits by administrators and technical staff to check their functionalities. Sensors are used to measure the building characteristics including humidity, temperature, and vibration. Such automated techniques for providing services that might require huge manpower and time invigorated the concept of smart cities. One more service towards smart cities is the waste management where it is a major issue for many cities. Sensors could be placed in garbage cans and when they are full, the system sends the garbage according to the shortest possible route. However, the most important and integral part of smart cities is crowd management with the increase of population. Different social events take place in different public places. To detect abnormal behaviors in crowd is very important for the concept of smart cities as a whole.

Crowd monitoring for behavior detection is an important application for smart cities where the system detects abnormalities during crowd gatherings. This helps to identify any problem source and enable the authority to monitor it. Other well investigated services in smart city traffic monitoring include handling traffic congestion, smart signals, and vehicles routing. Cameras and sensors could be sources to track congested area in crowd and report it back to the authority with less effort compared to the traditional methods where police officers are asked to report any accident of congested crowded area.

For crowd analysis, smart city services may require the interaction between many elements on the roads, homes, and government buildings, and other infrastructures. Regardless of the provided services in the smart cities, effective mechanisms are required to route the collected information from one place to another; other competent mechanisms are mandatory to manage many of the heterogeneous devices as well as many of the newly produced sensors. This raises the concept of internet of things (IoT) and its role towards the implementation of smart cities for crowd analysis to detect abnormal behaviors.

IoT is a network that accomplishes the seamless integration among sub networks such as sensor networks for public places, vehicular networks, and mobile networks. The interaction among these networks allows ease of data exchange between their elements. For example, a mobile node can exchange data with a vehicle in seamless manner as they are on the same network. However, for a seamless integration among the IoT networks, an elegant model is compulsory. A few models are proposed in the literature including the ones in [5-8]. However, most of the existing models were concern about certain networks. For instance, references [9-11] were focusing on RFID connectivity to IoT environment. Other models are proposed in [12] and [13] for connecting mobile, vehicular and sensor networks in the IoT environment. In both articles, a gateway including access point is proposed as a mean to integrate these networks together. In addition, authors of [14] reported that IoT gateway deployment as a message ferry outperforms placing the gateway based on petro graphical area.

One of the convoluted networks in smart cities applications in general is the Vehicular Ad Hoc Network (VANET) [15]. VANET consists of three main components which are On Board Unit (OBU), Application Unit (AU), and RoadSide Unit (RSU). OBU is a device that is installed in a public place to pass the collected data from sensors or applications in public places to the other OBU in other places or to RSU. AU is a device resides in a public or crowd place that can be utilized for the crowd surrounding information. The final component is RSU, where it is mainly installed along the roads, intersection areas, or crowd public places. The main role is to collect the information that is passed from the OBU to a control center. Based on the transferred data via RSU, the authority in the control center can come up with any crowd related decision.

Research deanship at the University of Ha'il, Saudi Arabia.

Based on the previous VANET facilities and new sensor capabilities, VANET could be one of the main players in IoT that could be exploited in many IoT services including the detection of abnormal crowd behavior in smart cities. Smart cars can identify the crowd congestion in order to facilitate rerouting in the smart city and it can also help to lower the occurrence of abnormal crowd behaviors. One of the challenges in VANET that contributes to the IoT is the detection of crowd behavior as a whole. Identifying crowd behavior could be used as early indicator of a dangerous situation that might lead to chaos in public places. In order to identify crowd behavior, sensors in public places and other units could be of help reporting different features/parameters used to do so. This paper investigates the detection of crowd behavior identification. The problem is not new; however, the previous work focuses on small set of parameters and use traditional methods for the same purpose.

Next section of the paper introduces the literature review to crowd behavior detection; proposed method is presented in Section III; results are elaborated in Section IV; discussion is presented in Section V and Section VI presents the conclusion.

II. Literature Review

Crowd behaviors have been significantly investigated in the literature including [15-22]. The importance of such investigations comes from the impact of such behavior on the people safety. Throughout this section, we summarize the state-of-the-art of crowd behaviors in some of the recent literatures. Table I presents the work done in this field recently. The goal of the studies has been classified in different categories. As can be seen, the table lists the used features, identification method, and the behaviors. The target of research done in [23-24] is to identify a particular crowd considering group of participants based on some features, while in [25-27], the main target is to identify the behavior of the crowd whether the crowd is normal or aggressive. Sometimes, the moderate style is used in exchange with normal style. Identification methodologies used are one of three categories which are neural networks [28-29], classification models [30-31], and statistical models [32-33].

The number of input parameters reported also in the Table I ranges from two to four. We argue that with increasing the number of input parameters to the identification method capturing all of the realistic information would increase the accuracy of crowd behavior detection. In fact, this is the main motivation behind the wok in this paper. Therefore, in our proposal, different parameters are used as input to the identification method namely crowd feature analysis, crowd coherency, social interaction, motion information, randomness in crowd speed, internal chaos level, crowd condition, crowd temporal history, and crowd vibration status along with time stamp. These inputs will be applied to deep learning identification method for crowd behavior identification.

TABLE. I. List of Published Work in the Field of Crowd behavior Identification

Ref	Features / Parameters	Identification Method	Behavior
[15]	- Crowd Speed - Exit Opening	Bayesian Probability	- Normal - Aggressive
[16]	- Speed and velocity - Average motion / Deceleration Profile - Turning speed vs. Radius of turn	support vector machine (SVM)	- Identify general behavior
[17]	- acceleration, - the speed, - Social interaction	neural network	- Crowd Scene analysis
[18]	- Chaotic behavior	Probabilistic ARX model	- Density estimation
[23]	- acceleration - randomness - the crowd speed - turning behavior	Decision Tree Clustering algorithm	- Identify coherent groups
[24]	- acceleration relative lane position	Statistical method of a Gaussian mixture model (GMM)	- Identify abnormal behavior
[25]	- Acceleration - Speed - position	fuzzy clustering algorithm	- Event planning
[26]	- position - velocity - Coherency	neural networks	- Crowd dynamics
[27]	- Feature integration	Bayesian network	- Crowd trajectories

III. Proposed Method

This section elaborates on the proposed VANET framework based on IoT architecture [15][34] for crowd analysis. The framework shows how VANET applications and protocols could perfectly fit the IoT architecture for crowd analysis. Fig. 1 shows the layered framework and the name of each layer. It basically consists of seven layers where the bottom layer is the data collection layer. Sensors, odometers, GPS information, etc. are sources of information in this layer. The second layer of the framework is the connectivity layer where different sensors might connect to each other. These sensors need to connect to RSU and RSU might need to connect to the data center. Therefore, all of the connectivity issues could be handled in this layer. This allows the collected information to be exchanged and forwarded to other components in the network until it reaches the data center helping out in decision making process. The collected raw data might be repeated, unsuitable, noisy, and unformatted. Therefore, layer 3 is responsible for data transformation and cleanness. Layer 4 is responsible for data analysis and storage since it is expected to have huge information increasing exponentially with time. This huge collected data needs to be aggregated to save the network bandwidth and removing the data redundancy. Many of the effective aggregation as well as fusion methods could be used in layer 5. Layer 6 is the application layer which involves many of the crowd applications as well as related data center applications. The decision making and process collaboration with other components of the system could be handled in layer 7.

Beside the vertical structure, edge software can be source for multiple layers as needed for crowd anomaly detection. For example, security aspects cannot be dedicated to one layer only. Particularly, security is needed at devices level; so they do not get hacked at the hardware level. It is also needed at application layer; so the software cannot be hacked as well. Another example is the scheduling for devices communication. It can be an important factor at devices level, so it can determine when to sense the data. It is also important at data transformation level; so it can be decided when to schedule the data exchange for crowd monitoring. This allows us to see the model, Fig. 1, from two sides, vertical and horizontal shape. The vertical one has a pre-determined function for each layer and the horizontal one can be an implementation for cross layer functions.

Table II briefly compare the VANET model to the IoT models proposed in [15][34]. As can be seen in the Table II, the proposed model perfectly fit the IoT model. For our work related to crowd behavior identification problem, this model can be extended to encode features and connectivity across multiple nodes in the network.

Nowadays, crowd behavior detection is the main focus of industry and significant progress has been made in this regard. In addition, the concept of IoT enforces many of the automation to take place. However, the behavior of crowd itself is not taking that much attention although it is very important for people safety and for the surrounding environment such as pedestrians. However, crowd behavior identification depends on many of the factors. Some of these factors were taking into consideration in the previous works but due to the complexity of the problem, some others were ignored.

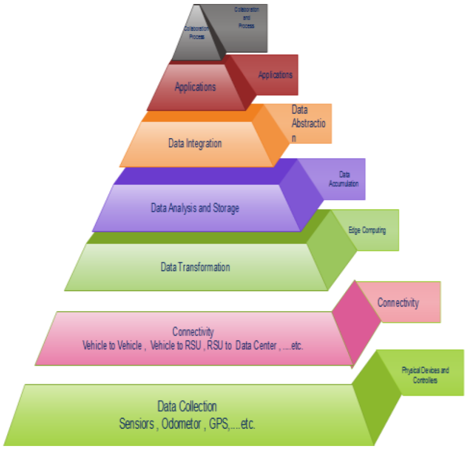

Collaboration Process
Collaboration and Process
Applications
Applications
Data Integration
Data Abstraction
Data Analysis and Storage
Data Accumulation
Data Transformation
Edge Computing
Connectivity
Vehicle to Vehicle , Vehicle to RSU , RSU to Data Center , ...etc.
Connectivity
Data Collection
Sensors , Odometor , GPS,....etc.
Physical Devices and Controllers

Fig. 1. VANET Framework based IoT Architecture.

TABLE. II. IoT LAYERS AND THEIR FUNCTIONS

Layer No.	Layer name	Function	Elements in VANET
1	Physical Device and Controller	Collect and sense the data from the surrounding environment	Sensors, odometer
2	Connectivity	Setup the connectivity between the network elements.	Zig bee, Bluetooth
3	Edge Computing	Exchange the collected data via the connected components	Public place unit, Sensor
4	Data Accumulation	Store the collected data in a dedicate data center after it travel though the connected devices	Data center
5	Data Abstraction	Abstract the stored data to ease the arrival of a decision in later layer	Artificial intelligent software
6	Application	Report the abstract to the user via application	Any mobile application
7	Collaboration and Process	The needed action based on the collected data can be processed in this stage.	People, business, sign in the places.

Here, we tend to take different parameters as input for accurate decision making about the crowd behavior including crowd coherency, social interaction, motion information, randomness in crowd speed, internal chaos level, crowd condition, crowd temporal history, and crowd vibration status along with time stamp. These parameters are easy to be captured by sensors in the data collection layer. In addition, the crowd behavior decision could be one of two behaviors namely aggressive and normal. Here, we extend the definition of aggressive crowd that is driven by abnormal motion patterns. Based on these concepts, investigating a well-designed solution to classify crowd behavior is a challenging problem. Part of the challenge is to capture data such as computer vision based features out of a video stream. In fact, this parameter requires accurate image processing algorithm as well as it might take long processing time. In the next section, we will explore the details of the proposed methodologies.

Following the VANET model presented in Fig. 1, the crowd behavior identification could be handled based on the same layers. For instance, the raw data is collected through the data collection layer, then transformed to a suitable format in the transformation layer followed by analysis and storage in the next layer, decisions have to aggregated in the aggregation layer to be passed to the sensor applications layer. Finally, the data is shared using the collaboration process and through the connectivity layer.

For the paper to be self-content, this section starts by brief description to the used methodologies. The following subsection explains the SIFT algorithm. The next subsection briefly explain two of the deep learning neural networks namely Recurrent Neural Networks (RNN) and Convolution Neural Networks (CNN).

To detect crowd behavior, we compute scale invariant feature transform (SIFT) [35][36] from each video frame of the crowd video. SIFT based frame matching is a foundation step of many challenging issues in the field of computer vision, including object and situation identification. Besides that, SIFT features have several important aspects that make them significant for associating various frames of the crowd behavior. It is worth noticing that the SIFT features do not change by changing other factors including scale, rotation, and illumination. In addition to that, these features are highly unique, which allows a single feature to be correctly associated with the relevant and matching feature in the consecutive video frame. SIFT features computation from a video frame occurs in four steps which are scale space extrema detection, keypoint localization, orientation assignment, and keypoint descriptor.

Scale-space kernel is mainly the Gaussian function. Therefore, the scale space of a frame of a video is the function L(x, y, σ) that it is generated from the convolution of a variable-scale Gaussian, G(x, y, σ), with the frame, I(x, y) as given in equations (1) and (2).

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \tag{1}$$

$$G(x,y,) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{2}$$

The scale-space extrema in the difference-of-Gaussian function is convolved with the video frame, D(x, y, σ), that is calculated from the difference of two nearby scales separated by a constant multiplicative factor k as provided in equation (3).

$$\begin{aligned} D(x,y,\sigma) &= (G(x,y,k\sigma) - G(x,y,\sigma)) * I(x,y) \\ &= L(x,y,k\sigma) - L(x,y,\sigma) \end{aligned} \tag{3}$$

To find the local maxima and minima in the difference frame D(x, y, σ), each feature point is compared to its eight neighbors in the same frame. It is detected and identified only if it is larger than all of these neighbors or smaller than all of them. The computation of this operation is not heavy because most similar features will be identified after the first few operations. The important characteristic is to consider the iteration of sampling in the video frame and scale domains that is needed to effectively identify the extrema. To this end, there is no minimum spacing of features that will choose all the extrema because they can be in near vicinity with high probability. Therefore, we must investigate a formulation that trades off efficiency and completeness. In is important to note that the extrema that are in near vicinity are susceptible to small noise of the video frame.

Additionally, the total number of detected features increases with increased sampling of scales and the total number of correct matches also increases. The feature matching with high probability depends substantially on the amount of accurately matched feature points. Hence, it is important to exploit a larger number of scale samples and features. Nevertheless, the load of calculations also increases with this criteria. To this end, the scale space difference of Gaussian function has a large number of extrema and we can identify the most reliable and effective subset even with a coarse sampling of scales.

Preprocessing the video frame before extrema detection significantly ignores the high magnitude frequencies in spatial domain. Hence, the streaming crowd video frame can be magnified to engender more features than were present in the original video frame. Therefore, the size of the crowd video frame is significantly increased by utilizing linear interpolation technique before establishing the first level of the pyramid. The other operations could have been carried out by considering various sets of subpixel-offset kernels on the original video frame. In fact, increasing the size of the video frame presents improved implementation.

Once a keypoint sample is detected by comparing a pixel to its neighbors, the next step is to carry out a matching process in the neighborhood region. In this way, the technique reject points that have low contrast or improperly localized along an edge. For reliability, ignoring keypoints with low contrast does not suffice. The difference-of-Gaussian procedure will present huge magnitude on edges. Therefore, the process is sensitive to small amount of noises in the neighborhood regions. An improperly detected peak in the difference of Gaussian function will have a large principal curvature which can be calculated from a 2 x 2 Hessian matrix computed at the location and scale of the keypoint as given below:

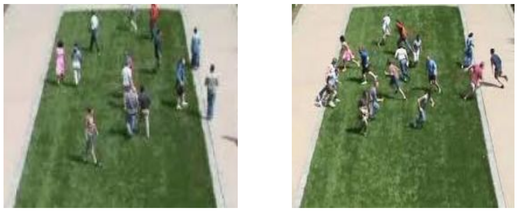
$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{4}$$

The magnitude H is formulated by subtracting neighboring features or keypoints. For each video frame, L(x, y) is calculated using pixel differences as given in equations (5) and (6).

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{5}$$

$$\theta(x,y) = \tan^{-1}((L(x,y+1) - L(x,y-1)) / (L(x+1) - L(x-1,y))) \tag{6}$$

A direction histogram is engendered from the gradient calculations of features within an area around the detected feature in the center. Once the keypoints on the consecutive frames are calculated, we determine the distances among the corresponding keypoints. If the distances of the corresponding keypoints are in a specific range and there are a set of such keypoints, the frame is considered as angry. Otherwise it shows the normal behavior of the driver. In Fig. 2, two frames are presented showing normal and panic crowd behavior from benchmark UMN dataset [37]. Our approach correctly detected the abnormal emotion of the crowd. If there a less than 10 SIFT features matching on the consecutive frames, we consider it panic behavior.



Normal behavior (b) Panic behavior

Fig. 2. Crowd behavior Analysis from UMN Dataset [37].

After recognizing partially the behavior of the crowd, we feed it as a parameter along with other parameters to Deep Learning model. It is worth noticing that Deep learning is the novel and significant trend in machine learning. It pledges general, powerful, and fast machine learning, moving us one step closer to artificial intelligence. In the Deep learning, input is passed through several non-linearities before being output. In comparison to Deep learning, traditional learning algorithms e.g., decision trees and SVMs and Naive Bayes are shallow. Deep learning excels on challenging problems where the inputs are not a few parameters but instead are parameter values in the form of images. In addition to ability of the deep learning to handle nonlinear data, deep networks also have other capabilities which set them apart from other traditional machine learning models. For example, we can modify them in different ways to fit to our problem.

The earlier important progress in Deep Learning was Deep Belief Networks [38-39] to pretrain deep networks. This method investigated that pretraining each layer is better for initial weights instead of random initialization. For example, Deep Belief Networks based on Restricted Boltzmann Machines [40], and Deep Autoencoders based on Autoencoders [41].

Autoencoders is driving more interest since it could be used as a way to pretrain neural networks. In fact, training neural networks is very difficult due to the reasons that the magnitudes of gradients in the lower layers and in higher layers are different. Moreover, the curvature of the objective function is difficult for stochastic gradient descent to locate the local optimum. Also Deep networks are multiple parameters oriented. That means they can remember training data but do not generalize well. The pretraining process cope with the aforementioned problems. In the pretraining step, a sequence of shallow autoencoders is trained layer by layer. The last layer is trained exploiting the supervised data. Finally, backpropagation is used to fine-tune the entire network using the supervised data. The significant progress in Deep Learning is the use of convolutional neural networks to obtain a paramount improvement. Actually the unsupervised learning in the autoencoders is less relevant when a lot of labeled data are available. In the convolutional neural networks, a technique called locality structures reduces the number of connections. Also weight sharing is used to reduce the number of connections. A few parameters in the model are constrained to be equal to each other in the weight sharing. The convolution networks also come with another layer known as the max-pooling layer. In this layer the max value of a selected set of output neurons is calculated from the convolutional layer and it is used as input to higher layers. In the convolution neural networks, explicit computation is performed with all the weights in the forward pass. In the backward pass, the gradient for all the weights is computed based on equation (7).

$$\frac{\partial J}{\partial w_1}, \ldots\ldots, \frac{\partial J}{\partial w_g} \tag{7}$$

The average of the gradients from the shared weights is used to update the weights as given in equations in (8).

$$\begin{aligned} w_1 &= w_1 - \alpha\left(\frac{\partial J}{\partial w_1} + \frac{\partial J}{\partial w_4} + \frac{\partial J}{\partial w_7}\right), \\ w_4 &= w_4 - \alpha\left(\frac{\partial J}{\partial w_1} + \frac{\partial J}{\partial w_4} + \frac{\partial J}{\partial w_7}\right), \\ w_7 &= w_7 - \alpha\left(\frac{\partial J}{\partial w_1} + \frac{\partial J}{\partial w_4} + \frac{\partial J}{\partial w_7}\right), \end{aligned} \tag{8}$$

For the max-pooling layer, in the forward pass, the layer that renders the max value is remembered, so that in the backward pass, the gradient for that layer is computed only. In Fig. 3 below, the diagram of the convolutional neural network is depicted that shows the layered architecture.

The input to the CNN may have multiple channels. Therefore, the CNN architecture can be modified to work with such input. In this modification, it is important to have a filter that processes multiple channels input. However, the weights are not shared across different channel. Each set of output produced by each filter is called a map. It is also possible to have a CNN architecture with multiple maps. To take this into account, the CNN can be modified to have multiple filters per location. In fact, the input to the CNN is generally two-dimensional. The CNN architecture cope with two-dimensional inputs. For this purpose, each filter has two dimensions. In case the inputs have many channels, then each filter is essentially three dimensional: row-column-channel.

It is also worth to notice that the CNN architecture considers input with fixed size. However, inputs may have many sizes. To deal with that, it is typical to crop the inputs at the center and convert all inputs to the desired size. In fact, many state-of-the-art CNN are sequences of processing blocks where each block is a combination of convolution, max pooling, and local contrast normalization. The variable size inputs are challenging in some cases. For example, predicting the stock market of any company. One possible solution to deal with the variable-sized input problem is to use convolutional neural network where the max pooling is applied to all of the output of the filters. However, it does not fix the problem. If the input is variable-sized, the output is fix-sized. The problem with this situation is that the CNN is invariant to translation. Considering such a large max-pooling, losing position information is inevitable. In the field of image processing and computer vision, translation invariance is acceptable since the output of a system should be invariant to translation. However, in the stock prediction model, this is an unwanted property, because we want to make use of the precise temporal information. The solution to deal with variable-sized inputs is to use a Recurrent Neural Network (RNN).

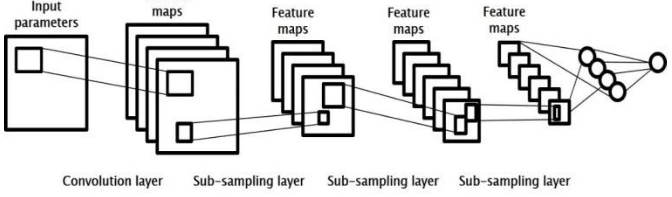


Fig. 3. Coding Process.

In the RNN, there are usually three sets of parameters: the input to hidden weights (W), the hidden to hidden weights (U), and the hidden to label weight (V). To this end, all the W's are shared, all the U's are shared and all the V 's are shared. It is due to this sharing property, that the RNN is suitable for variable-sized inputs. Considering these notations, the hidden states are iteratively calculated as shown in equation (9).

$$f(x) = Vh_T$$
$$h_t = \sigma(Uh_{t-1} + W_{x_t}),$$
$$.....$$
$$h_O = \sigma(Wx_O) \qquad (9)$$

The cost function can be minimized then to get the appropriate weights. To calculate the gradient of the RNN, the backpropagation algorithm can be used which is called Backpropagation through time (BPTT). While computing the gradient for the RNN, it could be either large or very small. Therefore, the entire training process will converge slowly. To improve the speed of the training process, it is suggested to truncate the gradient at certain values.

Based on the proposed parameters for crowd behavior identification, we believe that a deep learning technique is suitable to analyze the crowd behaviors. However, deep learning has been widely used in many of the image classifications as well as audio and showed a successful results over different datasets. It is also designed in such way to accommodate certain type of inputs. Therefore, our solution in this paper works in stages, starting by the collected data as shown in Fig. 4.

The input of the designed system is real crowd videos along with other parameters coming from sensors. Videos are framed and the status of the crowd is captured at each second. However, this raw information has to be preprocessed to have equal weights and meaningful information to the deep leaning. Part of the preprocessing is to have weights for each type of parameter. Some other statistics are applied on the collected data for each parameter such as Max, Min, Mean, Median, Variance, and Standard Deviation.

These statistical information expresses the changes in the frame information. Since the collected data will be huge to be used with the deep learning, the crowd behavior can be captured through multiple frames where the collected frames are divided into overlapped segments. Each segment is divided into a number of slices. Again, these slices are overlapped to avoid losing any information during processing. These slices are also considered as deep learning widows.

Another step before providing the input to the selected deep learning model is the fitting and normalization over the training data. The fitting is done by collecting the statistics (mean/stdev) from the training data. Then the testing dataset is also normalized using the statistics calculated from the training dataset. The paper utilizes two different deep learning techniques and their variants which are Convolutional neural networks (CNN) [42] and Recurrent neural networks (RNN) [43]. The dataset is adapted to work with both models during the processing phase. For instance, RNN is better working with sequence learning, the input is formatted accordingly.

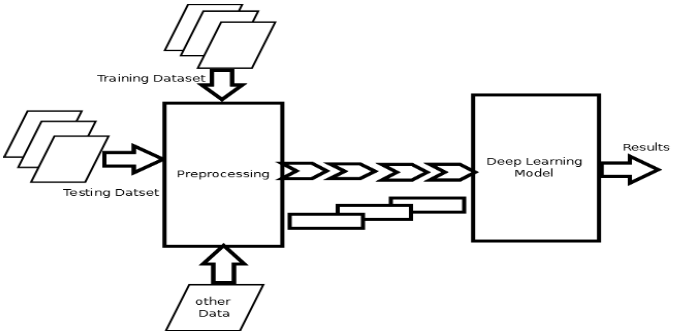


Fig. 4. Solution Process.

IV. Results

This section shows some of our test cases as a proof of concept to the utilization of deep learning with this large number of input parameters. First, in order to have all of the suggested parameters in hand, we consider two popular benchmark datasets which are called UMN [37] and UCSD [44] for panic detection and non-pedestrian entities detection in crowded scenes. Our criteria in measuring the performance of our approach are accuracy, precision, recall, and F1 score. F1 score could be computed by equation (10):

$$F_1 = 2.\frac{1}{\frac{1}{Recall}+\frac{1}{Precsion}} = 2.\frac{Precsion\,.Recall}{Precsion+\,Recall} \qquad (10)$$

We divide the video from each dataset into segments and slices. A frame consists of multiple parameters and statistical values. These data are preprocessed to be normalized and fitted according to the requirements of the deep learning model. The training dataset is 70% of the overall dataset and 30% is the testing dataset. For performance measure Conventional Neural Networks (CNN) with/without pooling layers, Recurrent neural networks (RNN), Pre-trained Recurrent neural networks (PretrainRNN) and Stacked Recurrent neural networks (StackedRNN) are examined. For RNN and StackedRNN, the input data is formatted to have one frame per row. Therefore, each row will be having different values ending with the correct label for training data. The frame information within a segment could be averaged to work on segments instead of frames. However, our experiments are done on frames instead for more accurate results. Rectified linear unit (relu) as an activation function and (softmax) are appended to the recurrent layer. Two layer StackedRNN (100 neuron each) are developed for the testing purpose as well. For CNN, the input format is different in terms of number of rows and columns. The number of rows in this case is 48 +8 row representing the main parameters and the generated statistics. However, the number of columns (frames per segment) is 128. Again, 1000 iterations are set Softmax last layer. CNN network is designed with six layers; three are convolution pooling and the other three are fully connected layers.

Table III shows the results collected for UMN dataset [37] from running the different methods using the training and test parts of the same dataset. We also extract other parameters from the dataset including crowd coherency, social interaction, motion information, randomness in crowd speed, internal chaos level, crowd condition, crowd temporal history, and crowd vibration status along with time stamp. These parameters are extracted manually with the help of 10 participants. As can be seen from the Table III, CNN with pooling layers overall performance is better than CNN without pooling layers. At the same time, RNN results, in general, is much better than CNN. However, the StackedRNN seems to be the best in terms of Accuracy, Precision, Recall, and F1 score. It was able to reach 76% accuracy. Based on our observations to the results, although the current dataset is representative but we believe that StackedRNN could perform more accurate with larger datasets.

Table IV shows the results collected for UCSD dataset [44] from running the different methods using the training and test parts of the same dataset. Again, we extracted other parameters from this dataset including crowd coherency, social interaction, motion information, randomness in crowd speed, internal chaos level, crowd condition, crowd temporal history, and crowd vibration status along with time stamp. As can be seen again from the Table IV, CNN with pooling layers overall performance is better than CNN without pooling layers. At the same time, RNN results, in general, is much better than CNN. However, the StackedRNN seems to be the best in terms of Accuracy, Precision, Recall, and F1 score. It was able to reach 75% accuracy. Based on our observations to the results, although the current dataset is representative but we believe that StackedRNN could perform more accurate with larger datasets.

We also present some results in comparison with other reference methods including spatio-temporal anomaly model (STA) [45], abnormal crowd behavior model (ACB) [46], and anomalous trajectory detection (ATD) [47]. The results are averaged over both the datasets and presented in Table V. As can be seen, our method outperforms all the reference methods.

TABLE. III. DEEP LEARNING METHODS PERFORMANCE FOR UMN DATASET

Method	Accuracy	Precision	Recall	F1 Score
CNN without pooling layers	0.65	0.54	0.49	0.52
CNN with pooling layers	0.71	0.48	0.54	0.56
RNN	0,67	0.52	0.46	0.50
PretrainRNN	0.69	0.67	0.47	0.53
StackedRNN	**0.76**	**0.71**	**0.60**	**0.66**

TABLE. IV. DEEP LEARNING METHODS PERFORMANCE FOR UCSD DATASET.

Method	Accuracy	Precision	Recall	F1 Score
CNN without pooling layers	0.48	0.47	0.55	0.51
CNN with pooling layers	0.57	0.58	0.49	0.52
RNN	0,60	0.61	0.56	0.61
PretrainRNN	0.65	0.66	0.66	0.63
StackedRNN	**0.75**	**0.75**	**0.67**	**0.69**

TABLE. V. DEEP LEARNING METHODS PERFORMANCE FOR BOTH THE DATASETS

Method	Accuracy	Precision	Recall	F1 Score
STA [45]	0.71	0.66	0.58	0.53
ACB [46]	0.67	0.61	0.51	0.51
ATD [47]	0,72	0.67	0.61	0.62
StackedRNN	**0.75**	**0.73**	**0.63**	**0.67**

Considering these results, we present overall analysis. First, the large number of parameters might affect the decision accuracy. At the same time, not taking all of the parameters into consideration might lead to wrong decision. So, it is a tradeoff between the two cases. Second, some of the parameters might be more effective than others. Part of our future work is to look into these problems and find an elegant solutions. For instance, fuzzy logic control could handle the uncertainty in the collected data and it might be utilized as well to rank the input parameters. Other extension to this work includes collecting big data related to real world crowd condition and data from sensors installed in different public places. Thus it would enable us to consolidate the current framework with more informative and useful information.

V. DISCUSSION

We have presented a new method for crowd behavior analysis method. We detect and identify panic situations and non-pedestrian entities. It is worth mentioning here that many methods have proposed previously for the same problem as we discussed in the literature review. However, those methods suffer from various problems including lack of generalization capabilities. Moreover, our proposed method is invariant to different key challenges as we mentioned in the introduction section. We carried out detail experimental analysis on two benchmark datasets which are considered very challenging for the same problem in the field of computer vision. Inspired by the concept of the Internet of Things, our method is enriched with robustness to deal with the difficult problem of crowd behavior analysis. In the experimental assessment, we used different performance metrics including accuracy, precision, recall, and F1 Score. Our method showed very performance considering both datasets and performance metrics. In fact, our work can be further extended to many other crowd behavior detection due to its generalization capabilities.

VI. CONCLUSION

In this paper, we have explored a new research direction in crowd behavior detection in smart cities using a novel and a comprehensive framework. Initially, SIFT features are considered to detect crowd behaviors. However, SIFT features are not used as standalone input. To make the framework robust, a number of other parameters are taken into account from the surroundings. Then a deep learning model is trained using the generated training data that detects the crowd behavior in the testing phase. CNN with pooling and without pooling, RNN, pretrained RNN, and Stacked RNN are used as deep learning methodologies. These methods are examined and their performances are validated. The results are promising especially with StackedRNN.

In our future work, we will develop novel deep learning architectures to further enhance the capabilities of our crowd behavior analysis method.

ACKNOWLEDGMENT

This research has been supported by the Research Deanship at the University of Hail, KSA, under the grant number 160778.

REFERENCES

[1] Q. Yi, D. Wu, W. Bao, and Pascal Lorenz. "The internet of things for smart cities: Technologies and applications." IEEE Network 33, no. 2 (2019): 4-5.

[2] C. Franco, A. Guerrieri, C. Mastroianni, G. Spezzano, and A. Vinci, eds. The Internet of Things for smart urban ecosystems. Cham: Springer, 2019.

[3] K. Moez. "Improving formal verification and testing techniques for internet of things and smart Cities." Mobile Networks and Applications (2019): 1-12.

[4] K. Moez, M. Lahami, O. Cheikhrouhou, R. Alroobaea, and A. Jmal Maâlej. "Security Testing of Internet of Things for Smart City Applications: A Formal Approach." In Smart Infrastructure and Applications, pp. 629-653. Springer, Cham, 2020.

[5] G. Kalpna, V. Puri, J. G. Tromp, N. G. Nguyen, and C. V. Le. "Internet of Things (IoT) and Deep Neural Network-Based Intelligent and Conceptual Model for Smart City." In Frontiers in Intelligent Computing: Theory and Applications, pp. 287-300. Springer, Singapore, 2020.

[6] C. Joseph, and J. Evans. "Informal urbanism and the Internet of Things: Reliability, trust and the reconfiguration of infrastructure." Urban Studies (2020): 0042098019890798.

[7] L. Wenwen, M. Batty, and M. F. Goodchild. "Real-time GIS for smart cities." (2020): 311-324.

[8] M. Ullah, H. Ullah, and F. A. Cheikh. "Single shot appearance model (ssam) for multi-target tracking." Electronic Imaging 2019, no. 7 (2019): 466-1.

[9] J. Hao, L. Xie, C. Wang, Y. Yin, and S. Lu. "CrowdSensing: A crowd-sourcing based indoor navigation using RFID-based delay tolerant network." Journal of Network and Computer Applications 52 (2015): 79-89.

[10] M. Ke, P. Zhang, and Z. Mao. "Study on large-scale crowd evacuation method in cultural museum using mutation prediction RFID." Personal and Ubiquitous Computing (2019): 1-15.

[11] M. Marwa F., A. E. Shabayek, and M. El-Gayyar. "IoT-Based Framework for Crowd Management." In Mobile Solutions and Their Usefulness in Everyday Life, pp. 47-61. Springer, Cham, 2019.

[12] L. Tie, J. Huang, S. S. Kanhere, J. Zhang, and S. K. Das. "Improving IoT data quality in mobile crowd sensing: A cross validation approach." IEEE Internet of Things Journal 6, no. 3 (2019): 5651-5664.

[13] G. Frank Joseph Lamont. "Internet-of-things (IoT) device/platform for crowd interaction processing." U.S. Patent 10,445,993, issued October 15, 2019.

[14] M. Yanlan, P. Gui, X. Luo, B. Liang, L. Fu, and X. Zheng. "IoT-based real time intelligent routing for emergent crowd evacuation." Library Hi Tech (2019).

[15] Vijayakumar, V., and K. S. Joseph. "Adaptive Load Balancing Schema for efficient data dissemination in Vehicular Ad-Hoc Network VANET." Alexandria Engineering Journal 58, no. 4 (2019): 1157-1166.

[16] D. Sayan, S. Burman, A. Mazumdar, and N. D. Roy. "Crowd Behavior Analysis and Alert System Using Image Processing." In Emerging Technology in Modelling and Graphics, pp. 721-729. Springer, Singapore, 2020.

[17] X. Yao, S. Liu, Y. Li, and X. Qian. "Crowd Scene Analysis by Output Encoding." arXiv preprint arXiv:2001.09556 (2020).

[18] K. Ajitesh, and M. Kumari. "Design and Analysis of IoT-Based System for Crowd Density Estimation Techniques." In Advances in Data and Information Sciences, pp. 307-315. Springer, Singapore, 2020.

[19] S. D. Khan, and H. Ullah. "A survey of advances in vision-based vehicle re-identification." Computer Vision and Image Understanding 182 (2019): 50-63.

[20] K. S. Daud, H. Ullah, M. Ullah, N. Conci, F. A. Cheikh, and A. Beghdadi. "Person Head Detection Based Deep Model for People Counting in Sports Videos." In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1-8. IEEE, 2019.

[21] B. Saleh, S. D. Khan, and H. Ullah. "Scale driven convolutional neural network model for people counting and localization in crowd scenes." IEEE Access 7 (2019): 71576-71584.

[22] H. Ullah, M. Ullah, and M. Uzair. "A hybrid social influence model for pedestrian motion segmentation." Neural Computing and Applications 31, no. 11 (2019): 7317-7333.

[23] W. Qi, M. Chen, F. Nie, and X. Li. "Detecting coherent groups in crowd scenes by multiview clustering." IEEE transactions on pattern analysis and machine intelligence 42, no. 1 (2018): 46-58.

[24] C. X. Han, and J. H. Lai. "Detecting abnormal crowd behaviors based on the div-curl characteristics of flow fields." Pattern Recognition 88 (2019): 342-355.

[25] B. Patrick C., T. M. Gibbs, and J. Lantz. "Crowd Management and Special Event Planning." In The Professional Protection Officer, pp. 283-293. Butterworth-Heinemann, 2020.

[26] G. Ioannis, P. Gavriilidis, N. I. Dourvas, I. G. Georgoudas, G. A. Trunfio, and G. C. Sirakoulis. "Accelerating fuzzy cellular automata for modeling crowd dynamics." Journal of Computational Science 32 (2019): 125-140.

[27] L. Wenxi, C. Y. Zhang, G. Liu, Y. Su, and N. N. Xiong. "Extraversion Measure for Crowd Trajectories." IEEE Transactions on Industrial Informatics 15, no. 12 (2019): 6334-6343.

[28] H. Dan, and T. Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." arXiv preprint arXiv:1903.12261 (2019).

[29] S. Adam, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. "A simple neural network module for relational reasoning." In Advances in neural information processing systems, pp. 4967-4976. 2017.

[30] W. Qi, J. Gao, W. Lin, and Y. Yuan. "Learning from synthetic data for crowd counting in the wild." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8198-8207. 2019.

[31] L. Noah, H. Gazula, S. M. Plis, and V. D. Calhoun. "Decentralized distribution-sampled classification models with application to brain imaging." Journal of neuroscience methods 329 (2020): 108418.

[32] C. John M., and T. J. Hastie. "Statistical models." In Statistical Models in S, pp. 13-44. Routledge, 2017.

[33] M. Bilal, M. Ullah, and H. Ullah. "Chemometric data analysis with autoencoder neural network." Electronic Imaging 2019, no. 1 (2019): 679-1.

[34] Building IoT Together by Cisco , [online] https://www.cisco.com/web/offer/emear/38586/images/Presentations/P11.pdf

[35] W. Xiangqian, Y. Tang, and W. Bu. "Offline text-independent writer identification based on scale invariant feature transform." IEEE Transactions on Information Forensics and Security 9, no. 3 (2014): 526-536.

[36] Lowe, G. "SIFT-the scale invariant feature transform." Int. J 2 (2004): 91-110.

[37] Unusual crowd activity dataset of university of minnesota, available from http: //mha.cs.umn.edu/movies/crowd-activity-all.avi.

[38] H. Geoffrey E. "Deep belief networks." Scholarpedia 4, no. 5 (2009): 5947.

[39] M. Abdel-rahman, G. E. Dahl, and G. Hinton. "Acoustic modeling using deep belief networks." IEEE transactions on audio, speech, and language processing 20, no. 1 (2011): 14-22.

[40] N. Vinod, and G. E. Hinton. "Rectified linear units improve restricted boltzmann machines." In Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807-814. 2010.

[41] S. Suvash, A. K. Menon, S. Sanner, and L. Xie. "Autorec: Autoencoders meet collaborative filtering." In Proceedings of the 24th international conference on World Wide Web, pp. 111-112. 2015.

[42] V. Andrea, and K. Lenc. "Matconvnet: Convolutional neural networks for matlab." In Proceedings of the 23rd ACM international conference on Multimedia, pp. 689-692. 2015.

[43] G. Alex, A. R. Mohamed, and G. Hinton. "Speech recognition with deep recurrent neural networks." In 2013 IEEE international conference on acoustics, speech and signal processing, pp. 6645-6649. IEEE, 2013.

[44] Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In: IEEEconference on computer vision and pattern recognition (CVPR), pp 1–8.

[45] O. Nitish, and A. Vaish. "Spatio-temporal anomaly detection in crowd movement using SIFT." In 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 646-654. IEEE, 2018.

[46] L. Yue, K. Hao, X. Tang, and T. Wang. "Abnormal Crowd Behavior Detection Based on Predictive Neural Network." In 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 221-225. IEEE, 2019.

[47] D. Deepan, and D. Mishra. "Unsupervised Anomalous Trajectory Detection for Crowded Scenes." In 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS), pp. 27-31. IEEE, 2018.

Short Poem Generation (SPG): A Performance Evaluation of Hidden Markov Model based on Readability Index and Turing Test

Ken Jon M. Tarnate[1], May M. Garcia[2], Priscilla Sotelo-Bator[3]
Computer Studies Department, College of Science
Technological University of the Philippines, Manila, Philippines

Abstract—**We developed a Hidden Markov Model (HMM) that automatically generates short poem. The HMM was trained using the forward-backward algorithm also known as Baum Welch algorithm. The training process was exhausted by a hundreds of iterations through recursion method. Then we used the Viterbi algorithm to decode all the best possible hidden states to predict the next word, and from the previous predicted word, it will generate another word, then another word until it reaches the desire word length that was set in the program. Afterwards, the model was evaluated using several kinds of readability metrics index which measure the reading difficulty and comprehensiveness of the generated poem. Then, we performed a Turing Test, which participated by 75 college students, who are well versed in poetry. They determined if the generated poems was created by a human or a machine. Based from the evaluation results, the highest readability score index of the generated short poem is in the grade 16th level. While 69.2% of the participants in the Turing Test, agreed that most of the machine generated poems were likely created by some well-known poets and writers.**

Keywords—Evaluation metrics; Hidden Markov Model; poetry generation; readability test; turing test

I. Introduction

Hidden Markov Model (HMM) has been successfully explored and applied in the fields of medical technology, military, forensics, bioinformatics, data security and even in arts and literatures. HMM models had also been widely used to create various types of applications software such as speech recognition, image and signal processing, and even text and poetry generation. As of today, Hidden Markov Model (HMM) has now become the based algorithm for creating a text generator, text summarizer, lyrics and music generator [1], [17] These applications is belong into one specific area of natural language processing called computational creativity, where the goal of the artificial intelligence (AI) is to change the nature of creative processes, where the machine will compete with human creativeness in terms of writing through the use of different mathematical models and algorithms.[2],[3] One specific product of this area is called "Poetry Generation." Where the machine will automatically generate poem(s) based from historical data or corpus data used to train the AI model. This complex task required a considerable amount of input knowledge (e.g. phonetics, syntax, semantics, grammar and rhymes). And currently, Hidden Markov Model (hmm) have been successfully conquer this specific topic area of the natural language processing [4], [5]. However, there are still room and subject for improvements, especially on the testing analysis and performance evaluation of the HMM model. Most of the published papers focus on the inner performance of the HMM model and used F1-score, precision and recall to calculate its accuracy. However, only few had considered to measure the outer performance of the HMM model and evaluated the content features of its generated output. In this paper, we tested the learning ability of the hidden markov based on the number of iterations performed before it produces a high quality machine generated poem. [6], [7] Then we evaluated the content of the generated poem by getting the readability score index. And lastly, we performed a standard Turing Test, to confirm the validity and authenticity of the generated poems. Where the participants are asked to determine, whether the generated poem was created by machine or human?

II. Related Works

There is an ample research on the application of Hidden Markov Model in the area of computational creativity and evaluated their model based on the quality and content of the generated composition (e.g. poems, lyrics, short story, novel and etc.) [8], [9].One of these researches is the "Wishful Automatic Spanish Poet" a program created by Pablo Gervás (2000). Where he tested and evaluated his HMM model by calculating the correctness of the generate poem based from the rhymes, syllables and word repetition [8]. Meanwhile, Hugo Oliveira (2008) created a platform for the automatic generation of poetry called "PoeTryMe." He evaluated his model using rhyme and lexical density. Rhyme density was defined as a quantitative measure of the technical quality of the composition. [10] While lexical density is getting the meaningfulness of the composition by identifying the basic parts of speech such as nouns, pronouns, verbs, adverbs, etc. [11], [14] For decades, the literature in poetry generation and the application of Hidden Markov Model in the computational creativity and natural language processing is still uprising and for many years it was already proven that Hidden Markov Model was suitable and efficient to use for text generation. One examples of this are; Polish language text generator [11] Steganographic text based [12] and Chinese couplets generation [13] However, researchers strongly recommended to explore more on the type of testing and what type of other evaluation metrics can be more suitable to evaluate the performance of the HMM model [6], [7]. For this purpose, we explored and used the other evaluation metrics used in computational creativity to evaluate our HMM model.

III. Research Methodology

A. Data Collections and Pre-Processing

A total of 1600 variety poems have collected from different sources; 500 poems extracted manually in the poemhunter.com website, 500 poems that talk about life which come from the poemsforfree.com website, and another 500 poems freeromanticlovepoems.net. We also included the Pablo Neruda collection which contains 100 different love poems. By combining all those poems in a single text file and make it as a corpus dataset. We pre-process the text by omitting unnecessary spaces, symbols and characters such as (&. /, *, "", (), etc.).

B. Content based Analysis on the Corpus Data

We used text analyzer software, which interpret the content of the collected dataset. Below are the results of the analysis.

C. Training of Hidden Markov Model

We used the forward-backward algorithm also known as Baum Welch algorithm. Then run the program with hundreds of iterations. And using the Viterbi algorithm it decodes the learning and predicts the next words based from the previous generated words. This process was repeated one hundred times to ensure that our model will give a better result compared to the previous results. Afterwards, the program will require an input seed word before it generates another word which will be the basis of the Hidden Markov Model to predict the next word (see Fig. 1 for the details of the model and see Fig. 2 for the actual sample output of the Hidden Markov Model.).

D. Testing of Hidden Markov Model

After the training of the model, we run the program for a hundreds of iterations, until the model produced a better quality of generated poems. In our experiments, we enter a word which are not present on the datasets and tested out if the Hidden Markov Model will still generate text.

Table II shows the numbers of iterations occurred before the Hidden Markov Model (HMM) generates a high quality of short poem. As we observed, the number of iterations is relevant for the generation of a good quality composition. And even though the seed words used in the experiment is not present on the data corpus, the Hidden Markov Model still able to generate short poems.

TABLE. I. Results of Text Analysis

Type of Test	Score
Total Word Count	19842
Number of different words	12727
Complexity factor	27.7%
Readability (Gunning-Fog Index)	7.3
Total number of characters	93521
Number of characters without spaces	52748
Average syllables per word	1.43
Sentence count	11055
Average sentence length	16.56
Max sentence length	188
Min sentence length	1

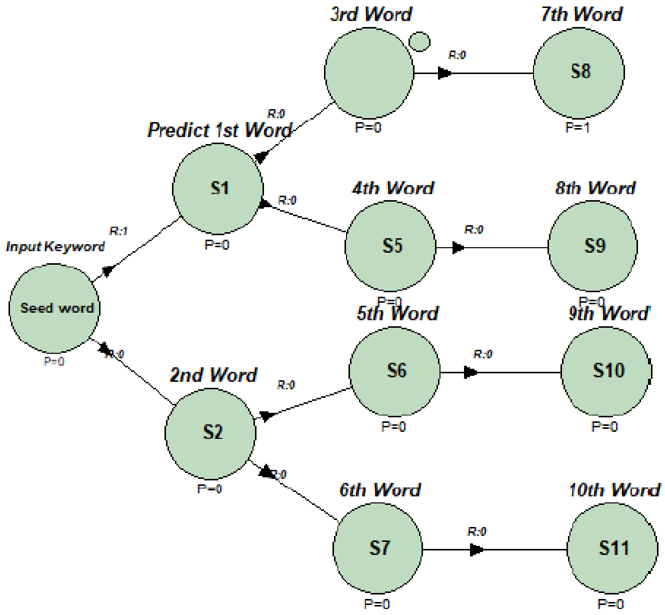


Fig. 1. Design of Hidden Markov Model for Generating Short Poem.

TABLE. II. Testing of Hidden Markov Model using a Seed Word which is not Present on the Corpus Dataset

Seed Word	Generated Poems	No. of Iterations
Education	Education is a room i fear it magnify myself, approaching to forever half of the mind already in my fantasy; it is the truth, that makes me smile, madly but sir, it is the room i fear room, after room, it	Best results after 58 iterations
Money	Money is riches it is up to you to take a rest or fall apart, by lips we call, the person might have joy or fear of it, the love for it is a road to be taken. a thirst from it lies a painful life or maybe a passionate one.	Best results after 110 iterations
Politics	Politics lonely is as love, it must be insist. one is eternal from wisdom trite let your will be different and feels all the gypsy of the men	Best results after 98 iterations
Envy	Envy is not touch not will. i feel them and recall, the truth about liv'd fires the joy of fire it brings harden yet is not touch, to cut is to clean if our hearts make us sea. to keep from pursuing the one i fierce	Best results after 127 iterations
Jealous	Jealous cannot choose you and vague time, and all the breeze, gets gentle the relishing air, fresh of love which rather i would flail, his admittance gained the distractions is more growing until the day	Best results after 128 iterations
Average Iterations		**104.2**

IV. Performance Evaluation

Evaluation of machine generated poem is a vigorously difficult task. Based from the previous researches, most of the machine learning model was evaluated using accuracy metrics such as F1-score, Precision and Recall However, in this study, we used Readability Index to examine the content and comprehensiveness of the generated poem and performed a standard Turing Test to examine the authenticity of the machine generated poem produced by Hidden Markov Model (hmm).

A. *Readability Index*

Readability Index is defined as the estimate difficulty of a text to read. By measuring the text's complexity by counting the different attributes present on the text such as word lengths, sentence lengths and syllables. There are standards readability metrics used in the United States both in public and private schools and reading centers (e.g. Gunning-Fog, Flesch-Kincaid, Smog, Coleman-Liau and Automated Readability Index) [15], [16].

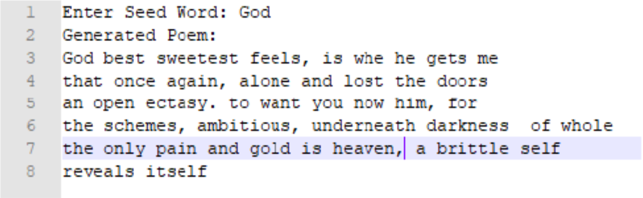
Based from the results of readability test, the sample generated poem was appropriate and suitable for the Grade 9th to Grade 10th students. This means that the reading difficulty of the entire sample text is in the Highschool level (see Fig. 3). Note: The automated graph in Fig. 3 was generated using text analyzer software powered by "analyzemywriting.com".

Table III shows the results of the readability index. Based from the results the average grade level difficulty of the text is suitable for grade 16th mostly fall for college students and working professionals. This means that the content of the generated poems was really deep and highly constructed.

B. *Turing Test*

This type of test is defined by Professor Alan Turing in his paper "Computing Machinery and Intelligence" where the AI machine will be tested based on how closely it can resemble or compete to human's intelligent [5], [14]. And finding human evaluators who are technically expert and well versed in writing is a tedious task. For this purpose, we sampled our machine generated poems to the 25 IT students, 25 IS students and 25 Computer Science students of Technological University of the Philippines. All participants were informed that the test was for a research project and instructed them about the Turing Test. We used the Table I sample generated poem as an actual sample to be tested.

As we observed in Table IV, the overall result was positive as the entire generated short poems were able to deceive a significant numbers of participants. Thinking that they were written by human poets, which is the primary goal of this experiment. The generated poem "Money" got the highest positive results. 83% of the participants agreed that this verse was created by some well-known poet or writer. Based from these results, the machine generated poem successfully deceived the human perspective and creativeness where 69.2% of the participants agreed that the poem generated by the Hidden Markov Model can compete to the human creativeness. (see Fig. 4, for the graph results of the conducted Turing Test.)



```
Enter Seed Word: God
Generated Poem:
God best sweetest feels, is whe he gets me
that once again, alone and lost the doors
an open ectasy. to want you now him, for
the schemes, ambitious, underneath darkness  of whole
the only pain and gold is heaven, a brittle self
reveals itself
```

Fig. 2. Sample Generated Poem Tested in a Text Analyzer Software.

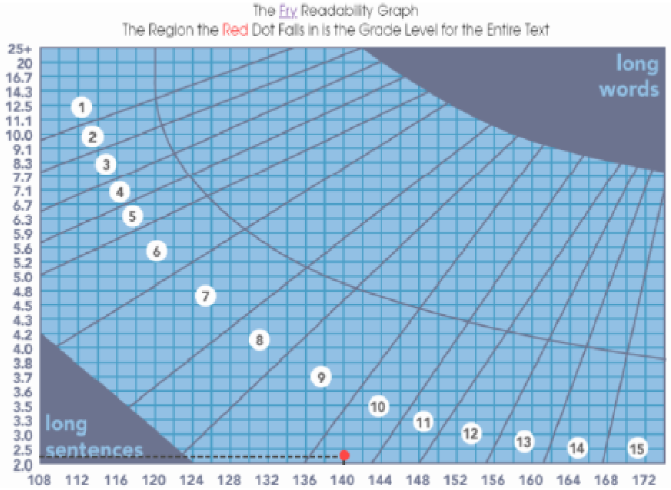


Fig. 3. Results of Readibility Test.

TABLE. III. Results of Readability Test

Type of Test	**Grade Level**
Gunning-Fog	20.67
Flesch-Kincaid	18.48
SMOG	13.02
Coleman-Liau	9.41
Automated	21.79
Average Grade Level:	**16.68**
Median Grade Level:	**18.48**

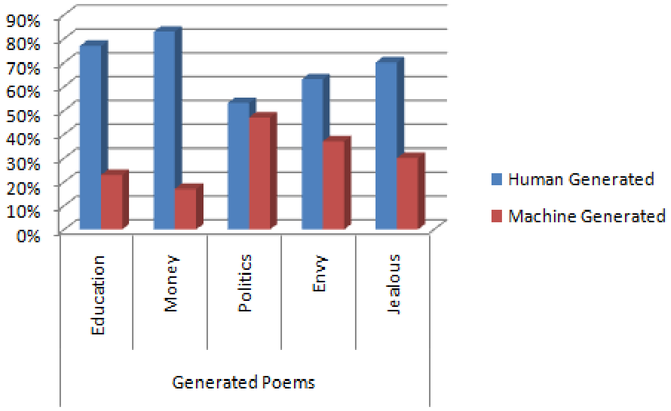


Fig. 4. Results of Turing Test.

TABLE. IV. RESULTS OF TURING TEST (IS THE POEM CREATED BY HUMAN OR BY THE MACHINE?)

Seed Word	Generated Poems	Human Generated	Machine Generated
Education	Education is a room i fear it magnify myself, approaching to forever half of the mind already in my fantasy; it is the truth, that makes me smile, madly but sir, it is the room i fear room, after room, it	77%	23%
Money	Money is riches it is up to you to take a rest or fall apart, and by lips we call, the person might have joy or fear of it, the love for it, is a road to be taken. a thirst from it lies a painful life or maybe a passionate one.	83%	17%
Politics	Politics is as love, its loneliness must be insist. one is eternal from wisdom trite let your will be different and feels all the gypsy of the men	53%	47%
Envy	Envy is not touch not will. i feel them and recall, the truth about liv'd fires the joy of fire it brings harden yet is not touch, to cut is to clean if our hearts make us sea. to keep from pursuing the one i fierce	63%	37%
Jealous	Jealous cannot choose you and vague time, and all the breeze, gets gentle the relishing air, fresh of love which rather i would flail, his admittance gained the distractions is more growing until the day	70%	30%
Average		**69.2%**	**30.8%**

V. CONCLUSIONS

The designed Hidden Markov Model has successfully generated short poems which able to pass the Turing Test and able to deceived the human mind. We able to achieved our objectives of examining the performance of the Hidden Markov Model using only the content features of its generated output and not relying on the accuracy metrics of the model which most of the researchers done. We attack a new type of approach of testing a machine learning model using the environment factors such as human, lexicons and comprehensiveness of the generated text. Currently, the outputs of our model are not yet perfectly correct in terms of grammars and semantics. As for the future works, exploring the semantic and syntactic relations of the words should be a good opportunity to look at, to develop new feature data engineering pipeline and approaches to improve the encoding and decoding process of the Hidden Markov Model.

ACKNOWLEDGMENT

This research was accomplished by the grace and power of our Lord Jesus Christ. All the glory and honor belongs to Him. Special thanks to: Dr. Dionisio A. Espression Jr., Prof. Fidela Q. Aranes, Prof. Fernando L. Renegado and Dr. Ira C. Valenzuela and the entire University Research Development Services (URDS) for making this research possible. – TUP for the World!

REFERENCES

[1] Addanki, K., & Wu, D. (2013, July). Unsupervised rhyme scheme identification in hip hop lyrics using hidden Markov models. In International conference on statistical language and speech processing (pp. 39-50). Springer, Berlin, Heidelberg.

[2] Besold, T. R., Schorlemmer, M., & Smaill, A. (Eds.). (2015). Computational creativity research: towards creative machines.

[3] Petrushin, V. A. (2000). Hidden markov models: Fundamentals and applications. In Online Symposium for Electronics Engineer.

[4] Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. Lund Working Papers in Linguistics, 53, 61-79.

[5] Fernandez, A. C. T., Tarnate, K. J. M., & Devaraj, M. (2018). Deep Rapping: Character Level Neural Models for Automated Rap Lyrics Composition. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-2SR. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Kamal, M. S., Chowdhury, L., Khan, M. I., Ashour, A. S., Tavares, J. M. R., & Dey, N. (2017). Hidden Markov model and Chapman Kolmogrov for protein structures prediction from images. Computational biology and chemistry, 68, 231-244.

[7] McCane, B., & Caelli, T. (2004). Diagnostic tools for evaluating and updating hidden Markov models. Pattern recognition, 37(7), 1325-1337.

[8] Gervás, P. (2000, April). Wasp: Evaluation of different strategies for the automatic generation of spanish verse. In Proceedings of the AISB-00 symposium on creative & cultural aspects of AI (pp. 93-100)

[9] Silva, E. D. S., Leão, R. M. M., & Muntz, R. R. (2010, October). Performance evaluation with hidden markov models. In International Workshop on Performance Evaluation of Computer and Communication Systems (pp. 112-128). Springer, Berlin, Heidelberg.

[10] Oliveira, Hugo Gonçalo. "PoeTryMe: a versatile platform for poetry generation." Computational Creativity, Concept Invention, and General Intelligence 1 (2012): 21.

[11] Szymanski, G., & Ciota, Z. (2002). Hidden Markov models suitable for text generation. In WSEAS International Conference on Signal, Speech and Image Processing (WSEAS ICOSSIP 2002) (pp. 3081-3084).

[12] Yang, Z., Jin, S., Huang, Y., Zhang, Y., & Li, H. (2018). Automatically generate Steganographic text based on markov model and Huffman coding. arXiv preprint arXiv:1811.04720.

[13] Pan, Z., Zhang, S., & Guo, Y. (2018). Easycouplet: Automatic generation of Chinese traditional couplets. In Transactions on Edutainment XIV (pp. 117-132). Springer, Berlin, Heidelberg.

[14] Tarnate, K. J. M., & Devaraj, M. (2019) Prediction of ISO 9001: 2015 Audit Reports According to its Major Clauses using Recurrent Neural Networks. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2.

[15] Fernandez, A. C. T. (2019). Computing the Linguistic-Based Cues of Credible and Not Credible News in the Philippines Towards Fake News Detection.

[16] Senter, R. J., & Smith, E. A. (1967). Automated readability index. Cincinnati Univ Oh.

Karaa, W. B. A., & Dey, N. (2017). Mining multimedia documents. Chapman and Hall/CRC.

A Novel Image Fusion Scheme using Wavelet Transform for Concealed Weapon Detection

Hanan A. Hosni Mahmoud
Department of Computer Science, College of Computer and Information Science
Princess Nourah bint Abdulrahman University, Riyadh, 11671 KSA
Dept. of Computer and Systems Engineering, Faculty of Engineering, University of Alexandria, Egypt

Abstract—**The aim of this paper is to detect concealed weapons, especially in high security places like in airports, train stations and places with large crowds, where concealed weapons are not allowed. We aim to specify suspicious person who may have a concealed weapon. In this paper, an Image Fusion technique using pixel alignment and discrete wavelet transform is proposed. It is mainly utilized for Concealed Weapon Detection. Image fusion can be defined as extracting information from two or images into a single image to enhance the detection. Image fusion allows detecting concealed weapons underneath a person's clothing with imaging sensors such as Infrared imaging or Passive Millimeter Wave sensors. A data fusion scheme for simpler sensors based on correlation coefficients is proposed and utilized. We proposed an image fusion scheme that utilizes fusion dependency rules using wavelet (WT) and inverse wavelet transform (IWT). The fusion rule is to select the coefficient with the highest correlation rate. The higher the correlation the stronger of the co-existed feature. Experimental results shows the superiority of the proposed algorithm both in quality and real time requirement. The proposed algorithm has a real time response time that is less than other comparable algorithms by 40%. At the same time it retains higher quality as shown in the experimental results. It outperforms other algorithms by superior PSNR of more than 10% of the comparable algorithms in average.**

Keywords—*Concealed weapon detection; image fusion; pixel alignment; wave sensors*

I. Introduction

Video surveillance systems acquire a video stream from the scene under monitoring from several sensors distributed across the area of interest. The analysis of the video stream begins with the detection of moving objects, and then recognition of the detected object is performed in order to classify it [1]-[3]. Then the object trajectory is identified in order to analyze the object's behavior or activities. Fig. 1 and Fig. 2 illustrate the processing flow in a visual surveillance system. The processing flow gives the first glance of the core challenges such as object identification. CCTV provides the visual sensors used for this type of systems; however, the CCTV cameras offer low resolution and low frame rate as well as varying quality due to environmental conditions such as changes in illumination. Tracking is another core challenge because of coordination required between different cameras [4]-[5].

A. *Processing of Visual Surveillance*

1) Object detection: Optical flow and background subtraction are usually utilized in object detection. Optical flow is computationally complex [6]. The main disadvantage of background subtraction technique is that a fixed background is required [7]-[8]. In outdoor environments, the high variability of environmental conditions requires robust adaptive background models that are computationally more expensive. In [9] the introduced a new technique to detect moving objects in MPEG videos utilizing modified optical flow and background subtraction algorithms.

2) Object classification: Object classification is a tough task in such low-resolution images. Surveillance footage has a quite poor resolution, resulting in an object of interest spanning in only a few pixels in each frame [10]-[11]. Two approaches have been used to solve this problem, histogram and model-based techniques. Color histogram of frames are calculated in the first approach. The model-based approach employs apriori geometrical knowledge of the objects of interest. The apriori knowledge can be constructed based on the appearance and features of the object. In [12] the authors used Convolutional Neural Network to analyze Low Resolution Thermal Images based on Embedded Platform.

3) Tracking of detected objects: Tracking introduces several challenges to the system, such as tracking in different lighting conditions. Shadow detection must be performed in order to avoid tracking a shadow instead of the real moving object. Another challenge is to track a person in a cluttered and dynamic background. The techniques employed for tracking can be classified in three groups: filtering techniques, statistical models, and multi-agents systems [13]-[14]. Filtering techniques such as Kalman filtering have been successfully employed in surveillance systems. Object tracking system utilizing Kernalized filters and Kalman predictive estimates was introduced in [15]. A Hidden Markov Model has been also used for tracking purposes; however, offline training data is required. Another technique used for tracking is Multi-agent systems, which is a well suited framework.

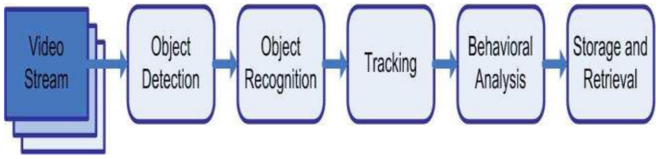


Fig. 1. Classical Processing Flow in Visual Surveillance Systems.

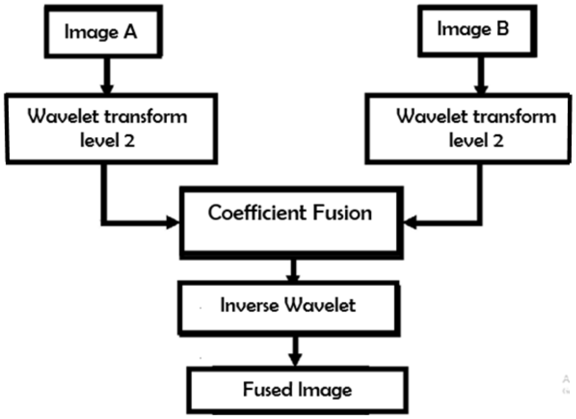


Fig. 2. The Proposed Image Fusion Scheme FMWFR.

4) Behavioral analysis: Behavioral analysis aims to describe the activity that is taking place in the area under monitoring. This is a classification problem for the feature data provided by the previous stages during some period of time. In [16], they introduced predicting unrest in social events using hidden Markov models.

B. Image Fusion Schemes

The processing of multiple images into a single image is defined as image fusion. It is very important the image fusion is done without employing final image distortion or decreasing entropy [11]. Image fusion methods vary and based on many techniques. Wavelet transform includes various frequency sub-bands. In [14], the authors introduced Multi-sensor image fusion utilizing empirical wavelet transform, showing high performance through experimentation of images came from different sensors. Normalized convolution framework through a multi-frame super resolution of digitized videos is introduced in [16]. Panchromatic Fusion utilizing Empirical Wavelet technique was introduced in [7]. In [13], they implemented image fusion algorithm in the wavelet domain and built a high performance FPGA to implement the proposed algorithm. Also, the authors in [15] proposed an Image Fusion Implementation. In [15], the authors proposed several image fusion Techniques that were based on discrete wavelet transform. Approaches other than wavelet transform are the techniques proposed in [7]. They proposed a dictionary entropy image fusion framework that was based on deep learning. In [8] the classification process was based on wavelet image fusion was introduced. In [9], texture features were used to perform image fusion.

II. Methodology

We propose an image fusion scheme that fuses images at the pixel level using a wavelet transform of the source images. Image decomposition by wavelet transform at different levels is shown in Fig. 3. The fusion rule is described by the dependency of the wavelet coefficient. The dependency metric will identify the stronger pattern.

A. The Proposed Image Fusion Scheme with Salient Feature Extraction FMWFR

We employed linear dependency metric to be performed on the pixels of the two block of dimension n × n. If the two blocks are unrelated then those pixels are linearly independent, and no relevant feature in the window. According to this metric, if the two blocks are related, then the pixels are linearly dependent, and pattern exists. We utilized Wronskian determinant formula to assess this metric. The problem is that background information of any two images are highly correlated. Therefore, in this paper we accommodate a simple preprocessing step, where background information in an image is detected from the frames of videos containing this image. Then background information is converted to all black pixels. For image B the background will be converted to all white pixels. Thus, no similarities of salient feature will be determined based on background similarities.

The LL band of the wavelet transform (WT) preserves the related features of the original image. LH, HL, HH bands hold information about the prominent feature of the original image. Fusion rule diagram is depicted in Fig. 4. The Wronskian determinant of the prominent features on both images deploys the dependency as follows:

$$E_{MWFR(A,B)}(m,n) = \sum_{y=m-1}^{m+1} \sum_{x=n-1}^{m+1} \left[\frac{[WT_A(y,x)]}{[WT_B(y,x)]}\right]^2 - \left[\frac{[WT_A(y,x)]}{[WT_B(y,x)]}\right] \qquad (1)$$

Where, WTA(y,x) is defined as the WT coefficient of image A with the position (y,x) in the 2D image.

Linear dependency is proven in equation 1. The similarity of the features in images A and B is detected by an increased value of EAB. EAB equals zero in the case of no features coexist in both mages A and B. The fusion algorithm determines the existence of the feature in image A which are more significant that the feature exists in B as depicted below:

$$F_{MWFR(A,B)}(m,n) = \begin{cases} WT_A(y,x) \; if \; E_{MWFR(A,B)}(m,n) \geq F_{MWFR(B,A)}(m,n) \\ WT_B(y,x) \; if \; E_{MWFR(A,B)}(m,n) < F_{MWFR(B,A)}(m,n) \end{cases} \qquad (2)$$

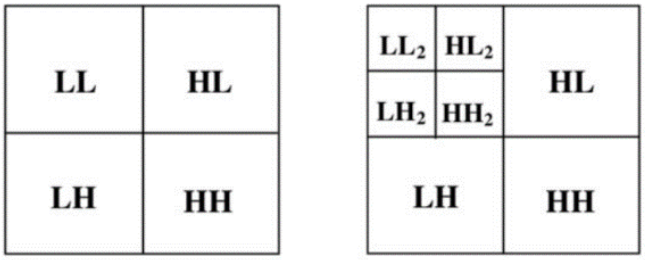


Fig. 3. Image Decomposition by Wavelet Transform at different Levels.

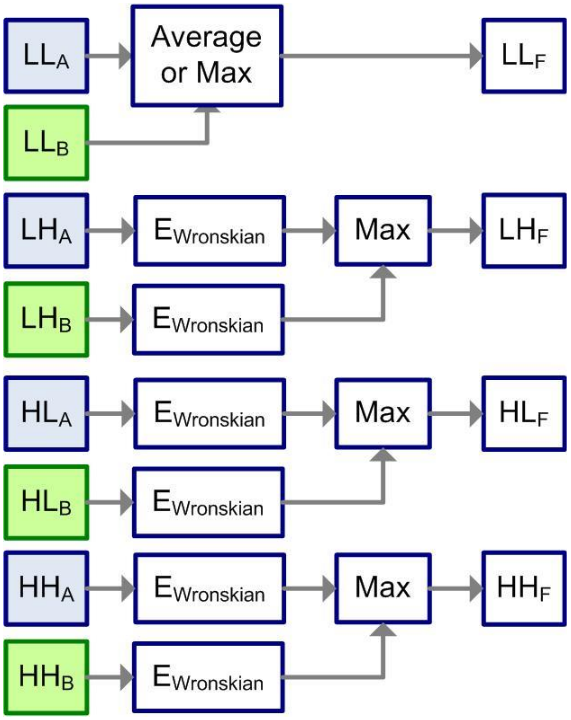


Fig. 4. Fusion Rule Diagram.

The main steps of the proposed algorithm of fusion of several bands of two source images is depicted in Fig. 5, and Fig. 6. The algorithms are illustrated as follows:

Algorithm 1. Fusion (Input: Image A; Image B, Output: Fused Image F)

Start

Preprocessing Phase:

a. Register the source images: source image *A and* source image *B.* from videos V_A and V_B
b. Detect background information in image *A* from the frames of videos containing this image.
c. Convert background information to all black pixels.
d. Detect background information in image *B* from the frames of videos containing this image.
e. Convert background information to all white pixels.

The Fusion Phase FMWFR

1. Resample images A and B in such a way that the pixels of the A and B are aligned.
2. Perform the discrete wavelet transform of A and B.
3. Calculate $W_A(y,x)$ (The WT coefficient of the original image *A* at position *y,x* in the spatial fixation).
4. Calculate $W_B(y,x)$ (The WT coefficient of the original image *B* at position *y,x* in the spatial fixation).
5. Calculate the Wronskian determinant of A and B using equation 1 as follows:

$$E_{MWFR_{AB}}(i,j)=\sum_{r=i-1}^{i+1}\sum_{c=j-1}^{j+1}\left[\frac{W_A(r,c)}{W_B(r,c)}\right]^2-\frac{W_A(r,c)}{W_B(r,c)}$$

The algorithm performs the fusion decision as follows:

```
If ($E_{AB}$ = 0)
Then
The feature $f$ does not exist in B
Elseif ($E_{AB}$ is greater value)
Then
{There is a feature $f$ exists in in A and B
Call Fusion rule algorithm}
```

Fusion rule algorithm

1. $$F_{MWFR}(i,j)=\begin{cases}W_A(i,j) & if \quad E_{MWFR_{AB}}(i,j)\geq E_{MWFR_{BA}}(i,j)\\ W_B(i,j) & if \quad E_{MWFR_{AB}}(i,j)< E_{MWFR_{BA}}(i,j)\end{cases}$$
2. Join WT of A and B spatially to get *AB.*
3. Apply inverse WT on *AB.*
4. final image is fused from A and B.

End

Algorithm 2. Fusion rule algorithm (Input: wavelet transform of image A; wavelet transform of image B, Output : LL_F, LH, HL, HH of the fused image F).

Start

1. *Compute LL_F= Average(LL_A, LL_B);*
2. *Compute LH_F= Max($E_{Wronskian}(LH_A)$, $E_{Wronskian}(LH_B)$);*
3. *Compute HL_F= Max($E_{Wronskian}(HL_A)$, $E_{Wronskian}(HL_B)$);*
4. *Compute HH_F= Max($E_{Wronskian}(HH_A)$, $E_{Wronskian}(HH_B)$);*

End

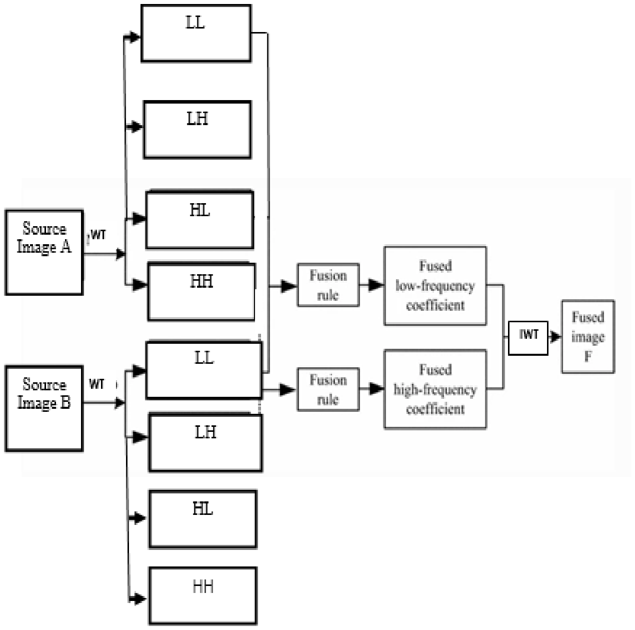


Fig. 5. Fusion of Several Bands of Two Source Images.

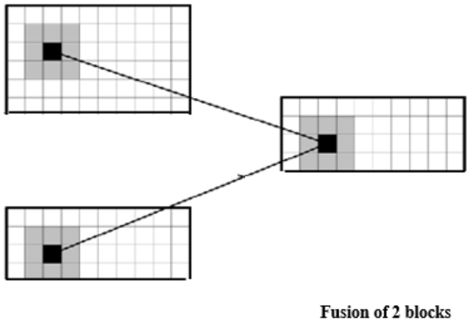


Fig. 6. Fusion of Two Blocks from Two Source Images.

III. Experimental Results

Simulation using test images that are extracted from [10] with permission are shown in Fig. 7. The results with the previous schemes as well as the proposed schemes are shown in Fig. 8 and Fig. 9 for images a and e respectively, the rest of the results are omitted due to space limit. The test images contained images with different exposition or focusing on different objects. Fig. 7 contains images coming from different sensors including Infrared, visible, MMW, and MRI. The methods tested are a combination of the rules explained above and are described in Table I.

Image Fusion at real time is very crucial to detect threats such as concealed weapon. Therefore, the evaluation criteria of any fusion algorithm should include fusion time. Image fusion scheme retains as much information as possible from the sources while introducing as few inconsistencies as possible. Therefore, another evaluation criteria should be done in terms of information, quality, and spectral efficiency. Qualitative measures by showing before and after images are shown in Fig. 7, 8 and 9. The assessment of the quality of the algorithm are performed with the help of reference image. Peak Signal to Noise Ratio (PSNR) are usually employed to compare a distorted image with a distortion-free image. PSNR establishes spectral differences within the fused image compared to reference image. The spatial criterion determined the spatial details in terms of maximizing the correlation between them. We compared all the algorithms in table.

In respect of fusion time and spectral quality through PSNR, the comparison is shown in Fig. 10 and 11. The results for Mutual Information show that WFR-MS: WFR using MS for the LL bands achieves the best PSNR, while the WFR-Ave achieves comparable PSNR as WFR-MS but wilt better fusion time as shown in Fig. 11. The spectral quality of fused images can be measured by several metrics such as Spectral angle mapper (SAM), Relative average spectral error (RASE). The Cross correlation (CC). In Table II, we are comparing the algorithms in Table I for images (a) and (b) in the first row of Fig. 9. Table III shows the same metrics calculating the averages of SAM, RASE and CC for all images in the first rows of Fig. 8 and 9. From the two Tables II and III, it shows the superiority of our proposed techniques WFR-MSWR, WFR-MS and WFR-Ave.

TABLE. I. The Methods Tested which are Combination of different Rules

Method	**Description**
DMR: The Decision Map Rule	DMR consists of Energy feature extraction Energy for all bands except LL with a maximum criterion. The LL band is fused using the average rule [15].
FBR: The Feature Based Rule	FBR employs the expected value of features for extraction of the feature and the selection of the coefficient is based on the maximum value of the feature. The LL band is fused by averaging the coefficients using the average rule [16].
Ave : The Average Rule	Ave does not extract any feature and the fusion rule consists of averaging the wavelet coefficients of the two source images [12].
MSWR: The Maximum Square Window Rule	MSWR uses the extraction rule and the fusion rule F_{Max} [13].
WFR-MSWR: The proposed algorithm case # 1	WFR- MSWR uses the Wronskian Fusion Rule combined with the proposed feature extraction FMWFR for all the bands except LL. The LL band was fused using maximum square window rule.
WFR-MS: The proposed algorithm case # 2	The Wronskian fusion rule uses the proposed feature extraction FMWFR for all the bands except LL. The LL band was fused by the maximum square rule.
WFR-Ave: The proposed algorithm case # 3	The Wronskian fusion rule uses the proposed feature extraction FMWFR for all the bands except LL. The LL band was fused by averaging.

TABLE. II. Spectral Metrics Comparison for Individual Images

	Spectral Metrics					
	Image a			***Image b***		
	RASE	***SAM***	***CC***	***RASE***	***SAM***	***CC***
DMR: The Decision Map Rule	58.36	0.176	0.87	56.48	0.158	0.91
FBR: The Feature Based Rule	55.38	0.211	0.78	56.01	0.252	0.81
Ave : The Average Rule	52.85	0.243	0.78	53.05	0.201	0.73
MSWR: The Maximum Square Window Rule	64.54	0.066	0.95	65.34	0.076	0.89
WFR-MSWR: The proposed algorithm case # 1	72.58	0.058	0.98	73. 88	0.058	0.97
WFR-MS: The proposed algorithm case # 2	72.57	0.250	0.97	72.46	0.276	0.95
WFR-Ave: The proposed algorithm case # 3	72.69	0.118	0.92	72.91	0.126	0.93

TABLE. III. AVERAGE SPECTRAL METRICS

Spectral Metrics			
	Average RASE	***Average SAM***	***Average CC***
DMR: The Decision Map Rule	44.547	0.159	0.87
FBR: The Feature Based Rule	43.352	0.282	0.78
Ave : The Average Rule	43.801	0.337	0.78
MSWR: The Maximum Square Window Rule	61.388	0.200	0.95
WFR-MSWR: The proposed algorithm case # 1	65.593	0.250	0.98
WFR-MS: The proposed algorithm case # 2	63.891	0.351	0.97
WFR-Ave: The proposed algorithm case # 3	64.520	0.206	0.92

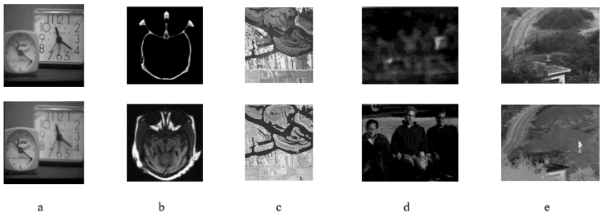


Fig. 7. Test Images from different Sensor Types or with different Exposition or Focus [10].

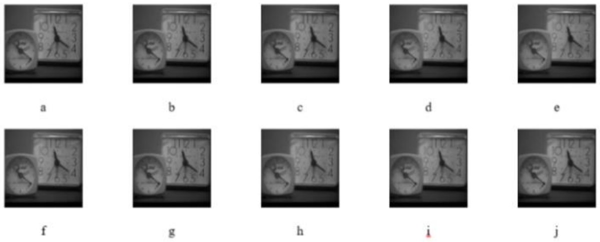


Fig. 8. Images Obtained after Fusing Two Images Focused in different Objects.

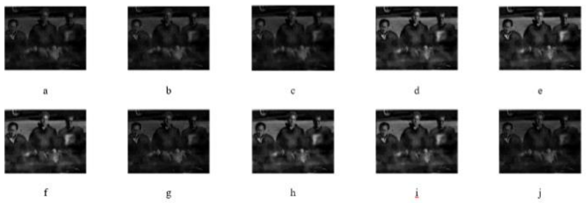


Fig. 9. Images Obtained after Fusing a MMW Image and a Visible Image.

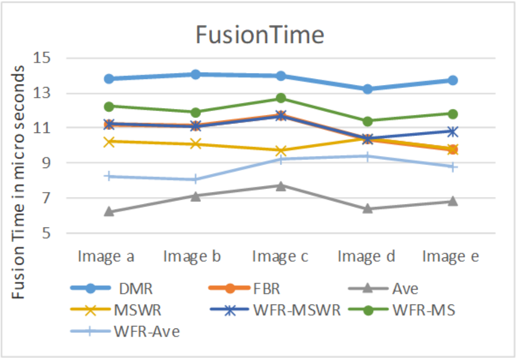


Fig. 10. Fusion Time for Images Obtained after Fusing a MMW Image and a Visible Image.

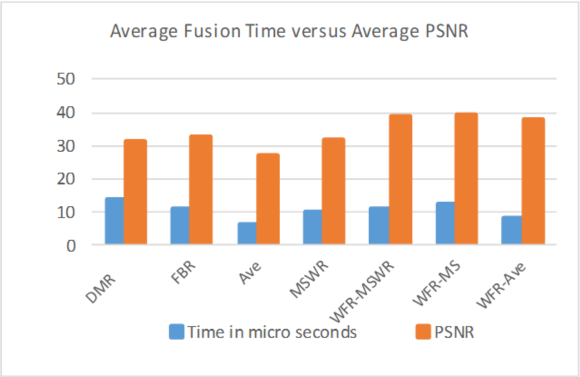


Fig. 11. Average Fusion Time Versus Average PSNR.

IV. CONCLUSION

Security systems emphasize on safety measures, one of them is the detection of concealed weapons. In the proposed architecture, the fusion of infrared images and visual information offers concealed weapon detection for secure-sensitive areas. Some of the results were shown earlier. In addition, the use of infrared technology for detection of concealed weapon raises privacy issues. In order to protect the privacy of the people crossing the scene under monitoring, a conditional image fusion can be done. First, a search for a concealed weapon should be applied to the infrared image alone. If the scene contains a suspected weapon, then fusion should be performed to identify the person carrying it. In the proposed technique, we employed linear dependency metric to be performed on the pixels of the two block of dimension n × n. If the two blocks are unrelated then those pixels are linearly independent, and no relevant feature in the window. According to this metric, if the two blocks are related, then the pixels are linearly dependent, and pattern exists. We utilized Wronskian determinant formula to assess this metric. Spatial and spectral metrics are applied to test images comparing the proposed algorithm with other well-known algorithms in the literature, all metrics show the superiority of our algorithm in its different cases. Also, Image Fusion at real time is very crucial to detect threats such as concealed weapon. Therefore, evaluation criteria of any fusion algorithm should include fusion time. We presented experimental results showing the superiority of the proposed algorithm w.r.t fusion time.

ACKNOWLEDGMENT

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Program.

REFERENCES

[1] Wei Liu, Wei Chen, "Recent Advancements in Empirical Wavelet Transform and Its Applications," Access IEEE, vol. 7, pp. 103770-103780, 2019.

[2] K J A S Sundar, V. Vaithiyanathan, "Design and Analysis of Fusion Algorithm for MultiFrame Super-Resolution Image Reconstruction using Framelet," Defence Science Journal, vol. 65, no. 4, pp. 292-299, 2015.

[3] K. J. A. Sundar, M. Jahnavi and K. Lakshmisaritha, "Multi-sensor image fusion based on empirical wavelet transform," 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), Mysuru, 2017, pp. 93-97.

[4] X. Lu and C. Xu, "Novel Gaussian mixture model background subtraction method for detecting moving objects," 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), Chongqing, China, 2018, pp. 6-10.

[5] Y. Zhang, "Generalized Wronskian Solutions for a Non-isospectral MKdV Equation," 2010 International Conference on Computing, Control and Industrial Engineering, Wuhan, 2010, pp. 65-67.

[6] Y. Wu, C. Chang and Y. Tao, "Closed-Circuit Television-Enabled Service: A Review of Security and Privacy Issues," 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops, Japan, 2013, pp. 306-309.

[7] Y. -L. Wu, Y. -H Tao, . C-J. Chang, and C. -W. Chang, Dilemma of CCTV adoption between security and privacy: The Taiwanese context, International Conference on Business and Information Conference, Sapporo, Japan, 2012.

[8] F. Conche, and M. Tight, Use of CCTV to determine road accident factors in urban areas, Accident Analysis and Prevention, vol. 38, 2006 pp. 1197-1207.

[9] S. Manchanda and S. Sharma, "Identifying moving objects in a video using modified background subtraction and optical flow method,"2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 129-133.

[10] Xu, Tuzhi, "Multisensor Concealed Weapon Detection Using the Image Fusion Approach" (2016). Electronic Theses and Dissertations.

[11] Gianmarco Cerutti, Bojan Milosevic, Elisabetta Farella, "Outdoor people detection in low resolution thermal images", 2018 3rd International Conference on Smart and Sustainable Technologies (SpliTech), pp. 1-6, 2018.

[12] G. Cerutti, R. Prasad and E. Farella, "Convolutional Neural Network on Embedded Platform for People Presence Detection in Low Resolution Thermal Images," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 7610-7614.

[13] M. Kivanc Mihcak, I. Kozintsev and K. Ramchandran, "Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising," 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258), Phoenix, AZ, USA, 1999, pp. 3253-3256 vol.6.

[14] D. Casasent, R. Patnaik, "Analysis of kernel distortion invariant filters", Proceedings of SPIE, vol. 6764, no. 2007, pp. 1.

[15] J. Luo and D. Chen, "Mapping Rules Based Data Mining for Effective Decision Support Application," 2008 International Seminar on Business and Information Management, Wuhan, 2008, pp. 506-509.

[16] A. Rao and K. Shah, "An optimized rule based approach to extract relevant features for sentiment mining," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 2330-2336.

The Extent to which Individuals in Saudi Arabia are Subjected to Cyber-Attacks and Countermeasures

Abdullah A H Alzahrani[1]
Computer Science Dept. Computing College at Alqunfuda
Umm Al Qura University, Makkah, Saudi Arabia

Abstract—**In light of the rapid development of technology and the increase in the number of users of the Internet via computers and smart devices, cybercrimes impact on enterprises, organizations, governments and individuals has been significant. Researches and reports on the impact of cybercrime and methods of prevention and protection are been introduced regularly. However, the majority focuses on the impact on organizations and governments. This paper aims to use a survey methodology in order to highlights the impact of cybercrimes on the individuals in Saudi Arabia and measure the awareness of cybersecurity among individuals. In addition, this research aims to investigate the common cybercrimes which target the individuals in Saudi Arabia and the countermeasure taken by them.**

Keywords—Cybercrimes; cybersecurity; identity theft; cyberattacks

I. Introduction

Cybersecurity can be defined as a prevention, protection, and restoration of information technology and data to ensure the availability, integrity, and confidentiality [1] – [3] while, Cybercrime is any crime that is committed using the Internet, for example, by stealing a person's personal or banking data or by infecting their computer with a virus [4]. A number of cybercrimes are widely addressed such as Identity theft, online shopping fraud, and phishing [5]. These crimes cost the individuals and organizations annually a high amount of money.

According to Cybint Solutions 2019 [6] [1] "It is expected that approximately $ 6 trillion will be spent globally on cybersecurity by 2021". This number highlights the importance of cybersecurity and issue of cybercrimes. This is due to the increasing number of cyberattacks and the more complicated they become. However, it is difficult to estimate the annual expenditure of an individual of cybercrimes [5].

This research focuses on the impact of cybercrimes on individuals in Saudi Arabia. Therefore, it is important to highlight the cybersecurity issues in this region. in 2017, around 12000 incidents of cyber-related were reported. While, it is estimated that the number of Internet users is around 27 million users of in Saudi Arabia [7]. Alongside with the aforementioned widely addressed cybercrimes, Communication and Information Technology Commission in Saudi Arabia pointed out a number of other cybercrimes in Saudi Arabia these are blackmailing, Privacy violation, and Unethical content [7].

According to McAfee Security Report 2014 [8] "In 2014, cyberattacks cost the kingdom of Saudi Arabia about 4 billion riyals, or about 0.17% of GDP (Gross domestic products)". However, this percentage is less than the acceptable percentage which is 2% of GDP. This makes Saudi Arabia in a non-alarming area of cybercrimes.

According to National Cyber Security Center (NCSC) in Saudi Arabia, in 2018 the threat alerts was higher by (13.5 %) s compared to the Fourth Quarter of 2017 [9][2]. NCSC revealed that the government and Education sectors were the main target of cyberattacks by around 71% of the attacks [9]. However, there is an insufficient of research or scientific investigation of cyberattacks on Saudi Arabia [10].

The importance of this research is derived from the fact that individuals are composing organizations. Individuals are clerks working in the terminal of the systems that are target for attacks, administrators who have high privileges on systems that are target for attacks, and technical who are working on the data of these systems. Moreover, sometimes these individuals are working on the systems from their personal devices. Therefore, neglecting investigating the impact of cyber-attacks on individuals is a gap that makes the picture of the impact of cyber-attacks incomplete. Furthermore, most of investigations (which will be shown in related work section in this paper) relies on the experts' opinions with orientation to specific domain of attacks. However, as it has been evident that individuals are the major cause of attacks [21], it is important to measure the impact of cyber-attacks on them as well as countermeasure they adopt.

In this paper, the impact of cybercrimes on individuals in Saudi Arabia is investigated. In addition, this research employs a survey methodology to collect data from individuals' respondents in order to provide an empirical result. The next section, will highlight a number of related studies. Then, research questions and research methodology will be explained in more details. Finally, the main finding of this research will be shown and discussed with some conclusions on them.

II. Related Work

Momein et al. 2010 [4] have conducted a survey based study to investigate the size and patterns of cybercrimes in Pakistan. The main findings were that the majority of

[1] https://www.cybintsolutions.com

[2] https://www.ncsc.gov.sa

cybercrimes to Internet users are related to privacy intrusion, sexual offenses, and e-commerce. The authors compared the international rate of the cybercrime impact on users with the local rate and offered several recommendations such as developing a national cybersecurity system and enforcement of a national cybercrimes law.

Bernik 2014 [11] carried out an analysis study of the organizations investment in protections from cybercrimes. The author followed a methodology that analyses a set of reports of costs of cybercrimes which are published by governments and organizations. This is to search for the actual causes of these costs. The author found that different models of calculating costs are implemented by organizations and that that the security experts are influencing the calculation processes. In addition, the author stated that the majority of the costs studies of the cybercrimes are exaggerated and unrealistic. Moreover, raising of organizational culture and awareness of protection is more efficient and inexpensive than other implementation of cybersecurity.

Riek et al. 2016 [5] believed that investigating impact of cybercrimes on individuals obtained insufficient attention with comparison to impact on organizations. The authors attributed this to those difficulties that may be encountered. Consequently, the authors have developed a survey methodology to study the costs of cybercrimes on individuals in 6 European counties. They focused on 7 different types of cybercrimes and found that Identity theft gained the highest impact on individuals in these 6 EU countries.

Kazmi et al. 2017 [12] investigated individuals' practices towards using internet banking. The research was conducted in three developing counties namely Saudi Arabia, Pakistan, and India. 1044 participants were asked to fill an online survey to measure their awareness of cybersecurity and internet banking threats. 272 participants were from Saudi Arabia. The authors found that there is a gap between the banks' expectations of individuals' practices and the individuals' actual actions. Consequently, the authors have introduced a two-part model that offers a set of security advices and instructions for the two parties (individuals – Banks).

National Cyber Security Centre (NCSC) in Saudi Arabia published a report in 2018 [9] that shows an empirical analysis of the cyber threats in Saudi Arabia. The report highlights the cyber threats for the years of 2017 and 2018. In addition, it has been noticed that the the cyber threats have increased since 2017 with 13%. As shown in Fig. 1, the majority of these threats targeted government and education sectors with percentage as 52% and 14% respectively. Furthermore, according to NCSC [9], malwares have been the observed to be the most common threats to all sectors. However, in the NCSC's report, the impact of cyber threats and attacks on the individuals was not highlighted in the investigation.

Another attempt of studying cybercrimes impact was carried out by Alelyani et al. 2018 [10]. The authors focused on the impact of cyberattacks on organizations in Saudi Arabia. In particular, the authors investigated the impact caused by Shamoon, Shamoon 2.0, and Ransomware on some well-known organizations such as Saudi Aramco. In addition, the aforementioned malwares were dissected. The authors relied on the results and solutions offered in [13]– [15] to provide a number of customized solutions and practices to be applied in the Saudi organizations. Raising awareness, heavy application of firewall, and decreasing the number of administrative accounts are the main solutions which are offered as customized solutions and practices to be applied in the Saudi organizations.

Harrell has published a series of investigations focusing on cybercrimes of identity thefts in US for years 2012, 2014, and 2016 [16] – [18]. In the latest one of investigations, the author studied the impact of identity theft on over 17.7 million persons whose data were brought from national authorities. It was found that the majority of surveyed persons have been impacted by financial losses and misuse of their data for other purposes. In addition, the majority of victims did not know how attackers collected their information. Furthermore, one out of ten of victims reported the incidents to authority. The author estimated the average financial losses of victims with 850$ and the median of 300$.

Important reports have been revealed by Symantec [19], [20] focusing on Internet Security Threat Report. In 2010, Symantec reported that an average of 260,000 identity theft or exposure is caused by a breach. In addition, although the majority of cybercrimes targeted government or organizations, individuals will be targeted in favor of identity thefts. Furthermore, Symantec estimated the individuals' financial loss associated to cybercrime with the around 100$ per year. Additionally, in 2019, Symantec revealed that more than 4,800 websites are exposed to attacks every month. Moreover, IoT (Internet of Things) devices and in particular smart phones are the favorite devices for attackers.

In conclusion, many studies have focused on the investigation of the impact of cybercrimes and attacks on governments and organizations. However, insufficient works have focused on the impact on individuals, in particular, in Saudi Arabia. In addition, the majority of these type of studies tends to offer expert opinions to mitigate the impact, whereas, a small number tends to study the actions taken by individuals against cybercrimes or attacks.

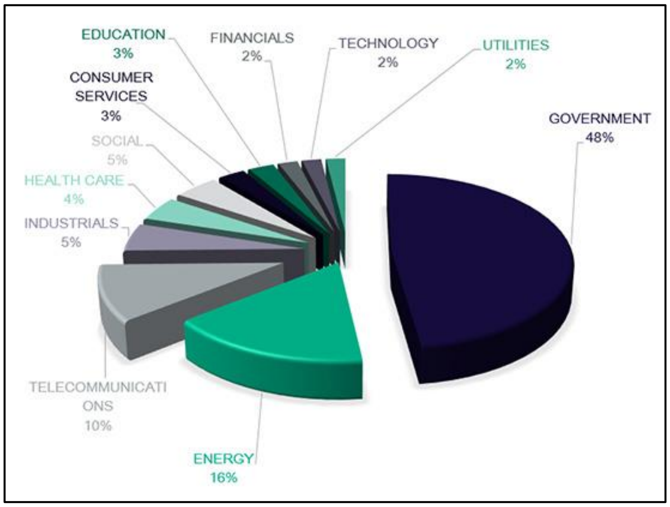


Fig. 1. Targeted Sectors According to NCSC [9].

The gap noticed is that the impact of cybercrimes on individuals in Saudi Arabia has not been sufficient covered. This includes the investigation of the actions and countermeasures which are applied by individuals in the case of cybercrime. In addition, it should highlight the matter of the individuals' knowledge of cybersecurity and protection methods. Although, the accurate estimation of the financial impact on individuals is difficult [5] in general, there has been no study that estimate a financial impact on individuals in Saudi Arabia. With all of the aforementioned, this research aims to fill this gap.

III. Research Questions

In order to achieve this research goal, three main research questions are framed. These questions are as follows:

RQ1. To what extent, is Internet network secure in Saudi Arabia? In order to answer this question, a set of factors was proposed that would imply the level of internet security in Saudi Arabia. These factors focus on knowledge of individuals of cybersecurity, and simple protection methods. Moreover, it was necessary to put some attention to the types and the use (public or private) of devices that are attacked when individuals are targeted.

RQ2. What effect does cybercrimes have on individuals in Saudi Arabia? To draw conclusion on this question, the experience of individuals with cybercrimes should be studied. This is in the light of questions focusing on the purposes of cyberattacks, recovery expenses of attacks, and attacks recurrences.

RQ3. What are the most common cybercrimes impacting individuals in Saudi Arabia and the causes of such attacks and what countermeasures are taken? To answer this question, the methods used in attacks should be investigated as well as the countermeasures against them. By carrying out such an investigation, a complete picture of the most common cybercrimes will be gained and the reactions taken by individuals will be highlighted. Therefore, respondents will be asked questions related to this as shown in the next sections.

IV. Research Methodology

Survey methodology has been adopted in this research. The survey was an online and links to it were sent to participants via emails and social media. This methodology is useful for gaining opinions from large number of respondents in Saudi Arabia. Thus, this study recruited 629 participants who are from different province, gender, qualification level, and work sector (public and private). The survey was divided into five main sections. The first section is intended to obtain the consent of participation and to gather some general information such as the province, gender, qualification level, and work sector.

The second section aims to identify the level of knowledge of participants in cybersecurity related topics. The third section focuses on measuring respondents' knowledge of cybersecurity related institutes and organizations in Saudi Arabia. Furthermore, it focuses on measuring their previous responses to these organizations, in addition to, their knowledge of the Cybercrime System in Saudi Arabia.

The fourth section is the main part of the survey and consists of 7 questions. These questions are to investigate the extent of effect of cybercrimes on individuals in Saudi Arabia. The questions cover the methods used in attacks, purposes of attacks, recovery expenses of attacks and countermeasures, and attacks recurrences. The fifth section focuses on the countermeasures taken by individuals when attacks occur. In addition, it uncovers the reactions of individuals towards attacks consequences.

Survey methodology has been adopted in this study in order to allow reaching an extensive number of respondents as fast as possible. In addition, a variety of tool supporting designing and spreading surveys are available such as Google Docs which has been used in this research. This allows conducting the research with cost efficiency and effectiveness.

The survey questionnaire was distributed online via emails and social media to 1000 respondents out of which 629 participated in the study and completed the questionnaire. The response rate of the survey is 61.9%. The participants of this research are individuals from Saudi Arabia who are using devices connected to the Internet. Out of the 629 participants, 70% are male and 30% are female. In addition, different levels of qualifications were noticed. The majority of the participants (50.40%) have Bachelor degree, while 33.23% are Postgraduate, 8.27% are High school diploma, 5.88% are Higher Diploma, and 2.23% hold Others education certificates.

The questionnaires were distributed to all of Saudi Arabia provinces. Participants were from all of the provinces as 63.59% of participants were from Makkah province, 13.99% were from Riyadh, 6.52% were from Eastern Province, 4.29% were from Madinah, 3.34% were from Asir, 2.23% were from Baha, 1.59% were from Jizan, 1.59% were from Tabuk, 0.95% were from Qassim, 0.64% were from Najran, 0.64% were from Northern Borders, 0.48% were from AlJawf, 0.16% were from Hail.

The majority of participants are employees who are working in both Government and Private sector as 69.79% are working Government sector and 13.51% are working Private sector. However, 16.69% of participants in this study stated that they are in Others sector of work. These might be students or unemployed people Finally, the ethical principles of research were followed in conducting the research keeping in mind the respect of the individual privacy and identity. Moreover, all data collected was for research purposes use only and participants were informed about this.

V. Results and Discussion

In this section, the results and their interpretation will be given. First the investigation results of the security of Internet in Saudi Arabia will be shown and discussed. Second, the results of the extent of effect of cybercrimes on individuals in Saudi Arabia will be illustrated and interpreted. Finally, the most common attacks and countermeasures of individuals in Saudi Arabia will be presented and discussed. This section aims to answer the research questions mentioned previously in this paper.

A. Security of Internet in Saudi Arabia

In order to study the effect of cybercrimes on individuals in Saudi Arabia, it is important to consider the security of the Internet in Saudi Arabia. In this research, it is not intended to evaluate technically the security of Internet in Saudi Arabia, rather than evaluating deductively from the users. Therefore, first, respondents were asked to express their thought of the Internet security in Saudi Arabia. As can be seen in Fig. 1 around 90% of Internet users believe that Internet is not secure. This result highlights the next question which draw attention on the knowledge of the users of cybersecurity, cybercrimes, and protection methods.

From Fig. 2, it is obvious that users have more knowledge in cybercrimes. However, they seem to know less about cybersecurity. Relatedly, increasing ignorance in protection methods can use to protect their devices and accounts from cybercrimes. Furthermore, surprisingly, around 65% of the respondents indicate very limited to limited knowledge in protection methods.

Protection method gives knowledge to probing practical questions. Therefore, respondents were asked about the basic methods of protection on their device and accounts. The focus was on the method of using Antivirus and regular change of passwords. Surprisingly, as shown in Fig. 3, around 60% of the respondents are not using an antivirus software. By having this practice, it can be understood that a considerable number of devices of individuals are exposed to attacks.

Moreover, Kazmi et. al. [12] described that a good practice of changing password is once every 3 months. However, Fig. 4 illustrates that 18% of respondents apply this practical method of protection. This leaves the majority of 82% jeopardizing their security in a simple way. Most importantly, is to highlight the high present of 43% of respondents are in an extreme risk of attacks as they never change their passwords.

Finally, with this result, it can be deduced that there is a considerable chance of attacks risks on individual's devices. However, the question of security of Internet network in Saudi Arabia rises. It can be concluded that security of Internet network in Saudi Arabia is competent. As with this results showing low level of individuals' knowledge of protection methods, around 67% of participants in this research have not experience cyberattacks. This will be illustrated and discussed more in the next section which focus on the extent of cyberattacks' effect on individuals.

B. Extent of Cybercrimes' effect on Individuals in Saudi Arabia

In this section, extent of cyberattacks' effect on individuals in Saudi Arabia will be illustrated and discussed. This will include the focus on respondents' experience with cyberattacks, knowledge of affected respondents of cybersecurity, purposes behind attacks, attacks recurrences, devices affected, and expenses related to cybersecurity.

It is essential to link results with each another. Therefore, in the previous section, the security level of Internet in Saudi Arabia was considered and it was concluded that it is competent level. This result was based on the low level of individual respondents' knowledge of cybersecurity and the high number of respondents who have not experienced cyberattacks. Fig. 5 illustrates that only 33% of respondents have experienced cyberattacks with around 14% of them are sure it was an attack and not only a device crash.

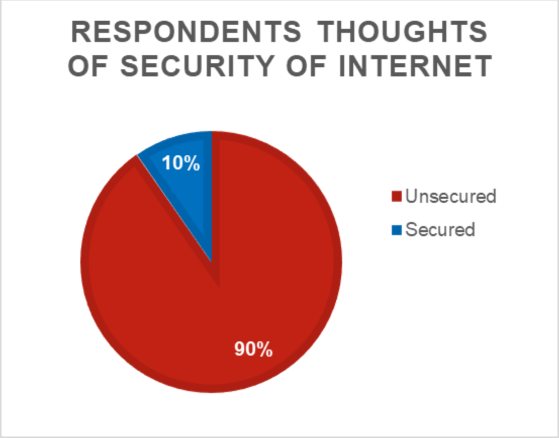


Fig. 2. Respondents thoughts of Security of Internet.

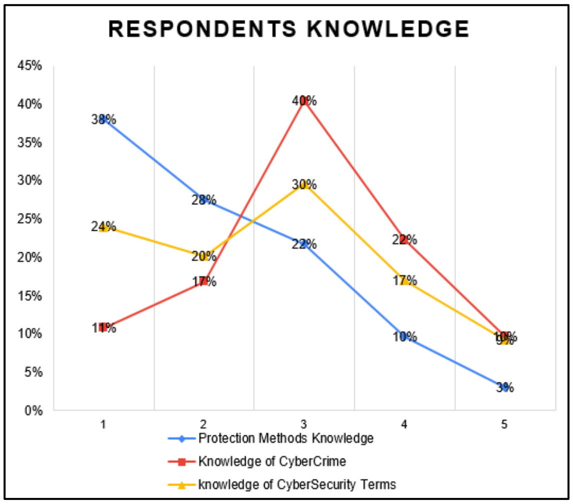


Fig. 3. Respondents Knowledge.

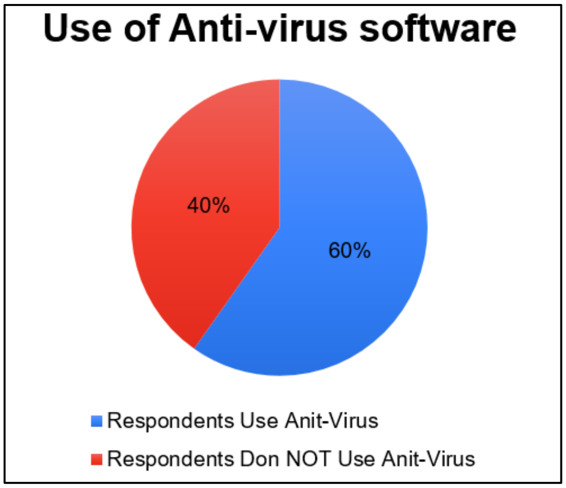


Fig. 4. Use of Anti-Virus Software.

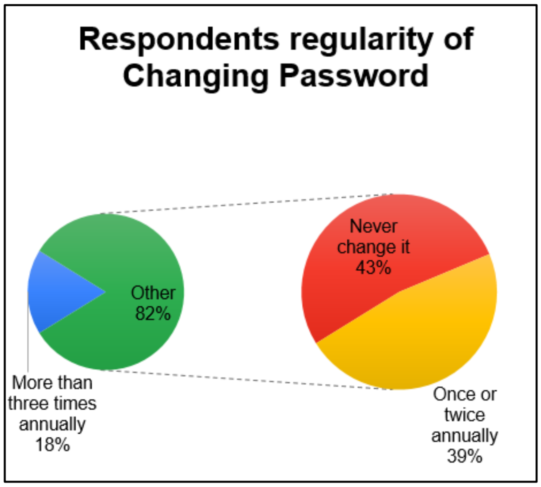


Fig. 5. Regularity of Changing Password.

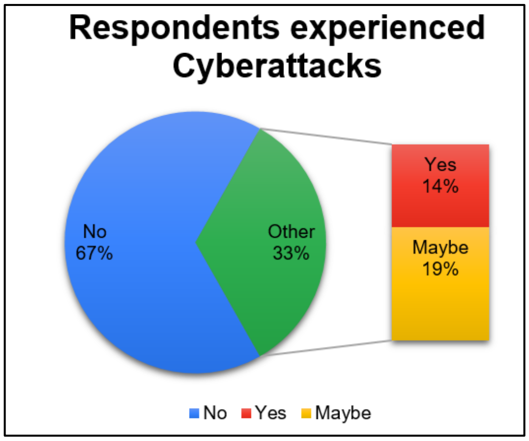


Fig. 6. Respondents Affected by Cyberattacks.

However, when it is about security, hesitated answers should be taking serious are attacks or threats, especially with the results of low knowledge of cybersecurity that is shown in the results previously. Therefore, in this research, affected respondents' percentage was determined to be 33% of total of 211 participants.

In order to investigate the extent of cyberattacks effects on individuals and to give a sharper focus to the study, the responses of affected people were separated. This is starting from the knowledge of this group of respondents in cybersecurity, cybercrimes, and protection methods. The results in Fig. 6 shows a similar result of the overall investigation results on the whole group of respondents as shown in Fig. 2. It seems that around 65% of the affected respondents of cyberattacks indicate very limited to limited knowledge in protection methods. This is along with the results that show users have more knowledge in cybercrimes, however, they seem to know less about cybersecurity.

It is reasonable to questions the clear effect cyberattacks had on the effected people. Fig. 7 illustrated that around 80% of affected respondents of cyberattacks in the study had their attitude toward using the Internet changed. They become more cautious and careful. This can be seen in Fig. 8 as 60% of them have not been attacked again. Importantly to notice in Fig. 8 is that 19% have been attacked again with different approach and for different purpose. This means that attackers either are seeking for different ways of attacks or new attackers with new approaches are introduced in the field.

In order to have sufficient information on the cyberattacks frequency or recurrence, effected respondents were asked to range the cyberattacks frequency for the past 5 years. Fig. 9 illustrates that 67% have been attacked less than 10 times. From such a result, it can be concluded here that the maximum 10 cyberattacks are impacting individuals. However, a considerable percentage of 22% are not sure about the number of times they have been attacked. This might be related to their low level of knowledge about the concepts cybercrimes and cybersecurity.

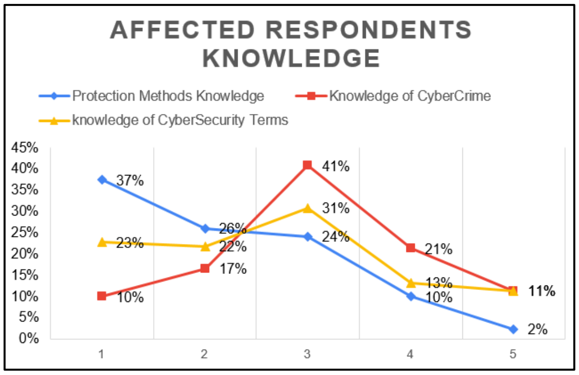


Fig. 7. Affected Respondents Knowledge.

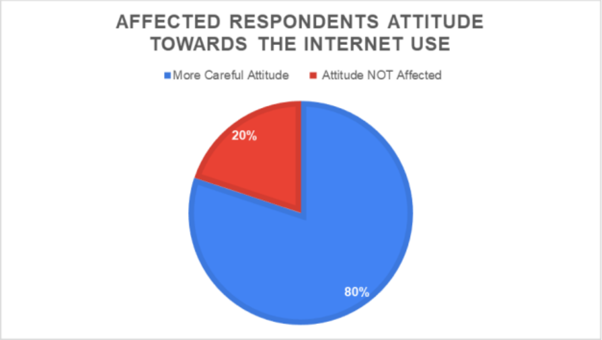


Fig. 8. Affected Respondents Attitude Towards the Internet use.

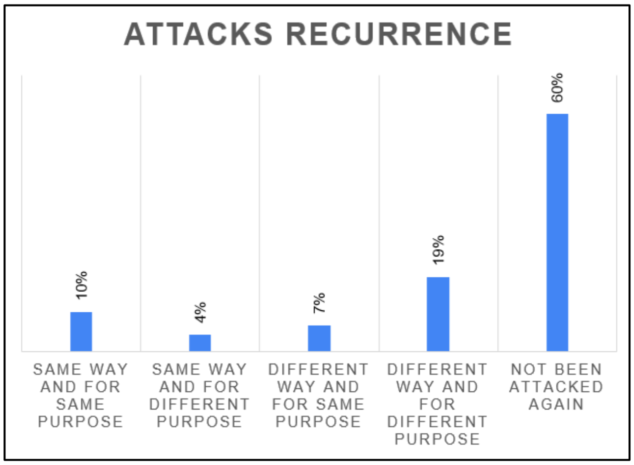


Fig. 9. Attacks Recurrence.

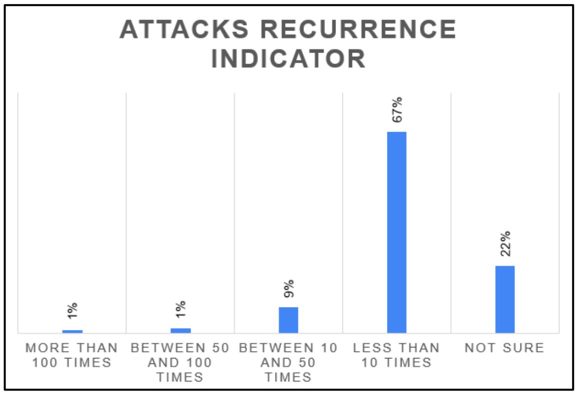


Fig. 10. Attacks Recurrence Indicator.

Investigating the knowledge, attitude after experiencing a cyberattacks, and and the frequency of attacks, leading the investigation towards the purposes of these attacks. Therefore, three categories of purposes were introduced to participants who are affected by cyberattacks. these categories are as follows: 1) identity theft; 2) blackmailing and money transfer; 3) damaging the devices and data. In addition, respondents were allowed to address their own interpretation of attacks purposes. However, an analysis of responses was conducted to check the respondents' entries to those aforementioned categories. The results of the analysis allow linking respondents' entries to those aforementioned categories.

From Fig. 10, it can be seen that around 42% of attacks are related to identity theft. This can be linked to possibility that attackers would use individuals as a breach to attack Saudi government sector, as a report published in [9] shows that 57% of attacks on Saudi network targeted government sector. Another possibility is that as Saudi Arabia is considered to be a wealthy country, identity theft would allow attackers to access and steal money. In addition, Fig. 10 illustrates that cyber-blackmailing gained a considerable amount of attention from attackers with that 29% of attacks experiences were linked to cyber-blackmailing.

Devices connected to Internet vary at the present. It is important to investigate the attacks are targeting which devices and whether these devices are for public or private use. Therefore, respondents experienced cyberattacks were asked to address the usual type and the purpose devices when attacks occur. From Fig. 11, it can be seen that around 51% of cyberattacks were experienced on Smart phones, however, closely 47% were experienced on computers. This is along with 90% of these devices are for private use as shown in Fig. 12.

Having 90% of cyberattacks on private devices leads to the question of individuals' expenditure on cybersecurity which means all spending on assurance of security such as antivirus software license etc. In addition, expenditure on cybersecurity can include dealing or recovering costs of cyberattacks. By addressing cybersecurity and expenses, it is reasonable to allow all respondents to participate in ranging their spending on it.

In this research, annual basis was opted. Therefore, respondents were offered three main categories of expenditure as follows: 1) Less than 500 Saudi Riyal; 2) Between 500 and 2500 Saudi Riyal; 3) More than 2500 Saudi Riyal. Fig. 13 shows that 64% of respondents spend maximum of 500 Saudi Riyal. This comes to equality of around 134$ (US dollar).

In conclusion, the main findings of the effects of cybercrimes on surveyed individuals in Saudi Arabia can be summarized as 33% of surveyed individuals have experienced cyberattacks on their private use smart phones or computers. In addition, the dominant cybercrime is Identity Theft with 42% of respondents. Although, the majority of 65% of attacked individuals have limited to very limited knowledge of protection methods, 60% of the attacked individuals have not experienced cyberattacks again, whereas 19% of them have been attacked again with different ways of attacks and different purposes of cybercrimes. In addition, 67% of surveyed effected respondents articulated that they have experienced cyberattacks less than 10 times. Finally, 64% of all surveyed individuals spend less or equal 500 Saudi Riyal (134$ US dollar) annually for cybersecurity reasons.

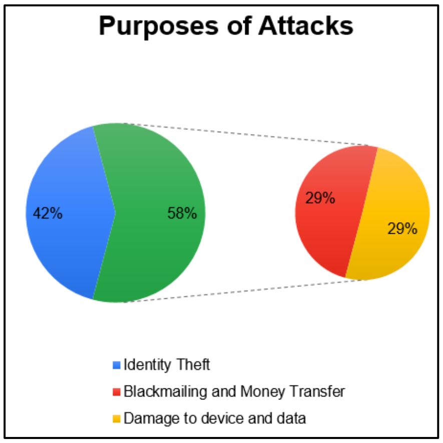


Fig. 11. Purposes of Attacks.

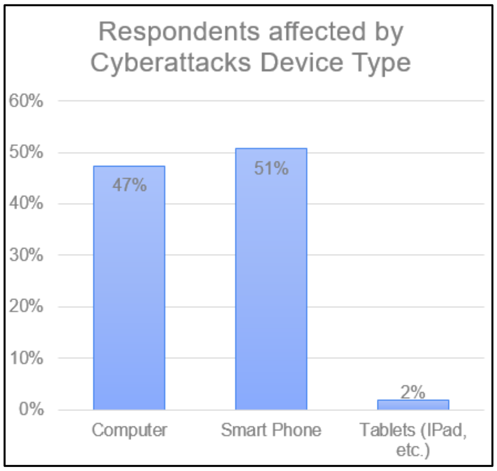


Fig. 12. Device Type.

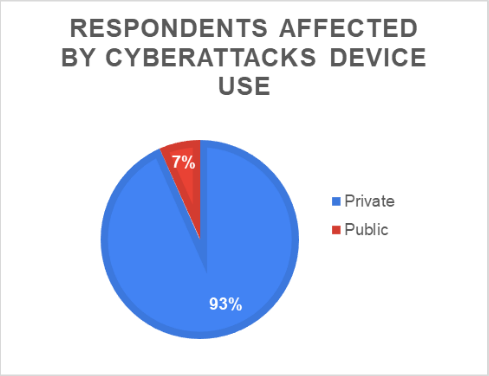


Fig. 13. Device use.

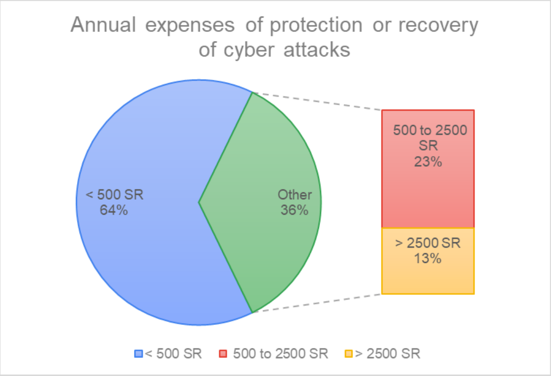


Fig. 14. Annual Expenses.

C. Most Common Cyberattacks Methods on Individuals

In this section, most common cyberattacks types on individuals will be investigated and discussed. In addition, reactions or countermeasures taken by the individuals will be studied. Finally, the linkage of each reaction to the attack methods will be made and discussed. This is to have a general view of the common cyberattacks types and the individuals' reactions and their priority to them.

First, Effected individuals have been asked to state the methods attackers used to them. From surveyed individuals' responses shown in Fig. 14, it can be seen clearly that the common attacks method on individuals is Tracking Ads links with 27% of responses. However, Surfing Untrusted Website, Downloading Unreliable Software, and Opening Unknown Senders' Emails gain a considerable attention of respondents with 21%, 19% and 18%, respectively.

Second, it is important to study the individual reactions toward any cyberattacks that might happen to them. Therefore, Fig. 15 shows the reactions of all and affected individuals. Three possible reactions respondents addressed as follows: 1) Format device and reset it to factory status; 2) Seek for an advice from a friend or a specialist; 3) Report to authority. As can be seen in Fig. 15, Seeking for an advice and Formatting device seems to be at the high priority of the individuals. However, individuals who have experienced cyberattacks seem to give more attention to actions like Formatting device. Interestingly, all individual respondents appear to deprioritize the action of reporting to authority to allow investigating on the cybercrime.

Third, it is meaningful to link each reaction to the attack methods. As can be seen in Fig. 16, individuals who have experienced cyberattacks tend to choose Formatting Device for all attacks in general. However, it seems that when an attack is suspected to be from an untrusted website, individuals tend to prioritize Seeking for an Advice over Formatting Device. Noticeably, Report to authority seems to be at the least priority to the individuals.

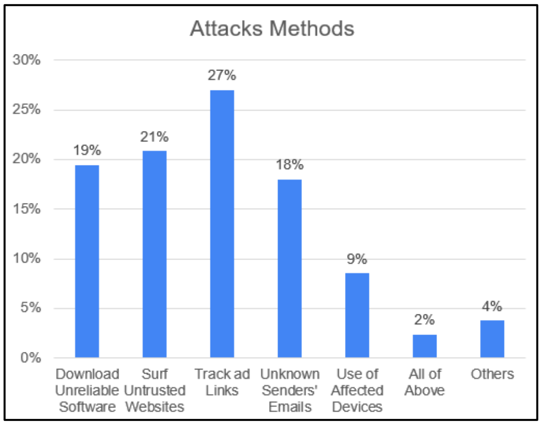
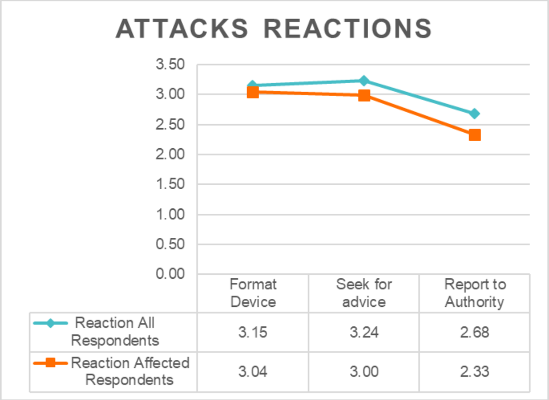


Fig. 15. Attacks Methods.



	Format Device	Seek for advice	Report to Authority
Reaction All Respondents	3.15	3.24	2.68
Reaction Affected Respondents	3.04	3.00	2.33

Fig. 16. Attack Reactions (Results is based on Average).

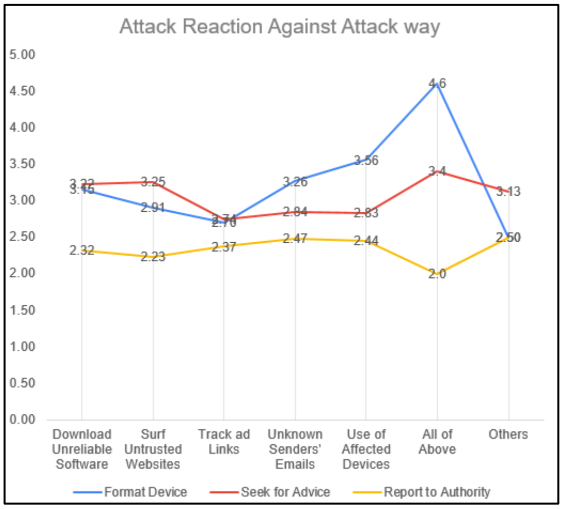


Fig. 17. Attack Reaction against Attack Method (Results is based on Average).

In conclusion, as can be seen in Fig. 17, Tracking Ads links seems to be the most common cyberattack methods that surveyed individuals have encountered. However, Surfing Untrusted Website, Downloading Unreliable Software, and Opening Unknown Senders' Emails are candidate methods used by attackers. In addition, reactions and countermeasures Seeking for an Advice, Formatting Device, and Report to authority are the common actions taken by individuals who either have or not have experienced cyberattacks.

VI. CONCLUSION

In the research the impact of cybercrimes on individuals in Saudi Arabia has been investigated. A number of findings were emerged. First, an undeniable ignorance of cybersecurity is noticed in individuals. In addition, Saudi networks is a competent secure network. This has been justified by the number of individuals who experienced cyberattacks with comparison to the low level of knowledge of protection methods. Furthermore, unlike organizations, individuals are not targeted by cyberattacks as the majority of 67% of respondents expressed that they have not experienced cyberattacks. However, they might be used to be the breach to organizations as Identity theft is the main purpose of cyberattacks on individuals.

Second, identity theft is the most common purpose of cyberattacks that targeted individuals. Furthermore, it was found that the majority of 64% of respondents spend less than 500 Saudi Riyal (134$) annually on cybersecurity which is less than the average annual spending of a US citizen. In addition, cyberattacks recurrences on individuals are less than 10 times over 5 years.

Finally, Tracking Ads links is the most comment approach used to attacks individuals, while, the Formatting device countermeasure is the popular action taken by individuals for most cyberattacks. Noticeably, report to authority action seems to be at the least priority to the individuals when attacks occur.

VII. FUTURE WORK

As the number of internet users is increasing in Saudi Arabia with 26 million as current number of user, more participants are needed in order to have more accurate and reliable conclusions. In addition, studying the impact on different age categories will enrich the research and provide clustering of responses to gain more conclusions. Finally, the financial impact on individuals has been estimated using one factor which is the participants' data. More resources are needed such as data from NCSC and national authority of cybersecurity.

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my University (Umm Al Qura University) which gave me the golden opportunity to do this research on the topic Software Engineering. Secondly I would also like to thank my friend (Dr Ahmad Albasheri) who helped me a lot in finalizing this research within the limited time frame.

REFERENCES

[1] Paulsen and P. Toth, 'Small business information security', US Dep. Commer. Doi, vol. 10, 2016.

[2] CNSS Instruction (CNSSI), 'Committee on National Security Systems', 2015.

[3] T. A. Johnson, 'National Security Presidential Directive/NSPD-43 Homeland Security Presidential Directive/HSPD-14', in National Security Issues in Science, Law, and Technology, CRC Press, pp. 651–654 , 2007

[4] F. A. Momein and M. N. Brohi, 'Cyber crime and internet growth in Pakistan', Asian J. Inf. Technol., vol. 9, no. 1, pp. 1–4, 2010.

[5] M. Riek, R. Böhme, M. Ciere, C. Gañán, and M. van Eeten, 'Estimating the costs of consumer-facing cybercrime: A tailored instrument and representative data for six EU countries', in Workshop on the Economics of Information Security (WEIS), University of California at Berkeley, 2016.

[6] D. Milkovich and Cybint Solutions, '15 Alarming Cyber Security Facts and Stats', 23-Sep-2019. [Online]. Available: https://www.cybintsolutions.com/cyber-security-facts-stats/., 2019

[7] Communication and Information Technology Commission, 'Digital security and user protection of Internet risk', 2018.

[8] Center for Strategic and International Studies and McAfee, 'Net losses: estimating the global cost of cybercrime: economic impact of cybercrime II', Jun. 2014.

[9] Saudi National Cyber-Security Authority, '2018 First Quarter Statistical Report about Cyber Threats and Risks', 2018. [Online]. Available: https://www.ncsc.gov.sa/wps/portal/ncsc/home/Reports/!ut/p/z1/hY5dC4JAEEV_jc8z64r5uhW4ZhAWmM2LbIvUhq5Z0se_z0V6tObpDPdyOUBQAFn1MCfVm9aqevgPFJaMLSPpx7iO5jOBGWKeiCTlWRrC_l-BhhgnTiDs1A1WYykJAskCP91sXcRzKXm0YBhzZ2EuXUcCSLe2r149FFbfdW2OHjrw8Nw21cilw7Kygxr9Hva_hWm9a1M83xzNB_MOTj8!/dz/d5/L2dBISEvZ0FBIS9nQSEh/. [Accessed: 11-Oct-2019]., 2018

[10] S. Alelyani and H. Kumar, 'Overview of Cyberattack on Saudi Organizations', J. Inf. Secur. Cybercrimes Res. JISCR, vol. 1, no. 1, 2018.

[11] I. Bernik, 'Cybercrime: The Cost of Investments into Protection.', Varstvoslovje J. Crim. Justice Secur., vol. 16, no. 2, 2014.

[12] Z. Kazmi, J. M. Alghazo, and G. Latif, 'Cyber Security Analysis of Internet Banking In Emerging Countries: User and Bank perspectives', presented at the 2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS), pp. 1–6. , 2017

[13] Shaunak and Gregory Paul, 'Detailed threat analysis of Shamoon 2.0 Malware', vinransomware, Feb-2017. [Online]. Available: http://vinransomware.com/blog/detailed-threat-analysis-of-shamoon-2-0-malware. [Accessed: 28-Oct-2019]., 2017

[14] Devika Jain, 'Shamoon 2: Back On the Prowl', NSFOCUS, Inc., a global network and cyber security leader, protects enterprises and carriers from advanced cyber attacks., 2017. [Online]. Available: https://nsfocusglobal.com/shamoon-2-back-on-the-prowl/. [Accessed: 28-Oct-2019]., 2017

[15] Codymercer, 'StoneDrill – Shamoon & Shamoon 2.0 Variant', 2017. [Online]. Available: https://blog.nsfocusglobal.com/categories/stonedrill-shammon-shammon-2-0-variant/. [Accessed: 28-Oct-2019],2017.

[16] E. Harrell and L. Langton, 'Victims of Identity Theft, 2012', US Department of Justice, Office of Justice Programs, Bureau of Justice, 2013.

[17] E. Harrell, 'Victims of Identity Theft, 2014', US Department of Justice, Office of Justice Programs, Bureau of Justice, 2015.

[18] E. Harrell, 'Victims of identity theft, 2016', US Department of Justice, Office of Justice Programs, Bureau of Justice, 2019.

[19] Symantec, 'Internet Security Threat Report (ISTR)', Symantec, 24, 2019.

[20] M. Fossi et al., 'Symantec internet security threat report trends for 2010', Symantec, 2011.

[21] M. Ovelgönne, T. Dumitraş, B. A. Prakash, V. S. Subrahmanian, and B. Wang, 'Understanding the relationship between human behavior and susceptibility to cyber attacks: A data-driven approach', ACM Transactions on Intelligent Systems and Technology (TIST), vol. 8, no. 4, pp. 1–25, 2017.

Towards Social Network Sites Acceptance in e-Learning System: Students Perspective at Palestine Technical University-Kadoorie

Mohannad Moufeed Ayyash[1*], Fadi A.T. Herzallah[2], Waleed Ahmad[3]
Department of Business Administration and E-commerce
Palestine Technical University, Kadoorie
Tulkarm, Palestine

***Abstract*—This study aims to examine social network sites acceptance in an e-Learning system and to propose a model encompassing determining factors that affect students' intentions to use social network sites in the e-Learning system. The proposed model was built based on the Technology Acceptance Model (TAM), perceived enjoyment, social influences, and perceived information security from the literature review. The quantitative method of data collection using a questionnaire survey is used in the current study. The data analysed using a structural equation modeling (SEM) approach through partial least square (PLS) software version 3. The results indicated that perceived ease of use, perceived usefulness, perceived enjoyment, social influences, and perceived information security has a significant and positive impact on student's acceptance of social network sites in an e-Learning system at Palestine Technical University-Kadoorie. Theoretical and practical implications discussed.**

Keywords—Social network sites; e-Learning system; perceived usefulness; perceived ease of use; perceived enjoyment; social influence; perceived information security; Palestine

I. Introduction

Social network sites are becoming a stable part of each individual's life, especially student societies [1]. It manages to transform the lifestyle of youth people while becoming one of the natural means of communication and entertaining [2]. Social network sites facilitating and creating knowledge sharing and eventually moving a speaking into a discussion, for example, by an institution to customers [3], [4]. Using social network sites in e-Learning has improved students study methods. Therefore, numerous academic institutions have exploited opportunities to use social network sites as a novel technology to cultivate their learning provision and stay competitive by attracts new students [5], [6]. In developing countries, e-Learning has the potential to snowballing demands for education as well to relieve a decrease in the competent tutors [7].

Furthermore, in the last era, research efforts have been augmented so that successful perceptions of social network sites use for learning undertakings could emerge [2]. Some researchers like [7], [8], and [9] found the positive effect of social network sites on the learning process that leading to a higher level of performance. On the other hand, the utilitarian nature of social network sites is still unclear [7], [10]. Also, despite the numerous possible advantages of including social network sites into learning, there is a massive difference between the level of positive perceptions of social network sites and the level of applied usage [11]. According to previous research, there is a dearth of studies on users' perspective on similar technologies in developing countries [12], more precisely in the Arab countries [13], [14]. Consequently, there is a need to assess the factors that can affect the use of social network sites in Arab countries [15].

The present study is considered a notable exertion since it determines factors impacting social network sites acceptance in e-Learning in Palestine as a developing country. Therefore, the objective of this study is to suggest a model for social network sites acceptance encompassing determining factors affecting students' intentions to utilise social network sites in the e-Learning system. Hence, the determination of the factors that can affect social media networks acceptance in e-Learning can help universities in enhancing students learning quality. Therefore, the researcher's stimulus to conduct this study includes the dire need for shaping the factors that affect student's acceptance of social network sites in e-Learning to meet the pioneering learning development. Hence, the critical question of this study is: what are the factors that are affecting student's acceptance of social network sites in the e-Learning system at Palestine Technical University-Kadoorie?

II. Literature Review

A. e-Learning in Palestine

The progression of Information and Communication Technology (ICT) makes this world a global village [16]. Therefore, education institutions are acceptance the cutting-edge (ICT) to be in line with the global development in education systems. The progress growth in ICT possibly increases creativity between students through e-Learning. Nevertheless, the implementation of ICT is still in its infancy stage across most of the Arab world [15]. Also, in the context of developing countries, e-Learning still faces a high rate of failure [15], [17]. Accordingly, Palestinian educational institutions are required to cover all difficulties faced by students in the acceptance of e-Learning.

Palestine is characterised as a fighting area in the Middle East district with mobility constraints [16], [18]. Palestine is separated into small portions divided by protection boundaries

[*] Corresponding Author

that harmfully impact the education scheme [18]. Education sector expansion faces numerous hindrances due to the constant of Palestinian Israeli conflict. Therefore, e-Learning is a priority instead of a luxury to develop the quality of learning for students in Palestine [16]. Consequently, technology incorporation is the most significant implication in conflict zones, with a lack of learning resources [17]. Palestinians consider the development of technology as a crucial tool for their existence, justifying their daily complications, enabling the crises, improving the equity of using technology between students in both public and higher learning [16], [18] [19]. Therefore, this research examines social network sites as an effective e-Learning tool, through determining the factors affecting student's acceptance of social network sites in e-Learning in Palestine at Palestine Technical University-Kadoorie.

B. Social Network Sites Acceptance and e-Learning

Social network sites have a profound influence on the communications way these days [20]. Social network sites offer desirable capabilities, comprising send messages, knowledge share, information access, research, and chat [21], [22]. Elkaseh [23] stated that social network sites usage in universities could have a positive influence on learners' educational results. Moreover, convinced researchers such as [24], [8] indicated that the use of social network sites in writing homework had a positive effect on their degree of education.

Owing to the introduction of ICT, it is now imperative that educational institutions around the world, especially in developing countries such as Palestine leverage social network sites usage in e-Learning. As well as, it has been noted that converting from traditional education to e-Learning has augmented in the present-day. Moreover, recent research has revealed a high level of willingness amongst students to use social network sites in learning [25], [26]. Therefore, this study focuses on the factors affecting a student's acceptance of social network sites in the e-Learning system in Palestine.

The rest of the paper is organised as follows: Section II presents the study literature review that revealed e-Learning in Palestine and social network sites acceptance, Section III reviews a theoretical foundation of the study, Section IV presents research model and hypotheses development, Section V presents method of the study, Section VI presents data analysis and results, Section VII discusses data analysis and findings. Finally, we conclude in Section VIII.

III. THEORETICAL FOUNDATION

A. Technology Acceptance Model

TAM is the most models used in predicting and explaining user's attributes that affect his/her acceptance behaviour to pioneering technologies because of its strength, ease, and suitability [27]. In TAM, perceived usefulness and perceived ease of use determine a person's perspective on using technology and drive a person's intention to accept the technology [28]. Additionally, the acceptance of information technology based on the user's perceived usefulness and of perceived ease of use of these technology [29], [30]. In brief, TAM main idea is summarised in individuals accept or reject information technology based on his belief that using this technology can help in completing jobs more successfully (perceived usefulness) and in less exertion (perceived ease of use) [30]. Overall, users' awareness perceived usefulness, and perceived ease of use are the critical factors in construction technology acceptance models [30], [31]. Additionally, before an understanding of the advantages of new technology, researchers need to recognise how students use them through TAM [32]. Therefore, the researchers need to understand student's acceptance of novel technology such as social network sites in the e-Learning system.

B. Perceived usefulness and Perceived Ease of use

Perceived usefulness is the degree to which the user believes that using a specific system is useful and will enrich performance [30]. In contrast, perceived ease of use refers to the degree to which a user believes that using a specific system is easy and not require much effort [30]. Previous studies found that perceived ease of use has positive effects on behavioural intentions to use [33], [34]. Perceived usefulness and perceived ease of use have been scrutinised to predict behavioural intentions to use social network sites [35]. Moreover, numerous research applied TAM in e-Learning and found that perceived usefulness and perceived ease of use have a significant influence on user behavioural intentions to use e-Learning [36]. Therefore, perceived usefulness and perceived ease of use are essential factors that affect student's intentions to use social network sites in e-Learning. Hence, it postulated that:

H1. Perceived usefulness will positively affect student's acceptance of social network sites in e-Learning at Palestine Technical University-Kadoorie.

H2. Perceived ease of use will positively affect student's acceptance of social network sites in e-Learning at Palestine Technical University-Kadoorie.

C. Perceived Enjoyment

Technology acceptance can be affected by the perceived enjoyment [37]. According to indicate how students perceive the diverse activity or services to be entertaining in them irrespective of any results that may be expected [38]. In the context of learning, a learner's particular feelings of enjoyment, inclination, and feeling play critical roles in clearing up user acceptance of e-Learning [39]. User behavioural intention concerning social network sites is often determined by the degree of perceived enjoyment when using social network sites [40]. Furthermore, enjoyment is a factor that determines the intention of users to use social network sites [41]. In this study, perceived enjoyment means that students use of e-Learning system through social network sites in a method that develops their learning competencies. Nevertheless, the student's perceived enjoyment through the use of social network sites is being scrutinised [42]. Therefore, it postulated that:

H3. Perceived enjoyment will positively affect student's acceptance of social network sites in e-Learning at Palestine Technical University-Kadoorie.

D. Social Influence

Social influence is an important factor affecting online user behaviour [43]. Furthermore, social influence affecting the user's intentions to use information technology [32], [44]. Therefore, social influence has existed extensively scrutinised in the information systems study. Wang [45] defined social influence as individual user's attitudes, beliefs and behaviour's are influenced by exhortation others. According to Maryam [46], social influence has positive effects on the user's intention to accept novel technologies such as social network sites. Based on the above discussion, it postulated that:

H4. Social Influence will positively affect student's acceptance of social network sites in e-Learning at Palestine Technical University-Kadoorie.

E. Perceived Information Security

Perceived information security plays a substantial role in information technology adoption and usage [47]. Luo et al. [48] defined perceived information security as individuals believe that private information will not be watched, kept and tamper throughout work sessions by illegal gatherings in a way reliable with their confident anticipation. Generally, the users of social network sites supply a lot of their individual information online, and the threat that this information could be misused is high in social network sites [49]. Moreover, perceived information security considered as the primary concern between users' through online existence [50], [51]. Therefore, perceived information security appears to lead students to leave the use of new technology, at which point e-Learning may be a failure. Therefore, educational institutions should take this dangerous concept into account to ensure students usage of the e-Learning system. Consequently, perceived information security is an essential factor affecting student's acceptance of social network sites in e-Learning. Based on the above discussion, it postulated that:

H5. Perceived information security will positively affect student's acceptance of social network sites in e-Learning at Palestine Technical University-Kadoorie.

IV. Research Model

Social network sites are an important tool for business organisations, so, it is crucial for education institutions to use this tool in learning. It is essential to realise that educational institutions need a strategy to increase their enrolment and enhance learning. Previous studies showed that students have a high level of willingness to use social network sites in learning [26], [27]. Social network sites are able to improve student's ability to achieve in the fulfillment of learning activities [31], [52]. The acceptance of social network sites have been scrutinised by various researchers using one of the numerous existing acceptance models [47]. Nevertheless, e-Learning adoption needs to be considered in the context of an information system [53]. Likewise, TAM has been mostly used in literature to ratify information systems adoption [53], [54], [55].

The emerging studies in the arena of social network sites are trying to identify the factors that help the users to accept social network sites as a novel technology in a different context such as e-Learning. However, according to Maryam [47], there is limited research in the acceptance of social network sites. Furthermore, no research has yet focused its caution on social network sites acceptance in e-Learning in Palestine via the combination of perceived usefulness, perceived ease of use based on TAM, and perceived enjoyment, social influence, and perceived information security based on literature review. Based on these aspects, this study derives the theories forming the base of the creation of the research model for social network sites acceptance including ascertaining factors that impact intention to use social network sites in the e-Learning at Palestine Technical University- Kadoorie. The research model is illustrated in Fig. 1.

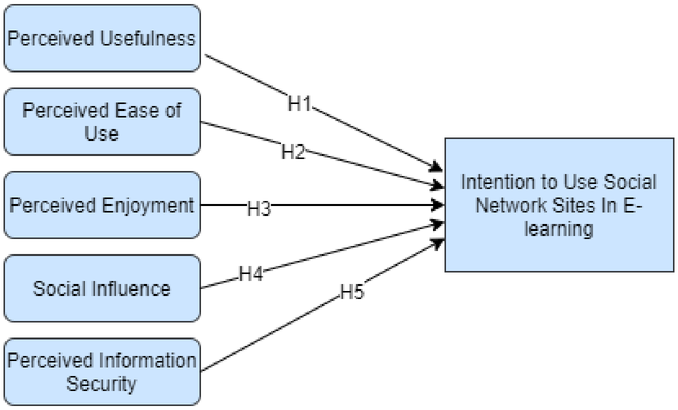


Fig. 1. Research Model.

V. Method

The aim of this study is to examine social network sites acceptance in an e-Learning system and to recommend a model for social network sites acceptance encompassing determining factors impacting students' intentions to use social network sites in the e-Learning system. This research sample involves the students of Palestine Technical University-Kadoorie under three branches in Palestine. The quantitative method of data collecting using a questionnaire survey is used in the present study. 370 Palestinian students were selected as the samples of the study using stratified random sampling. Regarding the number of the samples, Structural Equation Modeling (SEM) analysis, need at least 100 samples [56]. Also, using smart PLS path modeling recommends that the size of the sample should be a least 30 to 100 cases [56]. So, 370 respondents are satisfactory.

The survey used for data collection was adopted from earlier studies on the context of the information system. It includes 30 items adapted to the context of this study. The items used to measuring perceive enjoyment adopted from [57], [58], and [59]. The items measuring perceived usefulness, perceived ease of use and intention to use social media were adopted from [30], [60], and [61]. The items were measuring social influence adopted from [62], [63]. The items to measure perceive information security adopted from [47]. The respondent requests to choose one of the five-point Likert-scale ranging from (1) strongly disagree to (5) strongly agree. The online survey dispersed to students in Palestine Technical University-Kadoorie in West Bank, Palestine. The questionnaire survey also translated into the Arabic language as a native language of students to increase their understanding and respond efficiently. However, only 350 questionnaires are for data analysis. The data were analysed using the SEM approach through smart PLS 3 software.

VI. Data Analysis and Results

The data analysis in this study was employed using SPSS version 22 and Smart PLS version 3. The SPSS was used to obtain the descriptive statistics of the sample while the investigation of the latent variable within the causal structure was employed using Smart PLS. The statistical analysis results are presented in the next subsections.

A. Descriptive Statistics

The features of the suspects and the constructs' descriptive statistics are presented in Table I and Table II, respectively. From Table II, the range of the mean of the entire construct was from 3.20 to 4.17, with standard deviations ranging from 0. .64 to .78. This showed a narrow spread of the values around the mean. Furthermore, the skewness values ranged from -.950 to -.159, while the kurtosis values ranged from -.268 to 1.839. The rule of thumb for skewness and kurtosis as established by Byrne [64] for normally distributed data are ± 3 and ±7, respectively and based on this recommendation, the data for this study were assumed to be suitable and regular for further analysis.

TABLE. I. Sample Characteristics

Sample Characteristics	**Items**	**Frequency**	**Percent (%)**
Gender	Male	64	18%
	Female	286	82%
Age	18-24 years	331	95%
	25-35 years	9	3%
	36-44 years	8	2%
	45 and above	2	1%
Education	Diploma	56	16%
	Bachelor	278	79%
	Master	16	5%
Branch	Tulkarm	268	77%
	Ramallah	45	13%
	Hebron	37	11%
The most used social network sites	Facebook	146	42%
	WhatsApp	85	24%
	YouTube	17	5%
	Instagram	83	24%
	Telegram	2	1%
	Twitter	3	1%
	Snapchat	13	4%
	Others	1	0%
Time spent on social network sites per day	Less than 1 h	10	3%
	2-3 h	86	25%
	4-5 h	121	35%
	6 h and above	133	38%

TABLE. II. Descriptive Statistics

Construct	**INTU**	**PEOU**	**PE**	**PIS**	**PU**	**SI**
Items	5	5	5	5	5	5
Mean	3.860	4.170	3.880	3.200	3.660	3.470
S.D.	0.780	0.560	0.660	0.780	0.580	0.640
Skewness	-.680	-.950	-.720	-.1590	-.279	-.562
Kurtosis	.986	1.839	1.331	-.268	.814	1.375
Cronbach's alpha	0.818	0.830	0.859	0.813	0.730	0.763

The measurement instrument was checked for internal consistency using Cronbach's Alpha, and the result of the assessment showed all the constructs to obtain Cronbach's Alpha values of more than 0.60, indicating their high internal consistency [65]. These values showed a good correlation between the collections of items replies used to measure the study constructs [66].

B. Model Analysis

The partial least squares version 3 (PLS 3) was used during the model verification to analyse the data. There are two stages in structural equation modeling using PLS; these are measurement model assessment and structural model assessment [67]. During the measurement model assessment stage, the constructs are examined for reliability and validity, while structural model assessment focuses on the verification of the model hypotheses.

1) Measurement model result: Hair [67] stated that the verification of the survey for the measurement model is an aspect of the PLS technique. This process is often implemented based on formative and reflective constructs. The goodness of measures is tested using two significant criteria - reliability and validity. Reliability implies testing the consistency of a given proposed instrument in measuring a specific aspect it was intended to measure. At the same time, validity implies testing the efficiency of a given instrument in measuring a specific concept it was designed to measure [68]. This study employing the three-elements procedure to assess the measurement model: Indicator items reliability, convergent validity, and discriminant validity. As per Hair [67], the minimum acceptable level of item loading must be > 0.60; however, for this study, the minimum acceptable item loading level was 0.60.

As depicted in Fig. 2, 30 reflective indicators were used to test the measurement model. Also, three items (PU2, INTU1, PIS5) were found to have a factor loading of 0.594 and 0.671, respectively. According to Hair [69], any item with a factor loading value in the range of 0.40 to 0.70 should be excluded as far as its exclusion will improve the composite reliability result (CR) above the recommended threshold value. Hence, the removal of the indicators was done in this study by performing the PLS algorithm test.

The Average Variance Extracted (AVE) was used as the basis to test the convergent validity for each construct shown in Table III. Convergent validity is a measure of the extent of the positive correlation between a measure and the other measures of the same construct [67]. This study adopted 0.5 as the minimum acceptable AVE value as earlier suggested by Hair [67]. The results proved that “Perceived Enjoyment” presented the highest AVE value (0.647) while “Social Influence” presented the lowest AVE value (0.516). These are all acceptable values concerning their convergent validity.

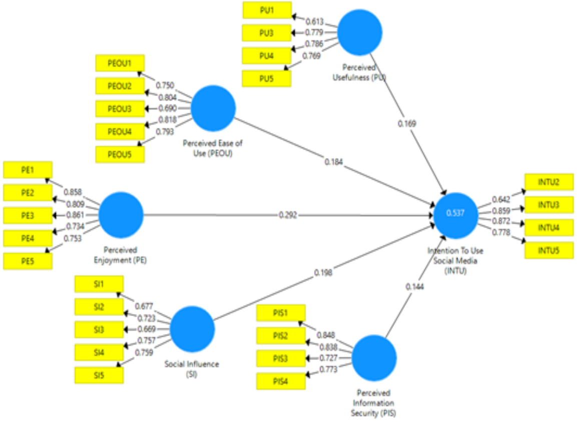


Fig. 2. Measurement Model.

TABLE. III. RELIABILITY AND VALIDITY RESULTS

Construct	Items	Factor	Construct	Items	Factor
Intention to Use (INTU)	INTU2	0.642	0.629	0.870	0.809
	INTU3	0.859			
	INTU4	0.872			
	INTU5	0.778			
Perceived Enjoyment (PE)	PE1	0.858	0.647	0.901	0.870
	PE2	0.809			
	PE3	0.861			
	PE4	0.734			
	PE5	0.753			
Perceived Ease of Use (PEOU)	PEOU1	0.750	0.597	0.881	0.841
	PEOU2	0.804			
	PEOU3	0.690			
	PEOU4	0.818			
	PEOU5	0.793			
Perceived Information Security (PIS)	PIS1	0.848	0.636	0.875	0.811
	PIS2	0.838			
	PIS3	0.727			
	PIS4	0.773			
Perceived Usefulness (PU)	PU1	0.613	0.548	0.828	0.715
	PU3	0.779			
	PU4	0.786			
	PU5	0.769			
Social Influence (SI)	SI1	0.677	0.516	0.842	0.776
	SI2	0.723			
	SI3	0.669			
	SI4	0.757			
	SI5	0.759			

TABLE. IV. FORNELL-LARCKER CRITERION AND HETEROTRAIT-MONOTRAIT RATIO

	INTU	PEOU	PE	PIS	PU	SI
Fornell-Larcker Criterion						
Intention To Use Social Media (INTU)	**0.793**					
Perceived Ease of Use (PEOU)	0.520	**0.772**				
Perceived Enjoyment (PE)	0.606	0.498	**0.805**			
Perceived Information Security (PIS)	0.427	0.322	0.308	**0.798**		
Perceived Usefulness (PU)	0.566	0.430	0.528	0.376	**0.740**	
Social Influence (SI)	0.542	0.364	0.453	0.354	0.556	**0.718**
Heterotrait-Monotrait Ratio						
Intention to Use Social (INTU)	-					
Perceived Ease of Use (PEOU)	0.626	-				
Perceived Enjoyment (PE)	0.726	0.578	-			
Perceived Information Security (PIS)	0.537	0.391	0.366	-		
Perceived Usefulness (PU)	0.739	0.528	0.658	0.495	-	
Social Influence (SI)	0.688	0.440	0.554	0.456	0.747	-

The discriminant validity of the examined constructs in this study was assessed using Fornell-Larcker and the Heterotrait-Monotrait Ratio (HTMT) criterion, as presented in Table IV [70], [71]. Table IV showed a higher square root of the AVE compared to the construct correlation, suggesting the establishment of the discriminant validity. This further strengthened the HTMT assessment result where the discriminant validity was established with HTMT0.90. Generally, the results suggested adequate convergent and discriminant validities of the model.

The results for the six constructs based were all considered valid measures of their constructs with respect to their statistical significance and parameter estimates. The results generally indicate adequate empirical support of the model for its reliability, convergent and discriminant validities.

2) Evaluation of the structural model: In this study, the structural model or inner model portrays the correlation between the examined constructs. Therefore, the evaluation of the structural model portrays the relationship between the research hypotheses and their effects on studied constructs. In this regard, the path coefficient (ß) criterion was used to test the postulated hypotheses in this study. The range of the standardised values for path coefficient is between -1 and +1; values closer to +1 implies a strong relationship between every two constructs and vice versa [67]. When assessing the significant level of relationships using path coefficient value, the t-value is usually higher than a specific critical value, indicating a significant coefficient at a given error probability. For instance, t-value > 1.96 indicates a significance level at p <0.05.

Table V showed that the results obtained from the research hypotheses tests were all acceptable. Specifically, the results of the first hypothesis (H1), which states that 'Perceived Usefulness (PU)' significantly influences on the 'intention to use social networks sites (INTU)'. This is based on evidence provided from that survey data with the result (ß = 0.169, t = 3.471, P-value<0.01) since the t-value is more than 1.96, the hypothesis is therefore accepted. Furthermore, the second hypothesis (H2) assumed that 'Perceived Ease of Use (PEOU)' had a positive influence on 'Intention to Use Social networks sites (INTU)' and this hypothesis was accepted (ß = 0.184, t = 3.873, p < 0.01). For the third hypothesis (H3), the significant influence of 'Perceived Enjoyment (PE)' on 'Intention to Social networks sites (INTU)' was also supported by the results (ß = 0.292, t=5.540, p < 0.01). Likewise, the fourth hypothesis (H4), which states that 'Social Influence (SI)' positively influences 'Intention to Use Social network sites (INTU)', was also supported by our survey data with values (ß = 0.198, t = 3.890, P-value<0.091). And (H5) which proposed a significant influence of 'Perceived Information Security (PIS) on the 'Intention to Use Social Media (INTU)' was also supported by the results (ß = 0.144, t=3.263, p < 0.01).

In this study, the coefficient of determination (R^2) value was also used to test the research hypotheses. As per Mitchell [72], R^2 values in the range of 0.01 - 0.09 are said to be low while those in the range of 0.09 - 0.25 are moderate; those in the range of 0.25 - 1 are high. Fig. 1 showed the results of the calculated R^2 values in this study. The calculated value (R^2) of the image was 0.537, indicating that 53.7% of the variances in Intention to Use Social Network Sites (INTU). Where explained by the five constructs, Perceived Ease of Use (PEOU), Perceived Enjoyment (PE), Perceived Information Security (PIS), Perceived Usefulness (PU), Social Influence (SI).

TABLE. V. HYPOTHESIS TESTING RESULTS

#	Hypothesis	Original Sample (O)	T Statistics (|O/STDEV|)	P. Values	Results
H1	PU→ INTU	0.169	3.471	0.001	Supported
H2	PEOU→INTU	0.184	3.873	0.000	Supported
H3	PE→INTU	0.292	5.540	0.000	Supported
H4	SI→ INTU	0.198	3.890	0.000	Supported
H5	PIS→ INTU	0.144	3.263	0.001	Supported

VII. DISCUSSION

This study examines student's acceptance of social network sites in e-Learning using perceived usefulness and perceived ease of use as adopted from TAM, in addition to the perceived enjoyment, social influence, and perceived information security based on literature review. All the postulates in the research model were supported.

The results of the factors that affect student's acceptance of social network sites in e-Learning revealed perceived usefulness (H1) significantly and positively affect the student's acceptance of social network sites in e-Learning (ß = 0.166, t = 2.667, p < 0.05). This result can be interpreted as when the students have more perceived usefulness; they will be more accepting of social network sites in e-Learning. Therefore, perceived usefulness is an indispensable factor to stimulate the students to accept social network sites in e-Learning. The results were consistent with the study of Al-Rahmi [42] who found a positive and significant association among perceived usefulness and social network sites in learning the Quran and Hadith. Sago [73] also found that the frequency usage of social network sites is positively impacted by the level of perceived usefulness provided by social network sites services. Likewise, Ramirez [74] found that perceived usefulness is positively associated with the use of social network sites. Furthermore, Al-Sharafi [79], found that perceived usefulness is positively associated with the user intention to use internet banking services.

Similarly, perceived ease of use (H2) was identified as a significant factor that is associated with the acceptance of social network sites in e-Learning (ß = 0.182, t =2.794, p < 0.05). This result can be interpreted as when the students have more perceived ease of use, and they will be more accepting of social network sites in e-Learning. Therefore, perceived ease of use is an indispensable factor to stimulate the students to accept social network sites in e-Learning. The results were consistent with the study of Al-Rahmi [42] who found a positive and significant association among perceived usefulness and social network sites in learning the Quran and Hadith. Similarly, Ramirez [74] found that perceived ease of use is a reliable prognosticator of social networks sites usage. Additionally, [76] indicated that perceived ease of use had positive effects on managers' intention to adopt e-commerce services.

Perceived enjoyment (H3) factor significantly and positively affect on the student's acceptance of social network sites in e-Learning (ß = 0.261, t = 4.633, p < 0.05). This result can be interpreted as when the students have more perceived enjoyment; they will be more accepting of social network sites in e-Learning. Therefore, perceived enjoyment is an indispensable factor to stimulate the students to accept social network sites in e-Learning. Therefore, if they find social network sites enjoyable, amusing, entertaining, and pleasant in completing their transactions effectively and efficiently, they will feel the impulse to use it in e-Learning. A result consistent with prior findings of Sledgianowski [75] found a positive influence of perceived enjoyment on social network sites adoption behaviour and Al-Rahmi [42] who found a positive and significant association among perceived enjoyment and social network sites in the context of learning Quran and Hadith.

The results as well showed that social influence (H4) supported the positive effect on the student's acceptance of social network sites in e-Learning (ß = 0.231, t = 4.828, p < 0.05). Based on this result, it is expected that an individual's beliefs in virtual communities may affect a student's intention to use social network sites in e-Learning. Thus, educational institutions policymaker can exploit the positive effect of social influence in supporting social media networks use in e-Learning. A result is harmonised with previous findings of [77], who found that social influence has a positive impact on behavioral intention. Moreover, perceived information security (H5) was identified as a significant factor that is associated with the acceptance of social network sites in e-Learning (ß = 0.204, t = 5.161, p <0.05). A result is matched with the study of [78], who found that information security positively impacts trust in E-government acceptance. Therefore, the students more perceived information security follows a tremendous student's acceptance of social network sites in e-Learning, meaning high perceived information security attracts students to use the e-Learning system. Consequently, perceived information security is a crucial factor to motivate the students to accept social network sites in e-Learning.

A. Theoretical and Practical Implications

The theoretical contribution is the particular implications of the results for the current theory associated with the social networks sites and e-Learning. In this study, a research model proposed to study the factors that influence students' acceptance of social network sites in e-Learning; thus, there are three theoretical implications. Firstly, the study made use of perceived ease of use, and perceived usefulness in addition to perceived enjoyment, social influence, and perceived information security based on a literature review to study the factors affects student's acceptance of social network sites in e-Learning. Secondly, there is a vast difference among the extent of positive insights of social networks sites and the extent of applied usage. Besides, there is a lack of studies on users' perspectives on similar technologies in developing countries.

Furthermore, no research until now has studied social networks sites acceptance in e-Learning in Palestine via the combination of perceived ease of use, and perceived usefulness, perceived enjoyment, social influence, and perceived information security. This study addressed this issue by derives the theories forming the base of the creation of the research model and proposing to determine the factors that affect the acceptance of social network sites in e-Learning at Palestine Technical University-Kadoorie. Thirdly, the research model can add to the existing social network sites and e-Learning literature.

Conversely, the current study also has practical implications. The first of which is the results of this study can be utilised by educational institutions policymakers to develop their educational strategies to benefit their institutions effectively. Secondly, technology developers at educational institutions may make use of the results and provide higher priority to perceived ease of use, perceived usefulness, perceived enjoyment, social influence, and perceived information security to increase student's acceptance of social networks sites in e-Learning. Thirdly, the survey used in this study agreed to the significant influence of (perceived ease of use, perceived usefulness, perceived enjoyment, social influence, and perceived information security) on student's acceptance of social network sites in e-Learning in Palestine. As well as, there is no research until now has studied social networks sites acceptance in e-Learning in Palestine. The result of this study should, therefore, support educational institutions policymakers and officials when taking stands on strategies to facilitate the effective use of social networks sites in e-Learning.

B. Limitation of the Study and Future Research

It is the interest of the current study to determine factors for the acceptance of social network sites in e-Learning. However, despite the contributions of the current study, some limitations still exist, which must be addressed. Firstly, the model used in the study includes two factors from TAM [30]; these factors are perceived usefulness and perceived ease of use with perceived enjoyment, social influence, and perceived information security as per the literature review. Regardless of the number of factors, their determination remains one of the limitations of the study because there are other factors that may affect a student's acceptance of social media networking in e-Learning. Secondly, the study sample was collected from undergraduate students in Palestine Technical University-Kadoorie. This sample may not adequately represent all the universities in Palestine.

Consequently, this result should not be generalised to the whole country in future studies. Thirdly, the present study scope is within the state of Palestine Technical University-Kadoorie; hence, the other universities, colleges, and educational institutions should be considered in future studies. Fourthly, this study cannot be generalised to all developing countries, as most of them may not share demographic features with Palestine. Therefore, further studies should be conducted in different countries to validate and strengthen the results of this study. The results of this study might help other researchers in shaping subjects to highlight, to hearten progress in social network sites acceptance in e-Learning.

VIII. Conclusion

Widespread usage of social network sites has attracted the care of learning institutions in the world. Social networks sites offer exceptional learning that relies on participative communications as alternates for traditional learning. Moreover, social network sites offer more great chances for sharing capabilities that see further experiments that exploit these advantages, encouraging the creative and innovative skills of learners. Social network sites thereby represent an appropriate environment for the types of sophisticated educational sceneries that can suit the necessities of contemporary learners, consistent with universal technical progress.

The significant contribution of the current research is the development of a model for social network sites acceptance, including shaping factors that impact students' intentions to use social network sites in e-Learning for studying the research objective. This model is a coherent model that can be used in future empirical studies in the same field; the model can also be extended in different directions. The model developed in this research can guide further studies on student's acceptance of social network sites in e-Learning.

The results showed a positive and significant relationship among perceived usefulness, perceived, ease of use, perceived enjoyment, social influence, and perceived information security and social networks sites acceptance in e-Learning. Based on these findings, the effect of these factors should be considered by researchers in the future and policymakers when trying to improve on the level of social networks sites acceptance in e-Learning in a developing country, including Palestine. Therefore, this study emphasised the prominence of social network sites as practical learning tools.

Acknowledgment

We would like to thanks Palestine Technical University-Kadoorie for funding and supporting the publication of this research.

REFERENCES

[1] J. Abbas, J. Aman, M. Nurunnabi, and S. Bano, "The Impact of Social Media on Learning Behavior for Sustainable Education: Evidence of Students from Selected Universities in Pakistan," *Sustainability*, vol. 11, no. 6, p. 1683, 2019.

[2] G. P.-K. Zachos, E.-A.; Anagnostopoulos, I., "Social Media Use in Higher Education: A Review," Education Science vol. 8, no. 4, 2018, doi: http://dx.doi.org/10.3390/educsci8040194.

[3] D. T. Hansen, The teacher and the world: A study of cosmopolitanism as education. Routledge, 2017.

[4] K. Bredl, Methods for analyzing social media. Routledge, 2017.

[5] E. Bagarukayo and B. Kalema, "Evaluation of elearning usage in South African universities: A critical review," International Journal of Education and Development using ICT, vol. 11, no. 2, 2015.

[6] P. Lakbala, "Barriers in implementing E-learning in Hormozgan University of Medical Sciences," Global journal of health science, vol. 8, no. 7, p. 83, 2016.

[7] A. Rhema and I. Miliszewska, "Analysis of student attitudes towards e-learning: The case of engineering students in Libya," Issues in informing science and information Technology, vol. 11, no. 1, pp. 169-190, 2014.

[8] J. A. Larusson and R. Alterman, "Wikis to support the "collaborative" part of collaborative learning," International Journal of Computer-Supported Collaborative Learning, vol. 4, no. 4, pp. 371-402, 2009.

[9] P. A. Ertmer, T. J. Newby, W. Liu, A. Tomory, J. H. Yu, and Y. M. Lee, "Students' confidence and perceived value for participating in cross-cultural wiki-based collaborations," Educational Technology Research and Development, vol. 59, no. 2, pp. 213-228, 2011.

[10] W. M. Al-rahmi, M. S. Othman, and M. A. Musa, "The improvement of students' academic performance by using social media through collaborative learning in Malaysian higher education," Asian Social Science, vol. 10, no. 8, p. 210, 2014.

[11] I. D. Keenan, J. D. Slater, and J. Matthan, "Social media: Insights for medical education from instructor perceptions and usage," MedEdPublish, vol. 7, 2018.

[12] M. N. Yakubu and S. Dasuki, "Assessing eLearning systems success in Nigeria: An application of the DeLone and McLean information systems success model," Journal of Information Technology Education: Research, vol. 17, pp. 183-203, 2018.

[13] N. Ameen, R. Willis, M. N. Abdullah, and M. Shah, "Towards the successful integration of e-learning systems in higher education in Iraq: A student perspective," British Journal of Educational Technology, vol. 50, no. 3, pp. 1434-1446, 2019.

[14] K. N. Shen and M. Khalifa, "Facebook usage among Arabic college students: preliminary findings on gender differences," 2010.

[15] A. Al-Azawei, "What Drives Successful Social Media in Education and E-Learning? A Comparative Study on Facebook and Moodle," Journal of Information Technology Education, vol. 18, 2019.

[16] Z. N. Khlaif and S. Farid, "Transforming learning for the smart learning paradigm: lessons learned from the Palestinian initiative," Smart Learning Environments, vol. 5, no. 1, p. 12, 2018.

[17] Y. K. Dwivedi et al., "Research on information systems failures and successes: Status update and future directions," Information Systems Frontiers, vol. 17, no. 1, pp. 143-157, 2015.

[18] K. Shraim and Z. Khlaif, "An e-learning approach to secondary education in Palestine: opportunities and challenges," Information Technology for Development, vol. 16, no. 3, pp. 159-173, 2010.

[19] S. Saidam, "Knowledge and e-governance building in conflict affected societies: Challenges and mechanisms," in Proceedings of the 1st international conference on Theory and practice of electronic governance, 2007, pp. 341-344.

[20] M. Ahsan and M. Kumari, "Rumors and their controlling mechanisms in online social networks: A survey," Online Social Networks and Media, vol. 14, p. 100050, 2019.

[21] M. Badri, A. Al Nuaimi, Y. Guang, and A. Al Rashedi, "School performance, social networking effects, and learning of school children: Evidence of reciprocal relationships in Abu Dhabi," Telematics and Informatics, vol. 34, no. 8, pp. 1433-1444, 2017.

[22] A. Abdulahi, B. Samadi, and B. Gharleghi, "A study on the negative effects of social networking sites such as facebook among asia pacific university scholars in Malaysia," International Journal of Business and Social Science, vol. 5, no. 10, 2014.

[23] A. M. Elkaseh, K. W. Wong, and C. C. Fung, "Perceived ease of use and perceived usefulness of social media for e-learning in Libyan higher education: A structural equation modeling analysis," International Journal of Information & Education Technology,vol.6,no.3,p.192, 2016.

[24] P. A. Ertmer, T. J. Newby, W. Liu, A. Tomory, J. H. Yu, and Y. M. Lee, "Students' confidence and perceived value for participating in cross-cultural wiki-based collaborations," Educational Technology Research and Development, vol. 59, no. 2, pp. 213-228, 2011.

[25] S. Manca and M. Ranieri, "Implications of social network sites for teaching and learning. Where we are and where we want to go," Education and Information Technologies, vol.22,no.2,pp.605-622, 2017.

[26] H. Karajeh et al., "Social media networks and pedagogy at the University of Jordan," Education and Information Technologies, vol. 23, no. 5, pp. 2073-2090, 2018.

[27] D. Z. Dumpit and C. J. Fernandez, "Analysis of the use of social media in Higher Education Institutions (HEIs) using the Technology Acceptance Model," International Journal of Educational Technology in Higher Education, vol. 14, no. 1, p. 5, 2017.

[28] E. Cho and J. Son, "The effect of social connectedness on consumer adoption of social commerce in apparel shopping," Fashion and Textiles, vol. 6, no. 1, p. 14, 2019.

[29] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS quarterly,pp.319-340, 1989.

[30] A. T. Shittu, K. B. Madarsha, N. S. N. AbduRahman, and T. B. T. Ahmad, "Determinants of social networking software acceptance: A multi-theoretical approach," Malaysian Online Journal of Educational Technology, vol. 1, no. 1, pp. 27-43, 2013.

[31] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," MIS quarterly, pp. 425-478, 2003.

[32] J. Kalin, "Doing what comes naturally? Student perceptions and use of collaborative technologies," International Journal for the Scholarship of Teaching and Learning, vol. 6, no. 1, p. 10, 2012.

[33] M. H. Fagan, S. Neill, and B. R. Wooldridge, "Exploring the intention to use computers: An empirical investigation of the role of intrinsic motivation, extrinsic motivation, and perceived ease of use," Journal of Computer Information Systems, vol. 48, no. 3, pp. 31-37, 2008.

[34] T. Ramayah and J. Ignatius, "Impact of perceived usefulness, perceived ease of use and perceived enjoyment on intention to shop online," ICFAI Journal of Systems Management (IJSM), vol. 3, no. 3, pp. 36-51, 2005.

[35] X. Deng, W. J. Doll, A. R. Hendrickson, and J. A. Scazzero, "A multi-group analysis of structural invariance: an illustration using the technology acceptance model," Information & Management, vol. 42, no. 5, pp. 745-759, 2005.

[36] S.-H. Liu, H.-L. Liao, and J. A. Pratt, "Impact of media richness and flow on e-learning technology acceptance," Computers & Education, vol. 52, no. 3, pp. 599-607, 2009.

[37] J. M. Curran and M. L. Meuter, "Encouraging existing customers to switch to self-service technologies: put a little fun in their lives," Journal of Marketing Theory and Practice, vol. 15, no. 4, pp. 283-298, 2007.

[38] H. Van der Heijden, "User acceptance of hedonic information systems," MIS quarterly, pp. 695-704, 2004.

[39] R. G. Saadé, W. Tan, and F. Nebebe, "Impact of Motivation on Intentions in Online Learning: Canada vs China," Issues in Informing Science & Information Technology, vol. 5, 2008.

[40] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "Extrinsic and intrinsic motivation to use computers in the workplace 1," Journal of applied social psychology, vol. 22, no. 14, pp. 1111-1132, 1992.

[41] C.-L. Hsu and J. C.-C. Lin, "Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation," Information & Management, vol.45, no. 1, pp. 65-74, 2008.

[42] W. M. Al-Rahmi and A. M. Zeki, "A model of using social media for collaborative learning to enhance learners' performance on learning,"

Journal of King Saud University-Computer and Information Sciences, vol. 29, no. 4, pp. 526-535, 2017.

[43] H.-T. Tsai and R. P. Bagozzi, "Contribution behavior in virtual communities: Cognitive, emotional, and social influences," Mis Quarterly, vol. 38, no. 1, pp. 143-164, 2014.

[44] R. L. Thompson, C. A. Higgins, and J. M. Howell, "Personal computing: toward a conceptual model of utilization," MIS quarterly, pp. 125-143, 1991.

[45] Y. Wang, D. B. Meister, and P. H. Gray, "Social influence and knowledge management systems use: Evidence from panel data," Mis Quarterly, pp. 299-313, 2013.

[46] A. Maryam, N. Maarop, R. Ibrahim, and M. Hasan, "Towards a model for studying social media adoption for the co-creation services domain," Indian Journal of Science and Technology, vol. 9, p. 34, 2016.

[47] S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," Electronic Markets, vol. 26, no. 2, pp. 173-194, 2016.

[48] X. Luo, A. Gurung, and J. P. Shim, "Understanding the determinants of user acceptance of enterprise instant messaging: an empirical study," Journal of organizational computing and electronic commerce, vol. 20, no. 2, pp. 155-181, 2010.

[49] I. Achumba, O. Ighomereho, and M. Akpor-Robaro, "Security challenges in Nigeria and the implications for business activities and sustainable development," Journal of economics and sustainable development, vol. 4, no. 2, 2013.

[50] S. Byabato and K. Kisamo, "Implementation of school based continuous assessment (CA) in Tanzania ordinary secondary schools and its implications on the quality of education," Developing Country Studies, vol. 4, no. 6, pp. 55-61, 2014.

[51] C. Jewitt, C. Hadjithoma-Garstka, W. Clark, S. Banaji, and N. Selwyn, "School use of learning platforms and associated technologies–case study: secondary school 1," 2010.

[52] S.-C. Yang and C.-H. Lin, "Factors affecting the intention to use Facebook to support problem-based learning among employees in a Taiwanese manufacturing company," African Journal of Business Management, vol. 5, no. 22, pp. 9014-9022, 2011.

[53] M. M. Abbad, D. Morris, and C. De Nahlik, "Looking under the bonnet: Factors affecting student adoption of e-learning systems in Jordan," The International Review of Research in Open and Distributed Learning, vol. 10, no. 2, 2009.

[54] Y.-C. Lee, "An empirical investigation into factors influencing the adoption of an e-learning system," Online information review, vol. 30, no. 5, pp. 517-541, 2006.

[55] S. Y. Park, "An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning," Educational technology & society, vol. 12, no. 3, pp. 150-162, 2009.

[56] M. Luthfihadi and W. Dhewanto, "Technology Acceptance of E-commerce in Indonesia," International Journal of Engineering Innovation and Management, vol. 3, pp. 9-18, 2013.

[57] M. MäNtymäKi and J. Salo, "Teenagers in social virtual worlds: Continuous use and purchasing behavior in Habbo Hotel," Computers in Human Behavior, vol. 27, no. 6, pp. 2088-2097, 2011.

[58] H. Nysveen, P. E. Pedersen, and H. Thorbjørnsen, "Intentions to use mobile services: Antecedents and cross-service comparisons," Journal of the academy of marketing science, vol. 33, no. 3, pp. 330-346, 2005.

[59] S. Moghavvemi, M. Sharabati, T. Paramanathan, and N. M. Rahin, "The impact of perceived enjoyment, perceived reciprocal benefits and knowledge power on students' knowledge sharing through Facebook," The International Journal of Management Education, vol. 15, no. 1, pp. 1-12, 2017.

[60] M. D. Dzandu, H. Boateng, G. Agyemang, and F. Quansah, "Social media adoption among University Students: What is the role of gender, perceived usefulness and perceived ease of use," International Journal of Social Media and Interactive Learning Environments, 2016.

[61] C. Lorenzo-Romero and M.-d.-C. Alarcón-del-Amo, "Segmentation of users of social networking websites," Social Behavior and Personality: an international journal, vol. 40, no. 3, pp. 401-414, 2012.

[62] M. Abdekhoda, A. Dehnad, S. J. G. Mirsaeed, and V. Z. Gavgani, "Factors influencing the adoption of E-learning in Tabriz University of Medical Sciences," Medical journal of the Islamic Republic of Iran, vol. 30, p. 457, 2016.

[63] K. Mzava and E. A. Kalinga, "Determinants of Nurses’intention to Use Elearning in Tanzania," Journal of Health Informatics in Developing Countries, vol. 11, no. 2, 2017.

[64] B. M. Byrne, Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming. Psychology Press, 2013.

[65] J. C. Nunnally, Psychometric theory 3E. Tata McGraw-Hill Education, 1994.

[66] D. Andrew, P. Pedersen, and C. McEvoy, "Research methods and design in sport management," Research methods and design in sport management, 2011.

[67] J. F. Hair Jr, G. T. M. Hult, C. Ringle, and M. Sarstedt, A primer on partial least squares structural equation modeling (PLS-SEM). Sage publications, 2016.

[68] U. Sekaran, and R. Bougie (2003). “Research methods for business, a skill building approach”, John Willey & Sons,” Inc. New York.

[69] J. F. Hair, C. M. Ringle, and M. Sarstedt, "PLS-SEM: Indeed a silver bullet," Journal of Marketing theory and Practice, vol. 19, no. 2, pp. 139-152, 2011.

[70] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," Journal of marketing research, vol. 18, no. 1, pp. 39-50, 1981.

[71] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," Journal of the academy of marketing science, vol. 43, no. 1, pp. 115-135, 2015.

[72] M. L. Mitchell, and J. M. Jolley (2012). Research design explained: Cengage Learning.

[73] B. Sago, "Factors influencing social media adoption and frequency of use: An examination of Facebook, Twitter, Pinterest and Google+," International Journal of Business and Commerce, vol. 3, no. 1, pp. 1-14, 2013.

[74] P. Ramírez-Correa, E. E. Grandón, M. Ramírez-Santana, and L. Belmar Órdenes, "Explaining the use of social network sites as seen by older adults: The enjoyment component of a hedonic information system," International journal of environmental research and public health, vol. 16, no. 10, p. 1673, 2019.

[75] D. Sledgianowski and S. Kulviwat, "Social network sites: Antecedents of user adoption and usage," AMCIS 2008 Proceedings, p. 83, 2008.

[76] A. Herzallah and M. Mukhtar, "The Impact of Perceived Usefulness, Ease of Use and Trust on Managers’ Acceptance of e-Commerce Services in Small and Medium-Sized Enterprises (SMEs) in Palestine," 2016.

[77] S. Chauhan, M. Jaiswal, and A. K. Kar, "The acceptance of electronic voting machines in India: a UTAUT approach," Electronic Government, an International Journal, vol. 14, no. 3, pp. 255-275, 2018.

[78] M. M. Ayyash, K. Ahmad, and D. Singh, "Investigating the effect of information systems factors on trust in e-government initiative adoption in Palestinian public sector," Research Journal of Applied Sciences, Engineering and Technology, vol. 5, no. 15, pp. 3865-3875, 2013.

[79] Al-Sharafi, A., Arshah, R. A., Alajmi, Q., Herzallah, A. T., & Qasem, Y. A., “The Influence of Perceived Trust on Understanding Banks' Customers behavior to Accept Internet Banking Services”, Indian Journal of Science and Technology, Vol 11(20), PP. 1-9, 2018.

Managing an External Depot in a Production Routing Problem

Bi Kouaï Bertin Kayé[1], Moustapha Diaby[2], Tchimou N'Takpé[3], Souleymane Oumtanaga[4]
Institut National Polytechnique Félix Houphouët-Boigny, Yamoussoukro, Côte d'Ivoire[1, 4]
Ecole Supérieur Africaine des TIC, Abidjan, Côte d'Ivoire[2]
Université Nangui Abrogoua, Abidjan, Côte d'Ivoire[3]
Laboratoire de Recherche en Informatique et Télécommunication, Abidjan, Côte d'Ivoire[1, 2, 3, 4]

***Abstract*—This paper addresses a production and distribution problem in a supply chain. The supply chain consists of a plant with no storage capacity that produces only one type of product. The manufactured products are then transported to a depot for storage. Customers demand is met by a homogeneous fleet of vehicles that begins and ends their trips at the depot. The objective of the study is to minimize the overall cost of production, inventory and transport throughout the supply chain. A Branch-and-Cut and a hybrid Two Phases Decomposition Heuristic using a Mixed Integer Programming and a Genetic Algorithm have been developed to solve the problem.**

***Keywords*—Production; inventory; distribution; transport; branch-and-cut; decomposition heuristic; MIP; genetic algorithm**

I. Introduction

In a supply chain, several activities are involved in meeting consumer needs. Producing, storing and distributing products are the main activities of the actors or companies that make up a supply chain. The problem of production or Lot Sizing (LSP) consists in determining the production schedule, the quantities to be produced when there is production and the quantities stored on the planning horizon. Planning production therefore means finding the best compromise between the production schedule and the quantities stored on the planning horizon. For a better visibility on production planning, readers are invited to see [1] . Another very important problem in the area of supply chain planning is the Vehicle Routing Problem (VRP). This is a NP-Hard problem as a particular case of the Travelling Salesman Problems (TSP) which is itself a NP-Hard problem [2]. It does not consider inventory and production decisions. It is mainly interested in the organization of the routes of each vehicle, knowing that the schedule for each customer's visit is known in advance. It therefore makes it possible to answer questions such as "Who to visit? "and "In which order should the visits be made? The objective of a VRP is to minimize the total cost of transport to meet customer needs over the planning horizon. The Inventory Routing Problem (IRP) is a supply chain planning problem based on the integration and coordination of inventory decisions and the determination of the best vehicle trips over a given planning horizon. The IRP is a generalization of the VRP which allows in addition, answer questions such as "When to deliver or collect products?" and "how much to collect or deliver for a given period of the planning horizon?". This is the place to implement a good replenishment policy and management practice such as the Vendor Managed Inventory (VMI). Among the replenishment policies [3] , the Order-up-to-level (OU) and Maximum level (ML) policies are the most used. In the OU replenishment policy, all customers have a maximum storage capacity and the quantity delivered is such that the maximum level of storage is reached at each delivery. While in the ML policy all customers have a maximum storage capacity and the quantity delivered is such that the maximum level of storage is not exceeded at each delivery [4] . Regarding the VMI, it is a practice in which the supplier decides when and how much to deliver to the customer based on customers' inventory information. He must also ensure that there is no stock shortage at the customers. This practice contrasts with the Retailer Managed Inventory (RMI) or Costumer Managed Inventory (CMI) practice in which customers decide the time and amount of replenishment, regardless of decisions made by other customers and the supplier. The CMI or RMI is therefore an appropriate practice for VRP. For a comparative study between the CMI and the VMI see [4]. In a supply chain, the production problem is solved independently of the decisions that characterize the VRP or the IRP. Similarly, the VRP or IRP problem is analysed without considering production decisions. However ,in a comparative study to analyse the importance of coordination of production and distribution [5] , it has been shown that savings of 3% to 20% on the total cost of production and distribution can be achieved by considering the supply chain as an integrated and coordinated system. As a result, integrating and coordinating production and distribution decisions has become a major concern for researchers in recent decades. The integrated and coordinated planning of the production and distribution functions of the supply chain is known as the Production Routing Problem (PRP). In order to facilitate the understanding of this work, the following organization is adopted. In Section II, a review of the PRP literature and these methods of solving are presented. Section III provides a description of the problem and its mathematical formulation. Sections IV and V propose a Branch-and-Cut algorithm and a Two-Phases Decomposition Heuristic, respectively. Test results can be found in Section VI. The conclusion of the work is presented in Section VII.

II. Literature Reviews

PRP is a NP-Hard problem considering that it contains the VRP. Then, several heuristics have been developed for its resolution. Without being exhaustive, there are the Greedy Randomized Adaptive Search Procedure (GRASP) [6], [7], the Adaptive Large Neighborhood Search (ALNS) procedure [8], [9], the Tabu research (TS) [10], the Variable Neighborhood Search (VNS) [11], the Memetic Algorithm [12] ,the heuristic based on Mixed Integer Programming (MIP) and iterative MIP [13], [14] .Decomposition heuristics are resolution approaches that consist in dividing the problem into several subproblems. the subproblems are then solved sequentially. An improvement procedure can be used to improve the final solution obtained. Several authors have developed a decomposition approach to solve the PRP problem. The first decomposition method for PRP problems was proposed by [5], [15] to solve a PRP problem with several products. The authors cut the problem into Capacitated LSP solved to the optimal and a distribution planning problem. In the work of [16], the PRP problem is solved in two phases. The first phase consists in solving the LSP in which distribution is considered through direct shipment. Once the decisions of LSP have been made and the customers to be visited have been determined, the second phase aims to determine the route of each vehicle over the planning horizon. The authors developed a VRP heuristic to solve the second phase of the problem. An improved version of the decomposition method has been proposed by [17]. The authors used the economic algorithm of Wagner and Whitin [18] to solve LSP and the algorithm of Clarke and Wright [19] for the construction of distribution plans for each period .They also used a local search procedure to improve the solution. Since the introduction of the Branch-and-cut algorithm (B&C) as an exact method to solve the integrated IRP [20], more and more researchers have been interested in the exact resolution of the PRP . the B & C is the most widely used of the exact methods of resolution of the PRP. It has been used to solve a PRP problem in which vehicle capacity and plant production capacity have not been considered [21]. For the model with the consideration of the capacity of a single vehicle used to transport a single type of product and in which production capacity is not considered, see [14]. The single-vehicle model has been extended to the multi-vehicle model in [8]. Qiu and al. have recently introduced three B&C algorithms to deal with various PRP. The first is a PRP in which reverse logistics and remanufacturing are considered [22] , the second is a problem of production and distribution of perishable products [23] and the third is a problem of production of several types of products with the use of several homogeneous vehicles and taking into account the setting cost when there is production or when moving from the production of one product to another [24]. For the exact method using Bender's decomposition see [25]. The difficulty in the exact resolution of the PRP lies in eliminating subtours in a vehicle's route. This difficulty is the same for any tour problem such as TSP, VRP or IRP. Two subtour elimination constraints (SECs) are increasingly used in PRP. On the one hand, there are the SECs resulting from the formulation of Boudia and al. [6], [17] and on the other hand, the SECs developed for selective TPS in [26] . Three methods or callable libraries have been proposed for the exact or heuristic separation of these SECs. Among these methods, there are four heuristics developed at the base for solving the Capacitated VRP (CVRP) with consideration of vehicle capacity [27] . An exact and heuristic separation by using the minimum S-T cut algorithm of the Concorde callable library have been developed by [28], [29] and a polyhedral approach to SECs separation has also been proposed in [30] . The exact algorithms used for PRP resolution in most cases use one of these SECs separations approaches. Table I presents the works on the use of exact resolution methods as well as the type of mathematical formulation of the model and the separation method used. The notation F|k refers to the formulation with vehicle index and the formulation F|nk refers to the formulation without vehicle index. These notations can be enriched by the precision of the type of procurement policy used in the mathematical formulation of the model. Thus, with the OR replenishment policy, one can obtain the notation of type F(OU)|k and F(OU)|nk and the notation of type F(ML)|k and F(ML)|nk for the ML replenishment policies. See [8] for details on the notation of mathematical formulations. Readers are also invited to see a very detailed literature review proposed in [31]. In this literature review, the authors presented the different types of integration problems such as the integrated problem of LSP and direct delivery, IRP, and the PRP.

TABLE. I. SUBTOUR ELIMINATION PROCEDURES

Works	**Model**	**Method of resolution**	**Separation method used**
Adulyasak and al.[25]	F|k	Bender’s decomposition	Applegate and al.[28]
Ruokokoski and al .[21]	F|nk	B&C	Applegate and al.[32] , [29]
Archetti and al.[14]	F|nk	B&C	Padberg and Rinaldi .[30]
Adulyasak and al.[8]	F|k and F|nk	B&C	Applegate and al.[28]
Qiu and al.[22]	F|k	B&C	Lysgaard and al.[27]
Qiu and al.[23]	F|nk	B&C	Lysgaard and al.[27]
Qiu and al.[24]	F|nk	B&C	Lysgaard and al.[27]

They also highlighted the different mathematical formulations of the PRP with resolution methods. A classification based on four criteria and covering 77 research studies from 1993 to 2016 was presented in a literature review [33] . The authors categorized the work according to the level of decision, the typology of the problem in the supply chain, the type of objective sought and the problem optimization model (resolution method). The problem studied in this work is to plan and optimize an integrated supply chain in which production decisions for a single type of product in a plant with no storage capacity and those for inventories in an off-plant depot are integrated and coordinated to meet the deterministic demands of several customers. We denote this problem by the "External Depot Production Routing Problem" (EDPRP), To the best of our knowledge, the PRP with External Depot has not been addressed before. In this EDPRP, a fleet of homogeneous vehicles leaves the depot for the distribution of products to customers or the collection of products at the plant. In the following section, more detailed description of the model and its mathematical formulation will be presented.

III. Description of the Problem and Mathematical Formulation

G= (N, A) is a complete graph in which N represents all the nodes formed by the plant, the depot and the customers with the index i ∈ {0...n+1} and A(N) = {(i, j) : i, j ∈ N, i ≠ j} all the arcs in N. The plant is represented by n+1, the depot is indexed by 0 and all customers are represented by {1, ..., n}. the graph of Fig. 1 represents the collection and distribution network for one plant, one depot and seven customers.

The sets

Let denote N_c= {1, …, n} the set of customers,

N_{dc}= {0, …, n} the set consisting of the depot and customers,

N_{cu}= {1, …, n+1} the set consisting of the customers and the plant,

N= {0, …, n+1} the set for depot, customers and plant,

T= {1, …, l} the set of periods (days) of the planning horizon,

K= {1, …, m} the set of homogeneous fleet of vehicles.

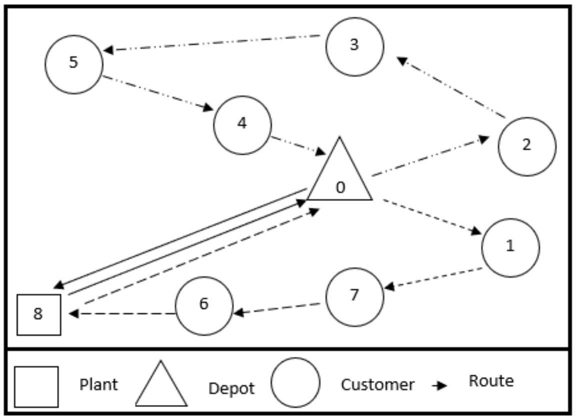


Fig. 1. Collection and Distribution Network.

The index:

i and j represent the index for the nodes of N,

t is the index of the different periods of the planning horizon,

k is the index for the homogeneous fleet of vehicles.

The parameters:

u : Unit Production Cost (UPC),

f: fixed cost of production,

h_i : Unit Inventory Cost (UIC) at node i,

c_{ij} : Unit Transportation Cost (UTC) from the node i to the node j,

d_{it} : the demand of the customer i in the period t,

C: plant production capacity,

Q: Maximum vehicle capacity,

L_i : Maximum or targeted storage capacity at node I,

I_{i0} : Initial stock available at node I,

MU_t ; = min { C, $\sum_{i\in N_c}\sum_{\tau=t}^{l} d_{i\tau}$ } ∀ t∈ T,

MC_{it} : =min{ L_i , Q, $\sum_{\tau=t}^{l} d_{i\tau}$} ∀ i ∈ N_1 ,∀ t ∈ T ,

Decision variables

p_t: quantity produced during the period t

I_{it}: level of stocks at the node i during the period t

q_{ikt}: quantity delivered to the node i by vehicle k during the period t

y_t : binary variable equal to 1 if there is production at the plant or 0 if not

Z_{ikt}: binary variable, equal to 1 if the i node is visited by vehicle k during period t or 0 if not

x_{ijkt} : binary variable, equal to 1 if vehicle k travels directly from the node i to the node j during period t or 0 if not.

Mathematical formulation

$$Z= \min \sum_{t\in T}(up_t + fy_t) + \sum_{t\in T}\sum_{i\in N} h_i I_{it} + \sum_{t\in T}\sum_{(i,j)\in A}\sum_{k\in K} c_{ij}x_{ijkt} \quad (1)$$

$$\text{S.t } p_t \le MU_t y_t \ \forall t\in T \quad (2)$$

$$p_t = \sum_{k\in K} q_{0kt} \ \forall t\in T \quad (3)$$

$$\sum_{k\in K} z_{n+1,kt} \le m y_t \ \forall t\in T \quad (4)$$

$$I_{0,t-1} + p_t = I_{0t} + \sum_{i\in N_c}\sum_{k\in K} q_{ikt} \quad \forall t\in T \quad (5)$$

$$I_{it-1} + \sum_{k\in K} q_{ikt} = d_{it} + I_{it} \ \forall i\in N_c , \forall t\in \quad (6)$$

$$\sum_{i\in N_c}\sum_{k\in K} q_{ikt} \le I_{0t-1} \ \forall t\in T \quad (7)$$

$$I_{it} \le L_i \ \forall i\in N_{dc}, \forall t\in T \quad (8)$$

$$\sum_{i\in N_c} q_{ikt} \le Q z_{0kt} \ \forall k \in K, \forall t \in T \quad (9)$$

$q_{0kt} \leq$ Q $z_{n+1,kt}$ ∀ k ∈ K, ∀ t ∈ T (10)

q_{ikt} ≤$MC_{it} z_{ikt}$∀ i ∈ N_c, ∀ k ∈ K, ∀ t∈ T (11)

$q_{n+1,kt} = 0$ ∀ k∈ K, ∀ t∈ T (12)

$\sum_{j\in N} x_{ijkt} = z_{ikt}$ ∀ i ∈ N, ∀ k ∈ K, ∀ t ∈ T (13)

$\sum_{j\in N} x_{jikt} + \sum_{j\in N} x_{ijkt}$ =2 z_{ikt}∀ i ∈ N, ∀ k ∈ K, ∀ t ∈ T (14)

$\sum_{k\in K} z_{ikt} \leq 1$ ∀ i∈ N_c , ∀ t∈ T (15)

x_{n+1jkt} = 0 ∀ j ∈ N_c, ∀ k ∈ K, ∀ t∈ T (16)

$\sum_{i\in S}\sum_{j\in S} x_{ijkt}$ ≤ |S|-1 ∀ S $\underline{C}$ N_c, |S|≥ 2, ∀ k∈ K, ∀ t∈ T (17)

$p_t, I_{it}, q_{ikt} \geq 0$ ∀ i∈ , ∀ k∈ K, ∀ t∈ T (18)

y_t, x_{ijkt},z_{ikt} ∈ {0,1} ∀ i, j ∈ , ∀ k ∈ K, ∀ t ∈ T (19)

The function (1) is the objective function. This function minimizes the total cost of production, inventory and transportation. Constraints (2) determine whether there is production at a given time in the planning horizon while determining the maximum amount to be produced for that same period. The constraints (3) indicate that all quantities produced at the plant must be transported to the depot to be stored. The constraints (4) limit the number of vehicles assigned to collect products at the plant when production occurs. Constraints (5) and (6) are the constraints of product flow conservation respectively at the plant and among customers. The constraints (7) indicate that customer deliveries in each period must be made from the depot stock of the previous period. Constraints (8) limit the stocks of each period by a maximum storage capacity both at the depot and at customers. Constraints (9) indicate that the amount of product delivered by each vehicle may not exceed the maximum capacity of the vehicle. The constraints (10) indicate that the amount collected by a vehicle over a period cannot exceed the maximum capacity of the vehicle. The constraints (11) limits the quantities delivered to each customer by each vehicle over each period. Constraints (12) indicate that each vehicle passing through the plant must be empty before entering the plant. (13) and (14) are vehicles flow conservation constraints. The constraints (15) indicate that each customer is visited no more than once by a vehicle during a given period. Constraints (16) indicate that no visit from a customer is allowed when leaving the plant. Constraints (17) are Subtour Elimination Constraints. The constraints (18) are non-negativity constraints. Two resolution approaches are used to solve the problem. At first, a B&C algorithm is developed to solve small instances and then a Tow Phases Decomposition Heuristic (TPDH) is developed for solving all instances of the problem.

IV. A B and C Algorithm for EDPRP Resolution

For this first resolution of the model, a B&C algorithm is used and following valid inequalities and steps are adopted:

A. Valid Inequalities

1) $\sum_{k\in K}\sum_{t\in T} z_{ikt} \geq 1$ ∀ i ∈ N_c (20). these constraints indicate that each customer must be visited at least once on the planning horizon.

2) $z_{ikt} \leq z_{0kt}$ ∀ i∈ N_{cu} , ∀ k ∈ K, ∀ t∈ T (21). these constraints indicate that if a vehicle k does not leave depot 0 in period t, then it will not visit any customers or the plant in the same period.

3) $x_{ijkt} \leq z_{ikt}$ and $x_{ijkt} \leq z_{jkt}$ ∀ (i, j) ∈A(N) , ∀ k ∈ K, ∀ t∈ T (22). these constraints indicate that no path will enter or leave a customer if the customer is not visited for a given period.

4) $x_{ijkt} \leq q_{jkt}$ and $x_{n+1,0kt} \leq q_{0kt}$ ∀ i∈ N_{dc} , ∀ j∈ N_c , ∀ k ∈ K, ∀ t∈ T (23). These constraints make it possible to avoid unladen visits of vehicles.

5) $x_{ijkt} + x_{jikt} \leq 1$ ∀ (i, j) ∈A(N_c), ∀ k ∈ K, ∀ , t∈ T (24) . with these constraints, each arc is crossed only once and in one direction by a vehicle.

6) $z_{0,k+1,t} \leq z_{0kt}$ ∀ k ∈ 1, …, m-1, ∀ t ∈ T (25) are the Symmetry-Breaking Constraints (SBC : valid for a homogeneous vehicles) [34], [35] . These inequities ensure that the k-1 vehicle cannot leave the depot if the k vehicle is not used.

7) $I_{i,t-s-1} \geq \sum_{j=0}^{s} d_{i,t-j}\left(1 - \sum_{k\in \mathrm{K}}\sum_{j=0}^{s} z_{i,k,t-j}\right)$
$\forall i \in N_c, \forall\, t \in T, \forall\, k \in K, s = 0,1,\dots,t-1$ (26) [14], [20].

B. B and C Algorithm

To solve the problem, constraints (20), (21), (22), (23), (24), (24), (25), (26) are added to the initial model defined by the objective function (1) and constraints (2) - (19) and let the SECs (17) drop .The relaxation of the linear program (LP) is then resolved. the approach to eliminate subtours is divided into two stages. At each node of the B&C tree, a check in the first step whether the tour of a vehicle k at a date t contains a subtour is done. Then, in the second step the corresponding SECs and Relinking Constraints (RCs) for this vehicle k at date t are added when a subtour is detected. In the remainder of this section, the use of "*" refers to a component of the LP solution at a node of the B&C search tree.

C. Subtour Detection

The subtour detection algorithm takes as input the vectors Z^* and x^* and produces as output the sets ST_{kt}^* of customers involved in a subtour for vehicle k at period t (if this set exists). The method for determining ST_{kt}^* is described as follows. For any vector Z^* and x^* of the LP solution at period t and for vehicle k ,the corresponding graph $G_{kt}^*(N_{kt}^*, A_{kt})$ is defined with N_{kt}^* = { i ∈ N : $Z_{ikt}^* > 0$ } and , $A_{kt}(N_{kt}^*)$ = { (i ,j) ∈ $A(N_{kt}^*)$: $x_{ijkt}^* > 0$ }. Let ST_{0kt}^* be the set of vertexes whose arcs form the Hamiltonian cycle of G_{kt}^* passing through i= 0. Note ST_{0kt}^* = {i ∈ N_{kt}^* : 0 ∈ N_{kt}^* } such a set. Building the sets N_{kt}^* and ST_{0kt}^* are described by the Fig. 2 and Fig. 3.

Then, let define ST_{kt}^*= {i ∈ N_{kt}^* : i ∉ ST_{0kt}^* } = N_{kt}^*/ ST_{0kt}^* with ST_{0kt}^* ∩ S T_{kt}^* ={Ø} and $N_{kt}^* = ST_{0kt}^* \cup ST_{kt}^*$. Two cycles have thus been defined. The main cycle ST_{0kt}^* and the ST_{kt}^* subtour to be eliminated. Detecting a subtour in the route of a vehicle k during period t is therefore equivalent to building ST_{kt}^*. if | ST_{kt}^* | ≥ 2 then a subtour is detected in the route of vehicle k during period t. To eliminate this subtour, the following procedure will be adopted.

```
$N^*_{kt}$ ←∅
    If ($Z^*_{0kt} > 0$) then
        foreach (i in N)
            If ($Z^*_{ikt} > 0$) then
                $N^*_{kt}$ ←i
            endif
        endfor
    endif
end
```

Fig. 2. The Procedure to Build N^*_{kt}.

```
$ST^*_{0kt}$ ←∅ ;  i←0
If ($Z^*_{0kt} > 0$) then
    repeat
        $ST^*_{0kt}$ ← i
        foreach (j in $N^*_{kt}$ and i≠ j)
            If ($x^*_{ijkt} > 0$ ) then
                i← j
            endif
        endfor
    until (i = 0)
endif
end
```

Fig. 3. The Procedure to Build Built ST^*_{0kt}.

D. Adding SECs

if $|ST^*_{kt}| \geq 2$ then the constraints $\sum_{(i,j)\in A_{kt}(N^*_{kt})} x_{ijkt} \leq |ST^*_{kt}| - 1 \ \forall k \in K, \forall t \in T$ (27) are added.

E. Adding RCs

if $|ST^*_{kt}| \geq 1$ then add the constraints $\sum_{j\in ST^*_{0kt}} x_{jikt} + \sum_{j\in N^*_{kt}/\{i\}} x_{ijkt}$ =2 z_{ikt} ∀ k ∈ K, ∀ t ∈ T, and i the first element of ST^*_{kt} (28) . Thanks to the constraints (28), the first element of ST^*_{kt} is connected to an element of the main tour (ST^*_{0kt}) and an element of N^*_{kt} until ST^*_{kt} be empty. the constraints (27) and (28) are used simultaneously according to $|ST^*_{kt}|$ for the total elimination of subtours and isolated vertexes.

F. Priority Order on Binary Variable

Tests on six branching orders allowed us to choose the following order: connection is made on the z_{ikt} variables first then to y_t variables and finally to the x_{ijkt} variables as in [24], [34].The B&C algorithm developed here can therefore be summarized in Fig. 4 as follows:

```
Initialize the upper bound U* and the incumbent solution.
Initialize the node pool N with the root node.
Generate and insert all known valid inequalities into
    the program at root node of the search tree.
  repeat
  Selection: Select the next node in N
             to evaluate and remove it from N
  Lower bound: Solve the LP relaxation
                    at the current node,
     let $U_l$ be the obtained lower bound
                  of the current node:
    if current solution is feasible then
        if $U_l$ >$U^*$ then
            go to the termination check.
        else
            U*←$U_l$ .
            Update the incumbent solution.
            Prune nodes with lower bound U > $U^*$
        end
   end
  Cut generation:
  foreach k in K
     foreach t in T
        if the current solution of
                   the LP relaxation
            contains $|ST^*_{kt}| \geq 1$ , then
            If  $|ST^*_{kt}| \geq 2$ , then
                Add corresponding SECs (27)
            endif
            Add corresponding RCs (28)
        endif
     endfor
  endfor
  Branching: if $U_l$ > $U^*$, go to
                  the termination check.
 until N=∅ or time limit is met (termination check)
Stop with the optimal solution and the corresponding
     cost $U^*$.
```

Fig. 4. The Procedure of Branch-and-Cut.

V. Two Phases Decomposition Heuristic (TPDH) for EDPRP Resolution

The decomposition method used in this work to solve the EDPRP follows the same approach as in [16]. the problem is solved in two phases. In the first phase, a LP of the LSP with direct shipment and direct collection (LSP_DS&DC) is resolved. At the end of this phase, production decisions (production and stored quantities) are set. Similarly, the quantities recovered at the plant or delivered by each vehicle to each customer over each period of the planning horizon are determined. The second phase is therefore to solve a TSP problem for each vehicle over each period of the planning horizon. So, a Genetic Algorithm (GA) is developed for the resolution of this second phase.

A. Phase I: Resolution of the LSP_DS&DC Model

$$Z= \min \sum_{t\in T}(up_t + fy_t) + \sum_{t\in T}\sum_{i\in N} h_i I_{it} \quad + \sum_{t\in T}\sum_{(i,j) j\in A(_N)}\sum_{k\in K} c_{ij}x_{ijkt} \quad (29)$$

$$\text{S.t } p_t \leq MU_t\, y_t \ \forall \text{ t} \in \text{T} \quad (30)$$

$$p_t = \sum_{k\in K} q_{0kt} \ \forall \text{ t} \in \text{T} \quad (31)$$

$$\sum_{k\in K} z_{n+1,kt} \leq \text{m}y_t \ \forall \text{ t} \in \text{T} \quad (32)$$

$$I_{0,t-1} + p_t = I_{0t} + \sum_{i\in N_c}\sum_{k\in K} q_{ikt} \ \forall \text{ t} \in \text{T} \quad (33)$$

$$I_{it-1} + \sum_{k\in K} q_{ikt} = d_{it} + I_{it} \ \forall \text{ i} \in N_c, \forall \text{ t} \in \text{T} \quad (34)$$

$$\sum_{i\in N_c}\sum_{k\in K} q_{ikt} \leq I_{0t-1} \ \forall \text{ t} \in \text{T} \quad (35)$$

$$I_{it} \leq L_i \ \forall \text{ i} \in N_{dc}, \forall \text{ t} \in \text{T} \quad (36)$$

$$\sum_{i\in N_c} q_{ikt} \leq \text{Q}\, z_{0kt} \ \forall \text{ k} \in \text{K}, \forall \text{ t} \in \text{T} \quad (37)$$

$$q_{0kt} \leq \text{Q}\, z_{n+1,kt} \ \forall \text{ k} \in \text{K}, \forall \text{ t} \in \text{T} \quad (38)$$

$$q_{ikt} \leq MC_{it}\, z_{ikt} \ \forall \text{ i} \in N_c, \forall \text{ k} \in \text{K}, \forall \text{ t} \in \text{T} \quad (39)$$

$$q_{n+1,kt} = 0 \ \forall \text{ k} \in \text{K}, \forall \text{ t} \in \text{T} \quad (40)$$

$$\sum_{k\in K} z_{ikt} \leq 1 \ \forall \text{ i} \in N_c, \forall \text{ t} \in \text{T} \quad (41)$$

$$x_{0ikt} + x_{i0kt} = 2\, z_{ikt} \ \forall \text{ i} \in N_{cu}, \forall \text{ k} \in \text{K}, \forall \text{ t} \in \quad (42)$$

$$x_{ijkt} = 0 \ \forall \text{ (i, j)} \in A(N_{cu}), \forall \text{ k} \in \text{K}, \forall \text{ t} \in \text{T} \quad (43)$$

$$p_t, I_{it}, q_{ikt} \geq 0 \ \forall \text{ i} \in \ , \forall \text{ k} \in \text{K}, \forall \text{ t} \in \text{T} \quad (44)$$

$$y_t, x_{ijkt}, z_{ikt} \in \{0,1\} \forall \text{ i, j} \in \ , \forall \text{ k} \in \text{K}, \forall \text{ t} \in \text{T} \quad (45)$$

To the mathematical model above, constraints (20), (21), (25), (26), $z_{jkt} \leq q_{jkt}$ and $z_{n+1,kt} \leq q_{0kt}$, ∀ j∈ N_c , ∀ k ∈ K, ∀ t∈ T (46) (to avoid unladen visits of vehicles.) are added.

The LSP_DS&DC model is solved according to the conditions set out in the experimental section. In the second phase," **" refers to components of the LP solution from phase 1 of the decomposition method. The results concerning the visit of the factory or each customer by each vehicle over each period (z_{ikt}^{**}) are reused to serve as an entry for phase II.

B. Phase II: Resolution of the TSP

The GA for each vehicle k at each period t (GA_{kt}) is described in Fig. 5. A vector of customers is used to model each chromosome. However, the depot and the plant are only considered in the evaluation of the fitness of the chromosomes. A Roulette Wheel is used for the selection operation. An operator with one crossing point is used to carry out the crossing of the chromosomes. The GA_{kt} is applied when | N_{kt}^{**} / {0, n+1} | > 2. And the total cost of transportation is equal to the sum of the transportation costs of each vehicle k for each period t. The cost of a direct collection to the plant or a trip for which 1≤ | N_{kt}^{**} /{0, n+1}| ≤ 2 (one or two customers visited) remains unchanged and therefore does not need an improvement by GA_{kt}. However, when |N_{kt}^{**}/{0, n+1}| > 2, the transportation cost for the vehicle k in the period t is equal to the fitness of the best chromosome of the last generation of the population of the Procedure of GA_{kt}.

```
Number_of_generation ← 0
Max_generation: initialise Max_generation
Termination criteria ← false
Generate initial random population
Repeat
    Evaluate fitness of each chromosome
    If Number_of_generation ≠Max_generation then
        Selection of parents for next generation
        Crossover of parents' chromosome
        Reparation of child's chromosome
        Mutation of chromosome
        Number_of_generation ←
                        Number_of_generation +1
    Else
        Termination criteria ← true
    End if
Until termination criteria = true
Choose the best chromosome
End
```

Fig. 5. The Procedure of of Genetic Algorithm on the Route of Vehicle k at Period t (GA_{kt}).

VI. Experimentations and Results

A. Experiments

The B&C algorithm described in Section 4 and the decomposition heuristic in Section 5 have been implemented in C++ with CPLEX 12.6 on a 64-bit Intel Pentium Dual Core 1.60 GHz, 1.60 GHz PC with 4 GB RAM. Only one thread was used in the experiments and the duration of each test is limited to 7200 seconds for the B&C and 3600 seconds for the decomposition heuristic. The purpose of these experiments is to evaluate the effectiveness of the model. The instances used in this work are derived from an adaptation of the instances used for the Multivehicle Production and Inventory Routing Problems (MVPRP). For more details on the instances used in Table II, see [34] .Since all deliveries to customers are made basing on the quantities in stock at the previous period in depot, an initial stock at the depot greater than zero is defined, unlike the basic instances. Moreover, the depot does not have the same geographical location as the plant. Thus, the initial stock levels at the depot for all instances are set to $I_{00} = 0.5 * (\frac{\sum_{t\in T}\sum_{i\in N_c} d_{it}}{l})$ and the plant's position is set to (0.0).

A total of 32 instances are applied to 4 classes of situations, i.e. a total of 32 x 4 = 128 instances. However, only the first 12 x 4=48 instances are used for the tests relating to the B&C algorithm. The first class consists of standard instances. The second and third class have the same characteristics as the first class. However, the second class has a high unit cost of production and the third class has a high transport cost. the fourth class is like the first and second class except for the unit inventory cost which is zero. The characteristics of the classes are summarized in Table III.

TABLE. II. CHARACTERISTICS OF THE INSTANCS FOR THE EPRP

n	l	m	C	L_0	Q
10	3/6	2	304	152	198
10	3/6	3	304	152	132
15	3/6	2	470	235	198
15	3/6	3	470	235	132
20	3/6	2	540	270	283
20	3/6	3	540	270	189
25	3/6	2	700	350	283
25	3/6	3	700	350	189
30	3/6	3	768	384	228
30	3/6	4	768	384	171
35	3/6	3	948	474	276
35	3/6	4	948	474	207
40	3/6	3	1256	628	360
40	3/6	4	1256	628	216
45	3	3	1438	719	360
45	3	4	1438	719	207
50	3	3	1348	674	360
50	3	4	1348	674	270

TABLE. III. INSTANCE CLASS CHARACTERISTICS

Classes	UPC	UIC	UTC
Class 1	Standard	Standard	Standard
Class 2	High	Standard	Standard
Class 3	Standard	Standard	High
Class 4	Standard	Null	Standard

B. Results

Let %diff represents the percentage difference between the cost determined by the heuristic method and the B&C method. diff is equal to the difference between the sums of the average costs obtained by the heuristic method and the B&C method divided by the sum of the average costs of the B&C method. In Table IV, each line corresponds to an average of 4 instance classes results for each number of customers n over l periods with m vehicles. the TOTAL Cost column represents the sum of the costs of the production cost (fixed and variable), the cost of inventories and the total cost of transport over l periods with m vehicles. The TOTAL Cost column represents the sum of the costs of the production cost (fixed and variable), the cost of inventories and the total cost of transport. the gap only available for the B&C refers to the percentage difference between the upper bound and the lower bound. the CPU defines the time taken to resolve the instances. 9 out of 48 instances have not been resolved and 19 out of 48 instances have not been resolved to the optimum. In the % Diff column, 7 out of 12 average results from the results of the B&C are better than those of the TPDH.

TABLE. IV. AVERAGE RESULTS OF B AND C VS TPDH

			B&C			**TPDH**				***%diff***
n	***l***	***m***	***TOTAL Cost***	***GAP***	***CPU***	***TOTAL Cost***	***CPU MIP***	***CPU AG***	***T.CPU***	
10	3	2	33581.75	10.06	4.19	34221.25	0.99	0.54	1.53	1.90
15	3	2	$23163^{\{2,3,4\}}$	13.57	18.27	20028	2.03	1.89	3.92	-13.53
20	3	2	$33700.5^{\{1,2\}}$	14.15	55.85	33463	1.15	2.18	3.32	-0.70
10	6	2	83259.5	0.34	5332.63^{2}	84116.5	9.97	0.91	10.88	1.03
15	6	2	112744.5	1.36	6698.81^{2}	111004.25	9.93	1.29	11.22	-1.54
20	6	2	$150748^{\{1\}}$	5.09	$7201.99^{(3)}$	151044	20.8	1.89	22.69	0.20
10	3	3	35600	3.18	12.9	35826.5	1.71	0.71	2.41	0.64
15	3	3	$48238^{\{3\}}$	0.59	126.18	47681	2.13	1.01	3.14	-1.15
20	3	3	$85802^{\{1,3\}}$	0.9	174.75	70477	1.51	1.52	3.03	-17.86
10	6	3	81877.5	2.56	7202.29^{4}	87065.25	13.03	0.76	13.79	6.34
15	6	3	112395.75	8.9	7201.79^{4}	115560.5	16.04	1.41	17.44	2.82
20	6	3	128464	15.36	7201.01^{4}	131567.5	2185.17	1.63	2186.8	2.42
Total			929574.5			922054.75				-0.81

[a] number of instances not resolved to the optimum

$^{\{-\}}$ unresolved instance class

However, the negative differences in instances 15_3_2 and instances 20_3_3 give an overall advantage to the results of the TPDH. The total difference between the TPDH result and the B&C is calculated as follows: TOTAL %diff = 100*((922054.75 - 929574. 5) / 929574.5) or TOTAL %diff = -0. 81. Based on the results of Table IV and TOTAL %diff it can be can said that the results of the resolved instances are globally close to 99.19% (100 -| TOTAL %diff |). Although the instances resolved with the B&C approach do not contain subtour, too many instances remain unresolved or unresolved to the optimum and the GAPs obtained are generally poor compared to the GAPs of the exact approaches described in Table I. However, it allows to make a comparison with the equivalent instances resolved by the TPDH. The Tables V, VI and VII describe respectively the averages results from the 128 instances with the TPDH, the percentages of production, inventory and transportation cost in the total production cost and the percentages of computation times. In these tables, Columns n, l and m respectively refer to the number of customers, periods and vehicles, the PROD column designates the total cost of production. this total production cost consists of the fixed production cost and the variable cost of production. Then, the INV column refers to the cost of inventories, the TRANS column represents the total cost of transport. The TOTAL Cost column is the sum of the costs of production, inventories and transportation. The MIP CPU, GA CPU and TOTAL CPU columns refer respectively to the time taken for the resolution of the LSP_DS&DC, the optimization of the transport part by the GA, and the sum of these two times. An average of 459,51 seconds (7.66 minutes) for computation time in the Table V is acceptable because EDPRP is a tactic levels problem. however, 99.32% of this calculation time is globally dedicated to the MIP and only 0.68% for the GA (Table VII). Thus, developing an GA or a memetic algorithm to solve the problem could considerably reduce the computation time of the different instances. The Table VI shows that a global percentage of 30.21% (8.59% + 21.62%) of the total cost is allocated to storage and transport.

TABLE. V. DETAILS OF AVERAGES TESTS RESULTS FOR TPDH

n	L	m	PROD	INV	TRANS	TOTAL Cost	CPU MIP	CPU AG	TOTAL CPU
10	3	2	23107.5	1827.75	9286	34221.25	0.99	0.54	1.53
10	3	3	23107.5	1827.75	10891.25	35826.5	1.71	0.71	2.41
15	3	2	30712.5	2693.25	13162	46567.75	2.64	1.73	4.37
15	3	3	30712.5	2693.25	15608.75	49014.5	2.02	1.03	3.05
20	3	2	35685	3321.75	13362.75	52369.5	1.24	1.98	3.22
20	3	3	35685	3321.75	15991.5	54998.25	1.38	1.44	2.82
25	3	2	39292.5	4389.75	15547.25	59229.5	1.96	2.79	4.75
25	3	3	39292.5	4389.75	16880.5	60562.75	1.69	2.60	4.28
30	3	3	45240	4209	19411	68860	5.93	3.71	9.64
30	3	4	45240	4209	22833	72282	10.85	3.35	14.19
35	3	3	66592.5	4662.75	21205	92460.25	4.10	3.64	7.74
35	3	4	66592.5	4604.25	26470.75	97667.5	4.68	3.68	8.35
40	3	3	69030	7533.5	22267	98830.5	3.66	5.34	9.00
40	3	4	69030	7533.5	26193.75	102757.25	4.34	4.69	9.02
45	3	3	91942.5	7550	24608.25	124100.75	5.21	6.36	11.57
45	3	4	91942.5	7874.75	31284.5	131101.75	7.07	5.05	12.12
50	3	3	73027.5	7767.75	29310.75	110106	8.56	8.91	17.47
50	3	4	73027.5	7767.75	33060	113855.25	9.69	5.82	15.52
10	6	2	63082.5	7430	13604	84116.5	9.97	0.91	10.88
10	6	3	63082.5	7451	16531.75	87065.25	13.03	0.76	13.79
15	6	2	82192.5	10974.75	17837	111004.25	9.93	1.29	11.22
15	6	3	82192.5	11055.75	22312.25	115560.5	16.04	1.41	17.44
20	6	2	93502.5	12653.25	20722.5	126878.25	18.20	2.01	20.21
20	6	3	94252.5	12502	24813	131567.5	2185.17	1.63	2186.80
25	6	2	104422.5	16560.25	21631	142613.75	19.72	3.23	22.94
25	6	3	104422.5	16281.25	26986.75	147690.5	41.22	3.01	44.23
30	6	3	118267.5	16857.75	33729.75	168855	140.01	3.03	143.04
30	6	4	118267.5	16728	40578	175573.5	1327.60	2.52	1330.12
35	6	3	145567.5	21020.5	36776	203364	2835.51	4.30	2839.81
35	6	4	145567.5	20479.5	45154.75	211201.75	3606.21	3.74	3609.94
40	6	3	177937.5	29374.25	39935	247246.75	1556.74	4.89	1561.64
40	6	4	177937.5	28849.25	44744.75	251531.5	2747.24	4.12	2751.36
Total			78748,5938	9887,33594	24147,82813	112783,7578	456,38	3,13	459,51

TABLE. VI. PERCENTAGE OF PRODUCTION, IVENTORY AND TANSPOTATION COST FOR TPDH

n	**PROD**	**%PROD**	**INV**	**%INV**	**TRANS**	**%TRANS**
10	689520	71.46%	74146	7.68%	201252	20.86%
15	903240	70.10%	109668	8.51%	275680	21.39%
20	1036500	70.84%	127195	8.69%	299559	20.47%
30	1308060	67.35%	168015	8.65%	466207	24.00%
35	1697280	70.17%	203068	8.40%	518426	21.43%
40	1975740	70.53%	293162	10.46%	532562	19.01%
45	735540	72.05%	61699	6.04%	223571	21.90%
50	584220	65.21%	62142	6.94%	249483	27.85%
Total	8930100	69.79%	1099095	8.59%	2766740	21.62%

TABLE. VII. PERCENTAGE OF CPU FOR TPDH

n	**CPU MIP**	**%CPU MIP**	**CPU AG**	**%CPU AG**
10	102.78	89.81%	11.661	10.19%
15	122.497	84.86%	21.852	15.14%
20	8823.906	99.68%	28.266	0.32%
25	258.308	84.74%	46.499	15.26%
30	5937.537	99.16%	50.422	0.84%
35	25801.949	99.76%	61.458	0.24%
40	17247.927	99.56%	76.156	0.44%
45	49.103	51.84%	45.626	48.16%
50	73.009	55.33%	58.942	44.67%
Total	58417.016	99.32%	400,882	0.68%

VII. CONCLUSION

This work focuses primarily on modeling a production routing problem in which the plant's storage capacity (depot) is geographically dissociated from the plant's location. The results of 48 tests with a B & C approach were compared with those of a decomposition heuristic method. The average results of all 128 instances are also presented in Table V. In this study, a depot-based distribution policy is adopted. However, the lack of storage capacity at the plant does not exclude the possibility to supply customers from the plant during production days. In future work, we will improve the B&C algorithm used in this paper by adding an initial solution generation heuristic to overcome the infeasibility problem and if possible, an improvement phase to reinforce the results. A GA or memetic algorithm will also be developed as an overall means of solving the problem instead of using a decomposition method.

REFERENCES

[1] Y. Pochet and L. A. Wolsey, Production planning by mixed integer programming. New York ; Berlin: Springer, 2006.

[2] J. K. Lenstra and A. H. G. Kan, “Complexity of vehicle routing and scheduling problems,” Networks, vol. 11, no. 2, pp. 221–227, 1981.

[3] L. Bertazzi and M. G. Speranza, “Inventory routing problems: an introduction,” EURO Journal on Transportation and Logistics, vol. 1, no. 4, pp. 307–326, Dec. 2012, doi: 10.1007/s13676-012-0016-7.

[4] C. Archetti and M. G. Speranza, “The inventory routing problem: the value of integration: The inventory routing problem: the value of integration,” International Transactions in Operational Research, vol. 23, no. 3, pp. 393–407, May 2016, doi: 10.1111/itor.12226.

[5] P. Chandra and M. L. Fisher, “Coordination of production and distribution planning,” European Journal of Operational Research, vol. 72, no. 3, pp. 503–517, Feb. 1994, doi: 10.1016/0377-2217(94)90419-7.

[6] M. Boudia, M. A. O. Louly, and C. Prins, “A reactive GRASP and path relinking for a combined production–distribution problem,” Computers & Operations Research, vol. 34, no. 11, pp. 3402–3419, Nov. 2007, doi: 10.1016/j.cor.2006.02.005.

[7] M.-S. Casas-Ramírez, J.-F. Camacho-Vallejo, R. G. González-Ramírez, J.-A. Marmolejo-Saucedo, and J.-M. Velarde-Cantú, “Optimizing a Biobjective Production-Distribution Planning Problem Using a GRASP,” Complexity, 2018. [Online]. Available: https://www.hindawi.com/journals/complexity/2018/3418580/abs/. [Accessed: 04-May-2018].

[8] Y. Adulyasak, J.-F. Cordeau, and R. Jans, “Formulations and Branch-and-Cut Algorithms for Multivehicle Production and Inventory Routing Problems,” INFORMS Journal on Computing, vol. 26, no. 1, pp. 103–120, Feb. 2014, doi: 10.1287/ijoc.2013.0550.

[9] Y. Adulyasak, J.-F. Cordeau, and R. Jans, “Optimization-Based Adaptive Large Neighborhood Search for the Production Routing Problem,” Transportation Science, vol. 48, no. 1, pp. 20–45, Feb. 2014, doi: 10.1287/trsc.1120.0443.

[10] V. A. Armentano, A. L. Shiguemoto, and A. Løkketangen, “Tabu search with path relinking for an integrated production–distribution problem,” Computers & Operations Research, vol. 38, no. 8, pp. 1199–1209, Aug. 2011, doi: 10.1016/j.cor.2010.10.026.

[11] Y. Qiu, L. Wang, X. Xu, X. Fang, and P. M. Pardalos, “A variable neighborhood search heuristic algorithm for production routing problems,” Applied Soft Computing, vol. 66, pp. 311–318, May 2018, doi: 10.1016/j.asoc.2018.02.032.

[12] M. Boudia and C. Prins, “A memetic algorithm with dynamic population management for an integrated production–distribution problem,” European Journal of Operational Research, vol. 195, no. 3, pp. 703–715, Jun. 2009, doi: 10.1016/j.ejor.2007.07.034.

[13] N. Absi, C. Archetti, S. Dauzère-Pérès, and D. Feillet, “A Two-Phase Iterative Heuristic Approach for the Production Routing Problem,” Transportation Science, vol. 49, no. 4, pp. 784–795, Nov. 2015, doi: 10.1287/trsc.2014.0523.

[14] C. Archetti, L. Bertazzi, G. Paletta, and M. G. Speranza, “Analysis of the maximum level policy in a production-distribution system,” Computers & Operations Research, vol. 38, no. 12, pp. 1731–1746, 2011.

[15] P. Chandra, “A dynamic distribution model with warehouse and customer replenishment requirements,” Journal of the Operational Research Society, vol. 44, no. 7, pp. 681–692, 1993.

[16] L. Lei, S. Liu, A. Ruszczynski, and S. Park, “On the integrated production, inventory, and distribution routing problem,” IIE Transactions, vol. 38, no. 11, pp. 955–970, 2006.

[17] M. Boudia, M. A. O. Louly, and C. Prins, “Fast heuristics for a combined production planning and vehicle routing problem,” Production Planning & Control, vol. 19, no. 2, pp. 85–96, Mar. 2008, doi: 10.1080/09537280801893356.

[18] H. M. Wagner and T. M. Whitin, “Dynamic Version of the Economic Lot Size Model,” Management Science, vol. 5, no. 1, pp. 89–96, Oct. 1958, doi: 10.1287/mnsc.5.1.89.

[19] G. Clarke and J. W. Wright, “Scheduling of Vehicles from a Central Depot to a Number of Delivery Points,” Operations Research, vol. 12, no. 4, pp. 568–581, Aug. 1964, doi: 10.1287/opre.12.4.568.

[20] C. Archetti, L. Bertazzi, G. Laporte, and M. G. Speranza, “A branch-and-cut algorithm for a vendor-managed inventory-routing problem,” Transportation Science, vol. 41, no. 3, pp. 382–391, 2007.

[21] M. Ruokokoski, O. Solyali, J.-F. Cordeau, R. Jans, and H. Süral, “Efficient formulations and a branch-and-cut algorithm for a production-routing problem,” GERAD Technical Report G-2010-66, 2010.

[22] Y. Qiu, M. Ni, L. Wang, Q. Li, X. Fang, and P. M. Pardalos, “Production routing problems with reverse logistics and remanufacturing,” Transportation Research Part E: Logistics and Transportation Review, vol. 111, pp. 87–100, Mar. 2018, doi: 10.1016/j.tre.2018.01.009.

[23] Y. Qiu, J. Qiao, and P. M. Pardalos, “Optimal production, replenishment, delivery, routing and inventory management policies for products with perishable inventory,” Omega, Jan. 2018, doi: 10.1016/j.omega.2018.01.006.

[24] Y. Qiu, L. Wang, X. Xu, X. Fang, and P. M. Pardalos, “Formulations and branch-and-cut algorithms for multi-product multi-vehicle production routing problems with startup cost,” Expert Systems with Applications, vol. 98, pp. 1–10, May 2018, doi: 10.1016/j.eswa.2018.01.006.

[25] Y. Adulyasak, J.-F. Cordeau, and R. Jans, “Benders Decomposition for Production Routing Under Demand Uncertainty,” Operations Research, vol. 63, no. 4, pp. 851–867, Aug. 2015, doi: 10.1287/opre.2015.1401.

[26] M. Gendreau, G. Laporte, and F. Semet, “A branch-and-cut algorithm for the undirected selective traveling salesman problem,” Networks, vol. 32, no. 4, pp. 263–273, Dec. 1998, doi: 10.1002/(SICI)1097-0037(199812)32:4<263::AID-NET3>3.0.CO;2-Q.

[27] J. Lysgaard, A. N. Letchford, and R. W. Eglese, “A new branch-and-cut algorithm for the capacitated vehicle routing problem,” Mathematical Programming, vol. 100, no. 2, pp. 423–445, 2004.

[28] D. Applegate, R. E. Bixby, V. Chvátal, and W. J. Cook, “The Concorde TSP Solver website. http://www.math.uwaterloo.ca/tsp/concorde.html.,” 2011.

[29] D. Applegate, R. E. Bixby, V. Chvátal, and W. J. Cook, “Concorde: A code for solving traveling salesman problems. http://www.tsp.gatech.edu/concorde.html.,” 2005.

[30] M. Padberg and G. Rinaldi, “A Branch-and-Cut Algorithm for the Resolution of Large-Scale Symmetric Traveling Salesman Problems,” SIAM Review, vol. 33, no. 1, pp. 60–100, Mar. 1991, doi: 10.1137/1033004.

[31] Y. Adulyasak, J.-F. Cordeau, and R. Jans, “The production routing problem: A review of formulations and solution algorithms,” Computers & Operations Research, vol. 55, pp. 141–152, Mar. 2015, doi: 10.1016/j.cor.2014.01.011.

[32] D. L. Applegate, R. E. Bixby, V. Chvatal, and W. J. Cook, The traveling salesman problem: a computational study The Traveling Salesman Problem: A Computational Study. Princeton University Press, Princeton, NJ. Princeton university press, 2007.

[33] Y. Kocaoğlu, A. Taşkın Gümüş, and B. Kocaoğlu, “Supply chain optimization studies: A literature review and classification,” 2018.

[34] Y. Adulyasak, J.-F. Cordeau, and R. Jans, “Formulations and Branch-and-Cut Algorithms for Multivehicle Production and Inventory Routing Problems,” INFORMS Journal on Computing, vol. 26, no. 1, pp. 103–120, Feb. 2014, doi: 10.1287/ijoc.2013.0550.

[35] L. C. Coelho and G. Laporte, “A branch-and-cut algorithm for the multi-product multi-vehicle inventory-routing problem,” International Journal of Production Research, vol. 51, no. 23–24, pp. 7156–7169, Nov. 2013, doi: 10.1080/00207543.2012.757668.

Machine Learning based Access Control Framework for the Internet of Things

Aissam Outchakoucht[1], Anas Abou El Kalam[2], Hamza Es-Samaali[3], Siham Benhadou[4]
LISER Laboratory, Hassan II University, ENSEM School, Casablanca, Morocco, IPI, Paris, France[1, 3, 4]
Cadi Ayyad University, ENSA School, Marrakech, Morocco[2]

Abstract—**The main challenge facing the Internet of Things (IoT) in general, and IoT security in particular, is that humans have never handled such a huge amount of nodes and quantity of data. Fortunately, it turns out that Machine Learning (ML) systems are very effective in the presence of these two elements. However, can IoT devices support ML techniques? In this paper, we investigated this issue and proposed a twofold contribution: a thorough study of the IoT paradigm and its intersections with ML from a security perspective; then, we actually proposed a holistic ML-based framework for access control, which is the defense head of recent IT systems. In addition to learning techniques, this second pillar was based on the organization and attribute concepts to avoid role explosion problems and applied to a smart city case study to prove its effectiveness.**

Keywords—Access control; internet of things; machine learning; security; smart city

I. Introduction

Access Control (AC) plays a pivotal role in the security world given its mission of protecting digital and physical accesses by delimiting and enforcing who has access to what and in which conditions [1]. However, most of the AC solutions we find in the literature tend to consider the IoT as a single block that is characterized mainly by the limited storage and computing capacities. In this paper, we will come back to this unfair and unrealistic view that slows down the elaboration of a holistic approach to address AC in IoT environments. Moreover, relying on a single technique to address an issue that is as complex as IoT is also a weakness that confines the performance of many IoT security-oriented models.

To fulfil the AC requirements, this paper will progressively build a global framework that not only focuses on policy management and AC models, but also digs deeper into the mechanisms that accurately fit them; which leads to a smooth and coherent Machine Learning (ML) integration going down to highlight what and where ML algorithm(s) should be implemented.

To do so, we first need to delimit the perimeter covered by the IoT by giving a much more representative definition of the term, which will allow us later to tackle the question of AC with a much more appropriate vision, and above all, will lead us to know where and how we can use the power of ML to take advantage of the large amount of objects and data we are handling.

Motivated by the above, we perceive the relationship between IoT and ML much like the relationship between the human body and its brain. Our bodies gather sensory input such as sight, sound, smell, taste and touch while our brains focus on gathering that data and making sense of it.

The remainder of this paper is presented as follows: Section II exposes an overview of ML applications in IoT scenarios; then Sections III and IV reveals the building blocks of the ML-based framework aiming to handle IoT AC, as well as all the required concepts to understand it. Next, Section V provides the theoretical and technical details of implementation which are applied to a smart city case study, before moving to the last section in which we discuss and evaluate the results.

II. Related Works

Basically, ML algorithms are computer programs that can essentially learn and improve their accuracy by looking at data without being explicitly programmed. In a more formal wording: “A Computer program is said to learn from an experience *E* with respect to some task *T* and some performance measure *P*, if its performance on *T*, as measured by *P*, improves with experience *E*” [2, 3]. In this section, we will be exploring the most promising and latest ML applications used in order to secure IoT environments.

A. Learning Algorithms for Constrained Environments

Despite the common sayings that build a delusive wall between ML algorithms and constrained nodes given the computational and storage limitations of the latter, many studies combined these two worlds in order to answer both application and security issues. In this section we are about to discuss the most recent and relevant ones.

Let us begin with a recent work [4] that combined the strengths of current neural and tree-based learning techniques in conjunction with ternary (-1, 0, 1) quantization to enable computation and size compression of NN models in IoT platforms. This technique outperformed the state-of-the art by 11.1%, 52.2% and 30.6% in the number of computations, the model size, and the overall memory footprint respectively, without losing too much in terms of accuracy.

Another study [5] focused on the IoT device side rather than the cloud. The proposed model is developed on a relatively simple tree learnt in a low-dimensional space for efficient prediction on IoT nodes like Arduino UNO or ATmega328P boards with 16 MHz, 2 KB RAM and 32 KB ROM. The authors executed their model using several datasets

This work was supported by the IGS Group, Paris, France

and proved that it was able to make predictions within milliseconds, had lower battery consumption than the state-of-art and could fit in KB of memory.

Additionally, a model called MorphNet was suggested by Google [6] to automate the design of Deep Neural Networks (DNNs). This approach is specifically adjustable to meet constrained environments' requirements without compromising the performance, it actually optimizes the DNN by iteratively shrinking and expanding. The study showed that MorphNet is simple to implement and fast to apply, which is why it is a better choice for IoT scenarios.

ShuffleNet [7] is another contribution in this direction. It is a particularly computation-efficient CNN designed specifically for mobile devices which are essentially characterized by their limitations in terms of computational power. It is mainly based on the power of 1×1 convolutions combined with channel shuffle with the aim of reducing the required cost for computation without neglecting the accuracy. Being implemented on an ARM-based mobile device, the model attains up to 13× actual speedup over AlexNet.

Besides, the authors in [8] suggested a NN-based implementation that takes benefit from the communications passed between IoT nodes. Theoretically, this work is founded on the UAT theorem affirming that a NN with a single hidden layer is enough to compute a bounded approximation of a generic continuous function. In fact, this remark has led to integrate intelligence into IoT constrained platforms by means of some (local and on-the-fly) computations as the data navigate between the IoT devices using the collective behavior of such networks.

A Mobile and Edge Computing (M/EC) solution was proposed in [9] to bring computation near the IoT end-nodes by applying CNNs, RNNs and RL at the edge of IoT networks. The very idea of this work is to implement Information-Centric Networking on top of the IoT via some techniques namely shared weights, pooling, and in-network caching to solve storage issues on IoT nodes. This approach led to remarkable reduction in latency for time-critical applications.

Another study [10] digs deeper into the technical hardware requirements to implement DL algorithms over IoT devices. The authors implemented several models inside three boards: Qualcomm Snapdragon 800 used for phones and tablets (4-core 2.3 GHz CPU, 1GB of RAM and 8MB DSP), Intel Edison principally oriented to wearables and form-factor sensitive IoT (500MHz dual-core CPU, 1 GB of RAM) and finally Nvidia Tegra K1 used for example in June IoT Oven [11], Nexus 9 phone and IoT-enabled cars (up to 1.7GB of RAM). The study proved, inter alia, the feasibility of implementing DL techniques on IoT oriented boards.

A more general approach was proposed in [12], in which the authors came with a semi-supervised deep RL model designed to fit smart cities. Its inference engine exploits Variational Auto Encoders (VAE) to generalize optimal policies. The model was implemented to handle localization issues in a smart building case study portrayed as an ensemble of labeled positions associated with the Received Signal Strength Indicator values from multiple iBeacons. It was able to learn better action policies with at least 23% improvement in terms of distance to the target as well as almost 67% more gathered rewards compared to the supervised Deep RL approaches.

B. Learning Applications for IoT Security

Now that we have seen many ML-based applications in the IoT, let us move to some studies that tackled security problems (always in IoT environments) through ML tools.

In fact, Support Vector Machines (SVMs) are one of the first and most used ML models. They represent standard classification models, generally known for splitting hyperplanes. Data sorting is achieved through maximizing the distance between the hyperplane and the nearby training samples of each class. SVMs are more adapted to datasets with a large number of features but a relatively small number of samples [13]. In the IoT world, a study [14] proposed a linear SVM-based Android malware detection system to secure IoT platforms. The comparison that was led between the performance of the model and other ML algorithms outbalanced the SVM method. Besides, SVM was also used to compromise cryptographic devices [15, 16]. However, one of the big challenges in multidimensional SVM problems is the tough task of selecting a suited kernel.

Another generic method is Random Forest (RF): an accumulation of Decision Trees (DTs), which means that they are built and trained in order to vote for the output class [17]. A study [18] over 17 IoT devices belonging to nine categories affirmed that RF (among other ML methods) presents significant improvements in the identification of unauthorized IoT nodes. Another ML-based study [19] was performed on IoT environments to detect DDOS attacks. In this regard, RF provided slightly superior results compared to other ML methods. That being said, it is important to emphasize that RF methods are not always feasible, specifically over large datasets as they require the construction of a -relatively- large number of DTs.

In another direction, UL is represented by the popular K-Means with the key objective of Data clustering (k being the number of clusters). The algorithm consists of assigning each data sample to one of the k clusters based on their (similar) features. Usually UL models are privileged when the dataset is not labelled. In IoT, k-means clustering was used to distinguish Sybil attackers from normal sensors through clustering the channel vectors in industrial WSNs [20]. Nevertheless, this technique has many limitations, namely the need to have roughly equal numbers in each cluster for the algorithm to properly work, as well as the non-trivial task of choosing k [21].

Now, let us move to the deep sphere, and begin our survey by Convolutional Neural Networks (CNNs). Actually, the basic idea of a CNN is to put a bit of structure in NNs [22] by shrinking the enormous number of connections between layers, thus optimizing the training time. One of the main benefits of CNNs is their end-to-end nature ensured by their built-in "features extraction" ability. Yet CNNs still have a high computational cost; hence the difficulty of implementing them on IoT constrained nodes. Many studies managed to bypass this limitation though, especially using distributed architectures [23]. Another study showed that CNNs could help in Android malware detection by means of raw sequence static analysis (RSSA) of disassembled programs [24].

One more giant pillar of ML nowadays goes under the name of Recurrent Neural Network (RNN). It is undoubtedly one of the ML big discoveries thanks to their particularity of having memory. They can read inputs $\boldsymbol{X}^{<t>}$ in sequence, and "remember" some information/context thanks to their hidden layer activations that get passed from a given time-step to the following [22]. Accordingly, RNNs can achieve excellent results in classifying network traffic and detecting malicious behavior. Besides, RNNs could be a good choice for IoT since it produces massive sequential data from different nodes. For instance, a previous work [25] proved the worth of using RNNs to detect network traffic behavior by modeling it as a sequence of states changing over time. Yet, vanishing and exploding gradient problems still the ultimate nightmare of RNNs.

In another direction, many researchers consider that there was various contributions in ML in recent times, but Generative Adversarial Networks (GANs) are the only contribution that could be called a breakthrough in the last decade. GAN trains two models at the same time: a generative model G to identify the data distribution, and a discriminative model D to predict the probability that a sample came from the training data rather than G [26]. A recent work [27] realized a GAN-based architecture in order to secure IoT systems by detecting abnormal behavior. GANs may have a potential application in IoT security especially in zero-day-like threats given their ability to learn diverse attack scenarios and then to generate innovative attacks beyond the existing ones. Though, up to now the training phase of GANs still unstable and a tough task [21].

Providing a large amount of training data is not always an easy task; hence, finding alternatives is a matter of serious concern for ML experts. Reinforcement Learning (RL) consists of learning behavior only through interactions between an agent (usually represented by the algorithm) and its surrounding environment; in fact this learning process consists of increasing the rewards it receives from the environment. Many researchers focus on the application of RL to IoT security; for instance, the work in [28] opted for an RL approach to learn a sub-band selection policy so that it could avoid both jammer signals as well as interference from other radios in wideband autonomous cognitive radios (WACRs). Two of our previous works [1, 29] tackled the Access Control (AC) in IoT scenarios, the two building blocks were: first taking into account the smart devices' context while making an AC decision; and proposing AC policies that can be improved and optimized over time. However, given the enormous and heterogeneous amount of data generated by IoT devices, the proposition benefits from the power of RL, to accomplish this task. The problem with RL algorithms is that they require a large number of practice run (given their trial-error nature) before they can make significant progress.

It is worth noting that, in addition to provide an explicit survey of the latest and relevant works in IoT and ML, one of the motivations of this related works section is to prove that ML is already used in the IoT world, and consequently to disprove the idea claiming that IoT and ML are two parallel universes. In the following section, the proposition of this paper is presented with all the necessary details.

Table I summarizes and compares these studies especially based on their achievements in a number of IoT situations.

TABLE. I. COMPARISON AND SUMMARY OF ML STUDIES FOR IOT

Application domains	Used algorithms	Achievements	Studies
Networks traffic optimization in indoor or smart city scenarios	Semi-supervised D-RL Auto Encoders D-NN, CNN, RNN	Bring computing next to IoT devices Reduction in latency More gathered rewards	[9], [10], [12]
(Relatively) Simple processing situations	D-NN, CNN Low-dimensional trees	Local & on-the-fly computations Up to 13× actual speedup over AlexNet Decreasing model size, memory consumption Prediction within milliseconds	[4], [5], [6], [7], [8]
Malware & Intrusion Detection	SVM, RF, CNN	Significant improvements in the identification of unauthorized IoT nodes Compromising cryptographic devices	[14], [15], [16], [18], [19], [23], [24],
Network threats & Network traffic behavior	RNN, GAN	Zero-day attacks Good results in time-based environments Data augmentation	[25], [27]
Security policy improvements	RL	Policy optimization Policy efficiency	[1], [28], [29]
New and unprecedented attacks	K-means, RL	Zero-day & Sybil attacks detection Avoid jammer signals	[1], [20], [28], [29]

III. Preliminaries

We begin this section by exposing the research questions standing behind our work, together with all the essential details required to understand our proposition.

A. Problematic

The nature of a large portion of IoT devices (e.g. Healthcare, critical infrastructures) makes security the number one priority: it is, literally, a matter of life and death. In addition, the density, heterogeneity and autonomy are intrinsic characteristics of these systems that not only expend the perimeter of potential attacks but also their magnitude [21].

However, it turns out that treating all security aspects in one proposition is not a reasonable task; hence, this paper is focusing on the AC cornerstone because of its nature of protecting the access to digital and physical resources by delimiting, managing and enforcing who has (has not, is obliged to have) access to what, when and under which conditions [1].

Furthermore, the abovementioned IoT features impose intelligent and dynamic management instead of traditional and impractical one; we believe that IoT must benefit from these features considered so far as obstacles, IoT nodes need to “learn to look after each other”. To do so, our research has confronted many speed bumps that we have tried to demolish in this article.

First, the misleading idea that reduces the IoT to constrained devices, but more importantly: The need to exploit the quantity of IoT devices and data as a catalyst for security to emerge.

Not to mention the necessity for models that go beyond simply defining AC policies to understand the context of each smart device and continuously improving these policies, without falling into the trap of static management or role explosion.

To the best of our knowledge, there seems to exist no previous work presenting a holistic ML-based framework for IoT answering these problematics.

B. IoT and Computation Paradigm

At first sight, IoT is a concatenation of two words: “Internet”, which refers to connectivity and communication aspects; and “Things”, which is a generic and global term that includes all kinds of objects, whether large or small, powerful or not. In that sense, every “thing” endowed with communication capacity is an IoT device. Put this way, one can easily classify the aforementioned constrained nodes, together with mobile phones, a Raspberry Pi board and cloud servers as IoT devices; and can also set traditional TVs, calculators or pillows outside the IoT scope (unless they are connected).

Actually, no one can deny the difficulty (sometimes even the impossibility) of implementing complex ML tools on several types of constrained nodes. Yet, these latters remain ambiguous especially given the recent innovations in ML-oriented chips, which are mainly due to the excessive demand and hot market of AI applications these days. This subsection exposes three reasons to motivate researchers –and investors– not to draw a spontaneous correlation between IoT and ML ineptness:

- IoT > constrained nodes: As explained before, one key idea to clarify when talking about IoT is that it is more than just a collection of constrained nodes. Not to confound with Wireless Sensor Networks.
- Hardware progress: The AI market is in an exponential growth, which leads to more investments, then to more innovations. This climate could only be beneficial for ML community. With this in mind, one can take a look at Amazon store for example to see how the ratio of hardware size to its storage capacity is decreasing faster than ever before. Regarding computation capacities the progress is astonishing as well, for instance, just few months ago, NVIDIA announced a 70mm x 45mm AI computer, 4 GB memory, Quad-core ARM® A57 CPU and 128-core NVIDIA Maxwel GPU [30].
- Software evolution: What is true for hardware also holds for software. Section II is an illustration of the active race in proposing new and suitable ML algorithms for IoT. Besides, many dedicated and extremely optimized ML libraries are already easing programmers’ life. For instance, Google’s Tensorflow Lite transforms heavy TensorFlow models into compressed flat buffers, which are then loaded into a mobile or embedded device [31].

Furthermore, there are several active research directions that could lead to further findings (even breakthroughs) in IoT-adapted ML applications: (i) parallel computing in training phase using Graphical Processing Units and Tensor Processing Units (GPUs/ TPUs), (ii) transfer learning in order to swiftly transfer the knowledge from pre-trained models, (iii) fog computing to decrease communications overhead size, data traffic, user-side latency, (iv) fast optimization algorithms [32].

C. Background

The aim of this section is to explore two concepts that are essential to understand our proposition, namely, AC and IoT architecture.

In fact, AC is of paramount importance being the entry point of every system after the identification/ authentication phase, thus securing any system must pass through (if not begin with) controlling its accesses. In the literature, tens of models are handling this issue, one of the most popular is Role Based Access Control [33] (RBAC); without going into too much details: instead of granting (or removing) a separate permission to every subject in the network, the model aggregates these subjects by roles and thus gets a lightweight version of its AC policy. Yet, end devices are not involved in AC decision, also, even with the aggregation of subjects into roles, IoT systems still have astronomical number of objects and actions to manage. Attribute Based Access Control [34] (ABAC) is another model that is taking up more and more space recently, its basic concept is to identify subjects and objects through attributes (characteristics), then the permissions are granted according to these attributes, which could be any relevant security-characteristics, this makes ABAC, unlike RBAC, more adapted to afford fine-grained AC highly valued in IoT circumstances. Still, it comes with a terrifying and intolerable drawback: complexity. Other models [35] tried to make many tradeoffs but generally fall into one of the aforesaid limitations.

Reasonably, Organization Based Access Control [36] (OrBAC) is a remarkable model that we believe it could be used as a foundation to an IoT-oriented solution. It inherits and extends the benefits of RBAC by proposing abstractions to all the AC elements (subject to role, object to view and action to activity – see Fig. 1), then it adds two more dimensions to the decision making process: context (crucial for IoT) and organization; this latter makes it a centralized model, but we will see in the next section how to turn this limitation into a strength.

In the other hand, and given what is been elucidated earlier in this paper, IoT is not a single homogenous block; hence, in order to propose reasonable solutions, there is no other way but to have a conception of the main building blocks composing standard IoT environments. To do so, of course one could suggest several subdivisions but since the intention of this paper is to conciliate IoT and ML, the proposed categorization needs be centered on computing capabilities.

In that sense, many researchers [37, 38] agree that every IoT platform could be molded into one or more of the following categories: C1: the constrained layer, here is where the constrained devices are located (physical constraints on many characteristics such as size, weight, available power and energy [39] which make them unable to accomplish more than basic tasks); C2: this category includes more powerful nodes, which are capable of executing relatively serious computations. In fact, the vast majority of the everyday smart devices fit into this category (e.g. smartphones, smart homes components); finally, C3: Computational or offloading layer, it is the most powerful one, it could be Cloud or local servers, computers, GPUs/TPUs and so on.

Generally, complex environments (like smart cities) are a combination of the three layers. Another point to underline is the absence of explicit and rigid boundaries between these layers; this is mainly due to the dynamic nature of the IT domain, in fact what we call powerful today (or for a given task) might be considered constrained tomorrow (or for another task) and vice versa. Based on some examples, Fig. 2 exposes the borders and intersections between these categories.

As shown above, typically what most people call IoT is in general the intersection between the three layers, it is where all sorts of devices are interacting with each other to create this large universe of smart devices.

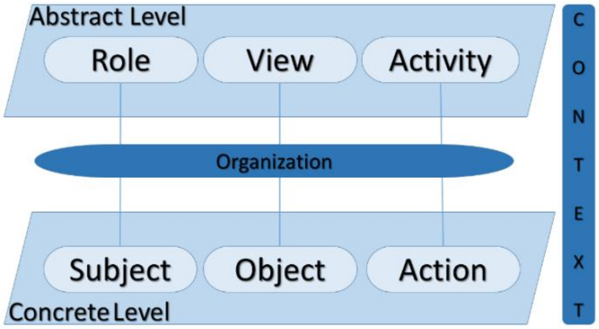


Fig. 1. Simplified Presentation of OrBAC Layers.

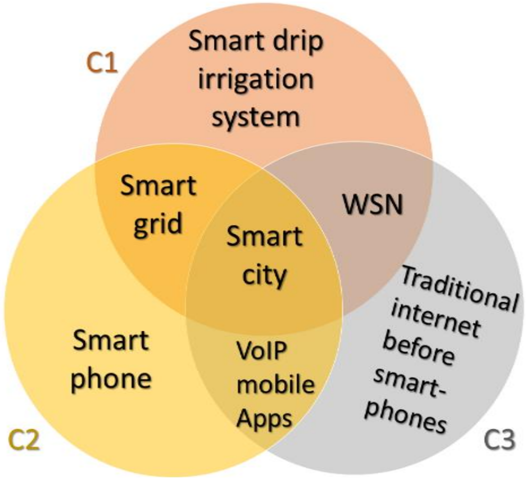

C1
Smart drip irrigation system
Smart grid
WSN
Smart city
Traditional internet before smart-phones
Smart phone
VoIP mobile Apps
C2
C3

Fig. 2. Intersections between IoT Layers.

IV. Contribution

In this section we present the AC framework that takes into consideration all the previously discussed requirements.

A. Global Questions Need Global Answers

One can fairly claim that IoT is the largest and most heterogeneous artificial network humans have ever made, it is becoming earth's nervous system. Therefore it would be naive, if not irrational, to search for elementary narrowed solutions to such a complex and multifaceted problem; instead what this paper is suggesting is a multi-layer AC solution, that exploits ML and OrBAC strengths to answer IoT burning questions exposed in Section III-A.

First thing to consider when treating AC in IoT is the overall architecture, whether it is distributed (like blockchain based solutions [40]) or centralized as long as it is equipped with the collaboration aspect. In fact, each of these two architectures has its advantages and drawbacks [29]; however, without loss of generality, this paper mainly focus on the second one for the following reasons: (i) IoT, at least as we know it today, requires that each device has an owner, whether a person, in the case of a smart home for example or an organization in industry or in smart cities. So practically an approach that takes this aspect into consideration will be closer to reality. (ii) Even in distributed architectures, there always should be an entity in charge of defining AC policies for the IoT nodes, unless the device is open to everyone (and therefore no need for AC at all). For simplicity reasons, in both scenarios this entity is given the name "object owner" (*OO*).

As shown in Fig. 3, the object owner *i* (OO_i) owns several IoT devices which can belong to any category (C1, C2 or C3). He defines their AC policies and stores them either locally, in a distributed manner on multiple servers or even in large networks (like blockchain) using smart contracts. Henceforth, the location of these policies is called Policy Information Point (PIP), as defined by the ISO/IEC standard for the access control framework [41] and the XACML related architecture [42].

As mentioned before, the choice of OrBAC as the background model to an IoT-oriented framework is defended by the following reasons: First, it has the concept of organization by design thus no need for extra dimensions to designate the *OO*. Also, OrBAC is distinguished by two other features crucial for IoT environments, namely an advanced level of abstraction required to alleviate the complexity produced by the colossal number of devices; together with the context incarnated in all OrBAC rules, which will ease the collect of real time contextual information from the end nodes for better AC decisions. In a more formal sense, the AC policy is stored in the PIP as rules presented in this form:

$$permission(org, r, v, ay, c) \quad (1)$$

Where *org* stands for organization or the view owner, *r* for role (aggregation of subjects), *v* for view as a collection of objects, *ay* for activity which is an abstraction of actions whereas *c* is the context. Thus the previous rule declare that: In the organization *org* the role *r* has permission to execute the activity *ay* on the view *v* under *c* circumstances.

1) Pre-request stage: Policy initiation: In step one, when the *OO* have to define a new rule, either the device *o* fits in one of the existing views *v* so the affectation : $use(org, o, v)$ (2) is executed; if not the role is automatically created when declaring a new permission and the subject and role get the same name.

Another key concept of this framework is the process of matching abstract and concrete entities, it also begins in phase one: In fact, besides the aforementioned rules, the PIP contains two types of match functions:

Usual correspondences in the form of $match(org, s_j, r_i)$ as shown in (2) are employed for the frequently used entities, while generic match functions based on attributes, for instance:

$$match(org, s^1, \ldots, s^n, r_i) \quad (3)$$

Which means if a subject *s*, has the attributes $s^1, \ldots, s^n$ then it belongs to the role r_i. Yet since the designation of s_j could as well be considered as an attribute of itself, we can use (3) for both.

2) Inference stage request processing: Now that AC policies are initiated and stored in the PIP, phase 2 begins: a subject s_i is willing to execute an action a_i on an object o_i, to do so the following request is sent to the Policy Enforcement Point (PEP):

$$get\_access(org, s_i, o_i, a_i, c_i) \quad (4)$$

When the PEP gets the access request, it triggers the process of matching (phase 3); it is the step where we go up from concrete to abstract entities in order to reduce the complexity. So the PEP transfers (4) to the Policy Decision Point (PDP), which in its turn requests the PIP by an:

$$get\_match(org, s_i, o_i, a_i, c_i) \quad (5)$$

After that, the PIP, responds by the corresponding matches:

$$back\_match(org, r_i, v_i, ay_i, C_i) \quad (6)$$

Up till now, the PDP has all the static features to take the decision. However, even if we have already handled two IoT major worries, namely context and complexity, the process still lacking dynamism. In fact, policies are statically stored in the PIP without any learning from past experiences. To answer this, (6*) will be coupled with an extra feature: a ratio reflecting the probability of a safe access granting given the aforementioned characteristics:

$$back\_match(org, r_i, v_i, ay_i, C_i, p) \quad (7)$$

To do so, (1*) needs to be joined with the probability feature which is set by default (in the policy definition phase) to *p=1*, then it is updated over time:

$$permission(org, r, v, ay, c, p) \quad (8)$$

Now, and based on a threshold fixed by the *OO*, the PDP can decide (phase 4) and inform the PEP, and consequently the requester, about the final decision. Note that this threshold has two essential benefits: first it gives the organization a better personalization of the framework, in addition to allowing it to define even several thresholds given the criticality of certain resources or context.

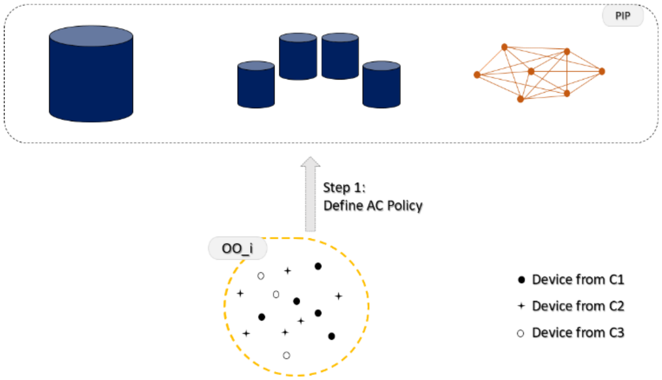


Fig. 3. Step 1: to define AC Policies.

3) Post-request stage learning and upgrading: The final phase -5- in our process is a post access one, it consists of calculating a feedback rating of the experience, which will be used by the PDP to generate AC policy updates leading to more accurate decision in the future (a concrete example is given in the case study section). The output will be stored in the "*learning matrix*" which will have 6 columns (organization, role, view, activity, context, feedback) and as many line as the number of experiences the system could store; while feedback is a rational number between 0 and 1.

The learning algorithms run in phase 5 varies according to the hardware resources of the system but also according to the layers defined in Section III-C. However, generally in complex and multifaceted IoT environments we propose using: RL and many resource consuming SL scheduled, for example, periodically in category C3; SL up to a reasonable size of the learning matrix in C2; while leaving normal equation stechnique or no ML at all to C1. Table II summarizes the previously detailed steps.

For simplicity reasons, we were focusing in permission formulas, however what goes for permissions is also valid for obligations and prohibitions [43].

B. The Algorithm

The steps discussed in Table II are compressed in the following Fig. 4 to explain the overall functioning of the algorithm. In fact, the framework could be segmented into three main time frames: (i) pre-request tasks, which handle the definition of the AC policies; (ii) request processing, involving all the actions triggered after an access request up till the subject receives back a permission/rejection; (iii) post-request actions that are responsible for the learning and policy improvements.

Note that the proposed framework is a decentralized one. In fact, the concept of organization is introduced to decompose complex IoT environments into reasonable and manageable groups, not to turn them into one giant centralized one. For instance, if we have to manage AC in a smart city situation, the framework will treat this multidimensional platform as a collection of organizations interacting and collaborating with each other.

Several studies have examined the collaboration issues in OrBAC [37, 44, 45, 46], either by creating further abstract entities, web services, or even through prior agreements between the involved organizations, however the definition of AC relationships using attributes that we saw in 1.b. (Table II) is a better alternative in IoT situations since with one tool we answer both intra and inter organizations AC concerns. An example of this scenario will be discussed in the following section, where we are exploring a smart city case study.

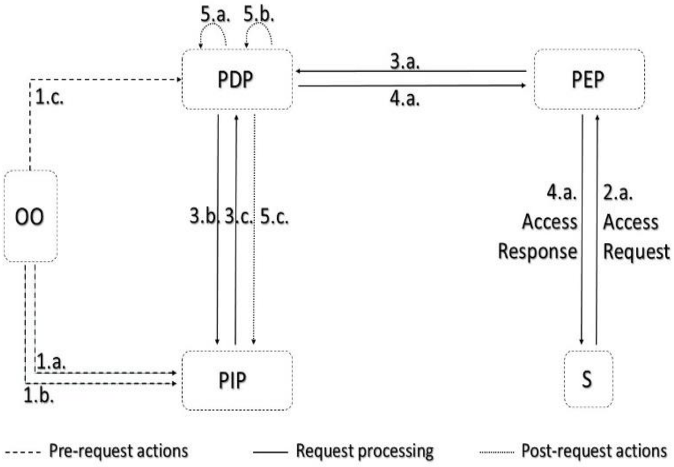


Fig. 4. Overall Functioning of the Framework.

TABLE. II. SUMMARY OF THE FRAMEWORK'S PHASES

	Action	Responsible	Location/ destination	Example
Phase 1	**1.a.** AC policy definition	OO	PIP	$\mathrm{permission(org, r, v, ay, C, p)}$
	1.b. AC relationship definition (either explicit or through attributes)	OO	PIP	$\mathrm{empower(org}, s^1, \dots, s^m, \mathrm{r)}$ $\mathrm{use(org}, o^1, \dots, o^n, \mathrm{v)}$ $\mathrm{consider(org}, a^1, \dots, a^p, \mathrm{ay)}$
	1.c. Threshold definition	OO	PDP	$tr = 1;$ $tr = 0.7$
Phase 2	**2.a.** The subject request executing an action over a resource/object	The requester/ subject	PEP	$\mathrm{get\_access(org, s, o, a, c)}$
Phase 3	**3.a.** Request for decision	PEP	PDP	$\mathrm{get\_access(org, s, o, a, c)}$
	3.b. Request matching information	PDP	PIP	$\mathrm{get\_match(org, s, o, a, c)}$
	3.c. Matching response	PIP	PDP	$\mathrm{back\_match(org, r, v, ay, C, p)}$
Phase 4	**4.a.** Make decision	PDP	PEP	$\mathrm{grant\_access(org, s, o, a, c)}$ $\mathrm{deny\_access(org, s, o, a, c)}$
Phase 5	**5.a.** Update learning matrix	PDP	PDP	$\mathrm{l\_matrix.append([org, r, v, ay, C, p])}$
	5.b. Learning process	PDP	PDP	$\mathrm{model.fit(l\_matrix)}$
	5.c. Send updates periodically	PDP	PIP	$\mathrm{permission(org, r, v, ay, C, p)}$

V. Case Study

The IoT is growing by leaps and bounds, thus creating this large, smart and autonomous system requiring less and less human intervention. Smart city (SC) was always the case in point that portrays this vision where devices not only interact but depend on (even control) each other. This section presents a SC situation to depict how the previously described framework could be implemented in such complex environment.

Actually, the choice of SC has been motivated by many reasons: First, because it covers many of the other IoT use cases. Secondly, in addition to the IoT world, it is also a typical case in the AI domain. Thirdly, it presents an active research field [47], and last but not least it is becoming a necessity in order to efficiently deal with the massive growth of urbanization that is estimated to reach 66% by 2050 [48].

To better illustrate this example, let us take three organizations in a SC, namely: a car rental agency (CRA), which represents the organization we are willing to secure, a smart parking (SP) and a police station (PoS).

As a CRA, our organization is mainly composed of self-driving cars (decomposed into two views: luxury and normal cars); its customers are generally normal clients, VIP clients or blacklisted ones (which makes respectively three roles: NC, VIP, BC); regarding the activities it is more realistic to categorize them by rental period (a1: 1 day, a2: between 1 and 3 days, and a3: more than 3 days). Finally, the context represents the time of the year during which the request is made: is it a peak season (peak) like summer for example, or off season (off).

In this case, each self-driving car is equipped with its own PEP, which means it is the car itself that receives the access request and responds the requester by the final decision. PIP may ideally be one (or several) local server(s) storing the AC policy. Regarding the PDP, in general it needs to be spread out over two layers: one part dealing with the steps (1.c., 3, 4 and 5.a.) that are executed within the smart car, and another part that needs to be run on C3 category (namely 5.b. and 5.c.). Nevertheless, as we will see later some lightweight versions of these two steps could also be run within the smart car. The following Fig. 5 portrays the building blocks forming this platform:

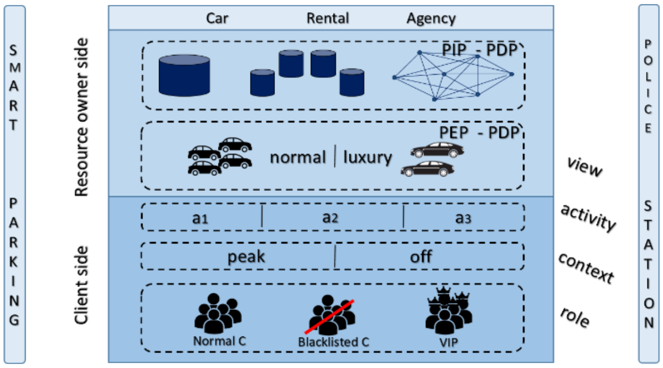


Fig. 5. The Building Blocks of the Case Study.

Now that we have exposed all the stakeholders, first thing to do as the owner of CRA is to define a primary AC policy, so in the PIP there will be two sorts of rules: (i) Those of the 1.a. step, which will be in the form of:

$$permission(*, VIP, luxury, a3, peak, 1)$$

$$prohibition(Competitor\_CRA, BC, normal, a1, off, 1)$$

* Star stands for *all*, i.e. any organization.

Then the definition of the AC relationships, which are responsible for matching abstract to concrete entities, for example a car is considered luxury if its release date comes after 2017 and its price exceeds 100,000$:

$$consider(*, time \geqslant 3days, a3)$$

$$empower(*, badge = True, VIP)$$

$$use(CRA, r\_date \geqslant 2017, price \geqslant 100{,}000\$, luxury)$$

For luxury cars the threshold is set to 1, so that access is not granted to any role without a 100% confidence about its previous experiences. While for the less critical view (*normal*) the threshold is reduced to 0.8 allowing more usability. These thresholds are set in the car side's PDP.

Let's move to phase 2. At the moment, an access request arrives to the PEP of the car id=79, it is a subject presenting his membership badge and demanding this resource for 5 days starting from August the 1st, 2019.

$$\begin{aligned} get\_access(org_A, badge = True, id = 79, \\ time = 5days, \\ starting\_time = 01/08/2019) \end{aligned}$$

After the previous request is forwarded to the PDP, this latter requests matching information from the PIP via:

$$\begin{aligned} get\_match(org_A, badge = True, id = 79, \\ time = 5days, \\ starting\_time = 01/08/2019) \end{aligned}$$

Then it gets the response:

$$permission(org_A, VIP, luxury, a3, peak, 1)$$

And since the threshold for luxury cars is set to 1, the PDP makes and informs the PEP about its final decision allowing the requester to access/use the car:

$$\begin{aligned} grant\_access(org_A, badge = True, id = 79, \\ time = 5days, \\ starting\_time = 01/08/2019) \end{aligned}$$

Now comes the learning phase. Actually, it is up to the organization either to rate the experience after its end (i.e. rate just the physical state of the car for example) or to perform an online rating so that it gets periodic feedback from the smart car to compute more sophisticated inferences (e.g. geolocation, fuel consumption, speed) and then the PDP ends up with a weighted average. To simplify our case, we use the first option.

Therefore, after the car is back, the PDP receives the required ingredients to compute the feedback. Let us consider that what is important for this CRA is whether the car is back in time and its mechanical situation, and that for both features this experience was negative, so the feedback was set to 0.6. The learning matrix is now updated:

$l\_matrix.append([org_A, VIP, luxury, a3, peak, 0.6])$

Now we can imagine that if after a while similar feedbacks comes from the same organization ($\mathrm{org_A}$) or under the same context or whatever, the algorithm will detect the pattern when the learning model is run over the collected data using $model.fit(l\_matrix)$, then the AC policy could eventually be updated, thus improved over time.

A final point that we want to highlight is about the attributes used in the matching phase. In fact, depending on their nature there are two ways to collect them; either explicitly as we saw in the previous example (i.e. as parameters within the access request) or by extracting them directly from the object. For instance, if we want to use the license plate number we could eventually use the cameras from the smart parking.

VI. Discussion and Conclusions

The motivation behind this work was to come up with a smart, decentralized and IoT-suited AC framework. In this section we discuss theoretical and practical contributions of this paper as well as their strengths over existing solutions.

First thing to remember is that IoT are not all constrained. To prove this we went back to define this paradigm and to delimit its boundaries and layers; in addition to overview several IoT propositions that have used ML techniques. However, to the best of our knowledge, the existing solutions are narrowed ones, each one focus either on proposing a model, managing the policy or tackling specific techniques. The problem with this approach shows up when an organization wants to put it all together, generally the concatenation of these uncoordinated solutions do not give acceptable results. Thus proposing a holistic framework to manage AC in IoT environments is another key contribution of this paper.

Equally important, the introduction of the notion of organization in IoT is notably benefic since it helps drastically decreasing the problem of role explosion, which is the number one problem challenging role-based and attribute-based AC solutions. In fact, an IoT environment could always be broken off into several organizations and therefore, depending on their mission, the roles will be manageable; and of course what goes for roles also goes for views, activities and context. Note that the organization aspect has by no means been a source of centralization, it is rather a push toward more decentralization and collaboration.

Another interesting and captivating point: the learning process we have introduced actually differs from the traditional procedure commonly used in the current AI applications, which consist first of a learning phase followed by the prediction phase. In our framework it is rather a minimum of basic knowledge is initiated in the beginning then learning came to fine-tune this expertise. It is actually how humans learn, they always have some innate skills before anyone comes to teach them anything.

Finally, we believe that having the ability to personalize the threshold, not forcing an immediate update after each experience, and allowing explicit as well as extracted attributes bring this framework with further flexibility and adaptability to fit the IoT requirements discussed in SectionIII-A.

References

[1] A. A. El Kalam, A. Outchakoucht, H. Es-samaali, “Emergence-Based Access Control: New Approach to Secure the Internet of Things”. DTUC ’18 Paris. 2018.

[2] T. M. Mitchell, “Machine learning”, 7th ed. NY, McGraw Hill, ISBN: 0070428077. 1997.

[3] Y. LeCun, Y. Bengio, G. Hinton, “Deep Learning”, Nature, volume 521, 2015.

[4] D. Gope, G. Dasika and M. Mattina, “Ternary Hybrid Neural-Tree Networks for Highly Constrained IoT Applications,” in the 2nd Conference on Systems and Machine Learning (SysML), 2019.

[5] A. Kumar, S. Goyal, M. Varma, “Resource-efficient Machine Learning in 2 KB RAM for the Internet of Things”; Proceedings of the 34th International Conference on Machine Learning, 2017.

[6] A. Gordon et al. “MorphNet: Fast & Simple Resource-Constrained Structure Learning of Deep Networks,” arXiv:1711.06798. 2017.

[7] X. Zhang, X. Zhou, M. Lin and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6848-6856, 2018.

[8] N. Kaminski et al., "A neural-network-based realization of in-network computation for the Internet of Things," IEEE International Conference on Communications (ICC), pp. 1-6, 2017.

[9] H. Khelifi et al., "Bringing Deep Learning at the Edge of Information-Centric Internet of Things," in IEEE Communications Letters, vol. 23, no. 1, pp. 52-55, 2019.

[10] N. D. Lane et al., "An Early Resource Characterization of Deep Learning on Wearables, Smartphones and Internet-of-Things Devices," in the Proceedings of the 2015 International Workshop on Internet of Things towards Applications, pp. 7-12, 2015.

[11] Nvidia. [online] Available at: https://blogs.nvidia.com/blog/2015/06/09/gpu-powered-june-oven/, [Accessed 2 December 2019].

[12] M. Mohammadi, A. Al-Fuqaha, M. Guizani and J. Oh, "Semisupervised Deep Reinforcement Learning in Support of IoT and Smart City Services," in IEEE Internet of Things Journal, vol. 5, no. 2, pp. 624-635, 2018.

[13] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, 2015.

[14] H.-S. Ham, H.-H. Kim, M.-S. Kim, and M.-J. Choi, "Linear SVM-based android malware detection for reliable IoT services," Journal of Applied Mathematics, vol. 2014, 2014.

[15] A. Heuser and M. Zohner, "Intelligent machine homicide," in International Workshop on Constructive Side-Channel Analysis and Secure Design, pp. 249-264, 2012.

[16] L. Lerman, G. Bontempi, and O. Markowitch, "A machine learning approach against a masked AES," Journal of Cryptographic Engineering, vol. 5, no. 2, pp. 123-139, 2015.

[17] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001

[18] Y. Meidan et al., "Detection of Unauthorized IoT Devices Using Machine Learning Techniques," arXiv preprint arXiv:1709.04647, 2017.

[19] R. Doshi, N. Apthorpe, and N. Feamster, "Machine Learning DDoS Detection for Consumer Internet of Things Devices," arXiv preprint arXiv:1804.04159, 2018.

[20] Q. Li, K. Zhang, M. Cheffena, and X. Shen, "Channel-based Sybil Detection in Industrial Wireless Sensor Networks: a Multi-kernel Approach," in IEEE Global Communications Conference, pp. 1-6: IEEE, 2017.

[21] M. Al-garadi, A. Mohamed, A. Al-ali, et al. “A survey of machine and deep learning methods for internet of things (IoT) security”. arXiv preprint arXiv:1807.11023, 2018.

[22] M. Ford, “Architects of Intelligence: The truth about AI from the people building it”, book, Packt Publishing, 2018.

[23] E. De Coninck et al., "Distributed neural networks for Internet of Things: the Big-Little approach," in International Internet of Things Summit, pp. 484-492, 2015.

[24] N. McLaughlin et al., "Deep android malware detection," in Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy, pp. 301-308, 2017.

[25] P. Torres, C. Catania, S. Garcia, and C. G. Garino, "An analysis of Recurrent Neural Networks for Botnet detection behavior," in Biennial Congress of Argentina, IEEE pp. 1-6, 2016.

[26] I. Goodfellow et al., "Generative adversarial nets," in Advances in neural information processing systems, pp. 2672-2680, 2014.

[27] R. E. Hiromoto, M. Haney, and A. Vakanski, "A secure architecture for IoT with supply chain risk management," in 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 1, pp. 431-435, 2017.

[28] M. A. Aref, S. K. Jayaweera, and S. Machuzak, "Multi-agent Reinforcement Learning Based Cognitive Anti-jamming," in Wireless Communications and Networking Conference (WCNC), IEEE pp. 1-6, 2017.

[29] A Outchakoucht, ES Hamza, JP Leroy, “Dynamic access control policy based on blockchain and machine learning for the internet of things,” in International Journal of Advanced Computer Science and Applications, Vol. 8, No.7, 2017.

[30] NVIDIA Corporation, Jetson Nano Developer Kit, March 2019, [online] Available at: https://developer.nvidia.com/embedded/jetson-nano-developer-kit [accessed 2 December 2019]

[31] Google Brain, Deploy machine learning models on mobile and IoT devices, 2019, [online] Available at: https://www.tensorflow.org/lite [accessed 2 December 2019].

[32] C. Zhang, P. Patras, and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," arXiv preprint arXiv:1803.04311, 2018.

[33] R.S. Sandhu, “Role-based access control,” Adv. Comput. 46, pp. 237–286, 1998.

[34] E. Yuan, J. Tong, “Attributed based access control (ABAC) for Web services,” in: IEEE Int. Conf. Web Serv., IEEE, 2005.

[35] A. Ouaddah, H. Mousannif, A. A. Elkalam, A. Ait Ouahman,

[36] “Access control in the Internet of Things: Big challenges and new opportunities”, Computer Networks 112, pp. 237–262, 2017.

[37] A. Kalam, et al., “Organization based access control,” in: IEEE 4th Int. Work. Policies Distrib. Syst. Networks, IEEE Comput. Soc, pp. 120–131, 2003.

[38] S. El Bouanani, M. A. El Kiram, O. Achbarou and A. Outchakoucht, "Pervasive-Based Access Control Model for IoT Environments," in IEEE Access, vol. 7, pp. 54575-54585, 2019. doi: 10.1109/ACCESS.2019.2912975.

[39] M.R. Abdmeziem, D. Tandjaoui, I. Romdhani, “Architecting the Internet of Things: State of the Art”, Robots and Sensor Clouds. Studies in Systems, Decision and Control, vol 36. Springer, 2015.

[40] C. Bormann, M. Ersue, A. Keranen, "Terminology for Constrained-Node Networks", RFC 7228, May 2014.

[41] A. Ouaddah, A. Abou Elkalam and A. Ait Ouahman, “FairAccess: a new Blockchain-based access control framework for the Internet of Things”, Security and Communication Networks, pp. 1-22, 2017.

[42] ISO/IEC JTC 1, Information technology, iso/iec 29146:2016, A framework for access management, 2016.

[43] eXtensible Access Control Markup Language (XACML) Version 3.0, OASIS Standard, January 2013.

[44] A. Ameziane El Hassani et al., “Integrity-OrBAC: a new model to preserve Critical Infrastructures integrity”, International Journal of Information Security, Springer, vol. 14, Issue 4, pp 367–385, 2014.

[45] A.E. Kalam, Y. Deswarte, “Multi-Orbac: a new access control model for distributed, heterogeneous and collaborative systems,” in: 8th IEEE International Symposium on Systems and Information Security, p. 1, 2006.

[46] A. Abou El Kalam, Y. Deswarte, A. Baïna, M. Kaâniche, “PolyOrBAC: a security framework for critical infrastructures,” Int. J. Crit. Infrastruct. Prot. pp. 154–169, 2009.

[47] I. Bouij-Pasquier, A. A. El Kalam, A. A. Ouahman, and M. De Montfort, “A Security Framework for Internet of Things,” Springer International Publishing, pp. 19–31, 2015.

[48] M. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, P. Sheth, “Machine learning for Internet of Things data analysis: A survey”. Digital Communications and Networks. 2017.

[49] U. Nations, “World urbanization prospects: The 2014 revision, highlights,” Department of economic and social affairs. Population Division, United Nations, 2014.

Enhancing the Bitrate and Power Spectral Density of PPM TH-IR UWB Signals using a Sub-Slot Technique

Bashar Al-haj Moh'd[1]
Department of Medical Engineering
Al-Ahliyya Amman University
Amman, Jordan

Nidal Qasem[2]
Department of Electronics and Communications
Engineering, Al-Ahliyya Amman University
Amman, Jordan

***Abstract*—Increasing the receiver's bitrate and suppressing the spectral line are issues of major interest in the design of compliant Time-Hopping Impulse Radio (TH-IR) Ultra-Wide Band (UWB) systems. Suppression of spectral lines has been commonly addressed by randomizing the position of each pulse to make the period as large as possible. Our analysis suggests that this influences the overall shape of a signal's Power Spectral Density (PSD) in a way that is useful for spectral line suppression or diminishing the PSD maximum peak power. A method for utilizing the system to generate a Dynamic-Location Pulse-Position Modulated (DLPPM) signal for transmission across a UWB communications channel is presented, and an analytical derivation of the PSD of a proposed DLPPM signal TH-IR UWB is introduced. Our proposed method can be applied without affecting the users of other concurrent applications. The theoretical model for DPLM TH-IR is compared with the PSD for conventional DPLM TH-IR. The results show that spectral estimation methods based on Fast Fourier Transform (FFT) significantly overestimate the continuous part of the PSD for small and medium signal lengths, which has implications for assessing interference margins by means of simulation. Another purpose of this paper is to improve a predesigned system by increasing the receiver's bitrate. This will be achieved by using the bits that control the sub-slot technique as information and designing a receiver capable of detecting them. The bitrate is effectively doubled. Finally, the proposed system for DPLM TH-IR has been built inside Simulink/MATLAB to test its results via a conventional DPLM TH-IR system.**

Keywords—Bitrate; FFT; PPM; PSD; spectral estimation; sub-slot; TH-IR; UWB

I. Introduction

Ultra-Wide Band (UWB) systems based on Impulse Radio (IR) are the leading candidates for communication systems with low power, low complexity, low rate, and high battery life, as well as immunity to multipath interference characteristics. Applications involving such systems range from collision avoidance automotive systems to sensor networks. IR-UWB technology (also referred to as impulse, baseband, and zero-carrier technology), uses very short pulses, which imply a large signal bandwidth, to convey information [1]. Various modulation techniques, such as Pulse-Amplitude Modulation (PAM), Pulse-Interval Modulation (PIM), Pulse-Shape Modulation (PSM), Pulse-Position Modulation (PPM), On–Off Keying (OOK), and Bi-Phase Shift Keying (BPSK), are used to transmit the information in such systems [2, 3]. In PPM, the information is determined by the position of one pulse [4].

In order to be able to deploy such applications, the interference from UWB-based devices to already-established narrowband deployments must be kept to satisfactory levels. Consequently, the Power Spectral Density (PSD) of IR-UWB-based devices must comply with regulatory spectral masks such as the one used by the Federal Communications Commission (FCC) [5]. In this context, simulation of the signals produced by UWB devices with their corresponding PSD estimation by Fast Fourier Transform (FFT) methods is an invaluable tool for the evaluation and improvement of such systems before building the physical prototype.

Conventional PPM TH-IR typically assumes a fixed timing offset between pulses in the signal set. This has several drawbacks for UWB systems. Therefore, there is a need for modulation schemes to realize the benefits of standard PPM while providing greater randomness.

Thus, a novel PPM scheme must be more robust to fix timing offset effects while allowing for a greater throughput in UWB systems than conventional PPM TH-IR schemes. Therefore, methods and systems for generating Dynamic-Location Pulse-Position Modulation (DLPPM) have been introduced [6].

In previous research [1,7,8], the behaviour of simulations with FFT-based PSD estimation of UWB signals was analysed by comparing the results with analytical and actual measurements. This paper is an extension of such works, whereby a comparison between the proposed and existing systems is introduced.

The use of FFT-periodogram estimation methods for spectrum analysis of random signals is well studied in the literature [9]. However, some constraints must be observed when using these methods for the purpose of assessing, via simulation, the PSD behaviour of a particular UWB system before implementation. In this work, the behaviour of such estimation methods for proposed DLPPM TH-IR as a function of the sample length is analysed by comparing with previous results of conventional DLPPM TH-IR and PPM TH-IR [1,6] obtained with a swept spectrum analyser. This allows the identification of several issues that must be considered to enhance the PSD of conventional PPM TH-IR UWB signals. The proposed DLPPM TH-IR utilised the subsections to double the bitrate in comparison with conventional DLPPM TH-IR, but the PSD is not affected.

This paper is organised as follows: Section 2 describes the basic system properties and derivation of the conventional and proposed DLPPM TH-IR UWB schemes. Section 3 shows analytical and simulation results of the described system performance, and the conclusions are given in Section 4.

II. Signal Representation and Analytical PSD of a DLPPM TH-IR System

A. Conventional DLPPM TH-IR System

Before discussing issues related to estimation of the spectrum power and increasing the bitrate for the proposed DLPPM TH-IR, we introduce the signal considered for analysing the proposed DLPPM TH-IR UWB system. The block diagram shown in Fig. 1 consists of six stages and manages to diminish the problem of the spectral lines and obtain a smooth PSD by observing the output at the spectrum analyzer [6].

The signal generated by the conventional DLPPM TH-IR system, shown in Fig. 1, is described by the formula [6]:

$$x(t) = \sum_m w\left(t - mT_r - \left(\beta_m + \frac{\alpha_m}{N_{ss}}\right)T_B\right) \tag{1}$$

where $w(t)$ accounts for the pulse shape, T_r is the mean pulse repetition rate, β_m is the m^{th}symbol from the rate of the 1⁄3 encoder taking values on the set $\{0,1,...,7\}$, T_B is the PPM modulation shift (modulation index), N_{ss} is the number of time sub-slots, and α_m is the position of the pulse inside the pulse repletion rate taking values on the set {0, 1, ..., N_{ss}-1}. Note that $x(t)$ is a random process.

B. Proposed DLPPM TH-IR System

Fig. 2 shows the proposed system, allowing comparisons between the conventional and proposed systems based on the DLPPM TH-IR technique. The introduced system differs from the one shown in Fig. 1 by utilising the extra sub-slot stage, which divides each slot by the number of sub-slots for holding new information, leading to an increased bitrate.

Fig. 1 shows the previous system block diagram. We can see that the data coming from the source through the convolutional encoder are used to choose the pulse-position slot of the PPM system. The convolutional encoder's role is to randomize the pulse position for a smooth PSD. The sub-slot position block is used to randomly determine the position of the pulse inside the slot. In other words, this block chooses the sub-slot position for maximum randomness of the positions for consecutive pulses.

In this paper, the position of the sub-slot is used to represent data in order to increase the bitrate and channel capacity. The block diagram of the proposed system is shown in Fig. 2. In this system, the data will be split into two parts each part contains three bits of data: the first part is to go through convolutional encoder-1 to select the proper slot in the frame, while the second part is to select the proper sub-slot inside the selected slot through convolutional encoder-2. This idea will be reflected in (1), where α_m on the old system was randomly chosen, and the position means nothing but just helps to randomize the pulse position to obtain a smooth spectrum. Whereas in the proposed system, this value has meaning, representing three bits of data as well as keeping the spectrum smooth.

In general, the PSD of the random TH-IR signal $\overline{X}(f)$ consists of continuous $x(t)$ as well as discrete components [1, 10] and is given by:

$$\overline{X}(f) = \frac{1}{T_r}|W(f)|^2 - \frac{1}{T_r}|W(f)|^2\left(\frac{1}{N_S*N_{SS}} + \frac{2}{(N_S*N_{SS})^2}\sum_{n=1}^{(N_S*N_{SS})-1}\left((N_s*N_{ss} - |n|)\cos(2\pi fnT_c)\right)\right) + \frac{1}{T_r^2}\left(\frac{1}{N_S*N_{SS}} + \frac{2}{(N_S*N_{SS})^2}\sum_{n=1}^{(N_S*N_{SS})-1}\left((N_s*N_{ss} - |n|)\cos(2\pi fnT_c)\right)\right)\sum_{-\infty}^{\infty}\delta\left(f - \frac{k}{T_r}\right) \tag{2}$$

where $W(f)$ is the Fourier transform of $w(t)$ and N_s is the number of time slots. The value of T_B is chosen to eliminate as many spectral lines as possible [11]. The proposed system is a new modulation technique based on conventional DLPPM TH-IR [6]. The proposed DLPPM TH-IR scheme maximizes the average separation between modulated pulses to achieve greater resistance to large delay spreads. In addition, the proposed DLPPM TH-IR randomizes the time offset between adjacent pulses to provide greater immunity to multiple access interference. Thus, the bandwidth efficiency of UWB communications systems is increased.

Fig. 3 is an illustration of an exemplary signal, in the time domain, modulated using the DLPPM TH-IR technique. The signal consists of a number of symbols, having a symbol period T_r. Symbols are transmitted with very short pulses. Each symbol represents M bits of binary information and has a value in the range of 0 to 2^M-1. Each symbol is divided into a number of slots, each having a duration equal to T_B. Each slot contains sub-slots, each having a duration equal to T_B/N_{SS}. For instance, if N_S =8 and N_{SS} =8, there are 8 slots each containing 8 sub-slots, which correspond to the symbol values {0, 1, 2, …, 7}. Finally, each slot is divided into ${T_B}/{8}$ time sub-slots, which correspond to the conventional DLPPM TH-IR code sequence.

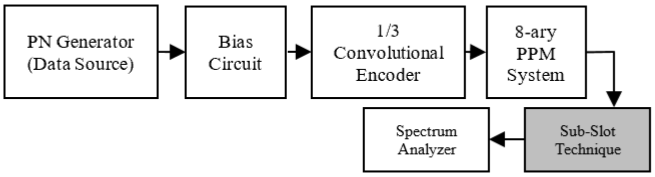


Fig. 1. Conventional DLPPM TH-IR System Block Diagram [6].

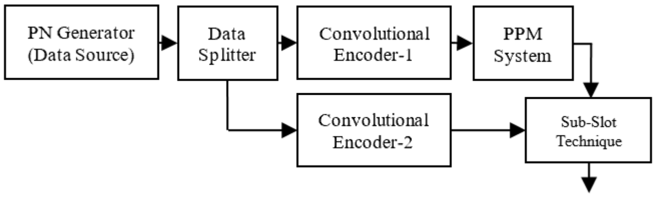


Fig. 2. Proposed System Block Diagram using DLPPM TH-IR Technique.

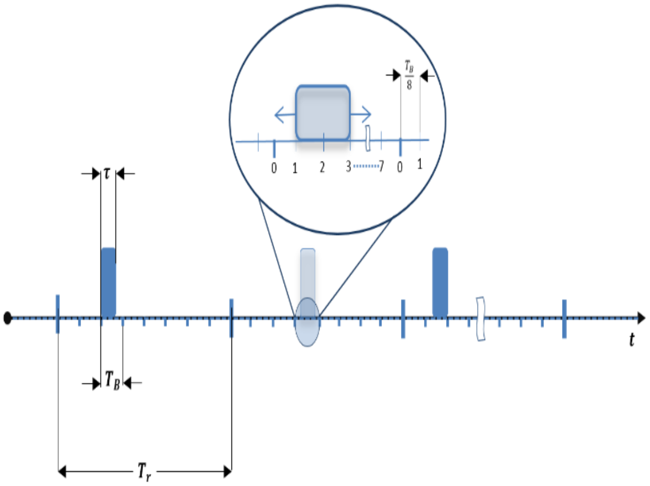


Fig. 3. An Illustrative Example of the Proposed DLPPM TH-IR Signal, where N_S=8 and N_{ss}=8.

Comparison between the conventional and proposed systems is described again in the signal shown in Fig. 3. This signal represents three bits of data when used in the old system, which exactly represents the data “011” in this example, as the pulse is in the third slot, where one frame of T_r has eight slots. As a result, moving this pulse inside the slot has one main function, which is to achieve maximum randomness and smooth PSD. In the proposed system, the signal can be used to represent six bits of data with no effect on the PSD smoothness. For instance, this signal represents the data “011001”, where the three bits on the left “011” represent the slot position exactly as they do in the old system, and the other three bits on the right “001” represent the sub-slot position, with this position determined from the (1/3) convolutional encoder-2.

III. Analytical and Simulation Results

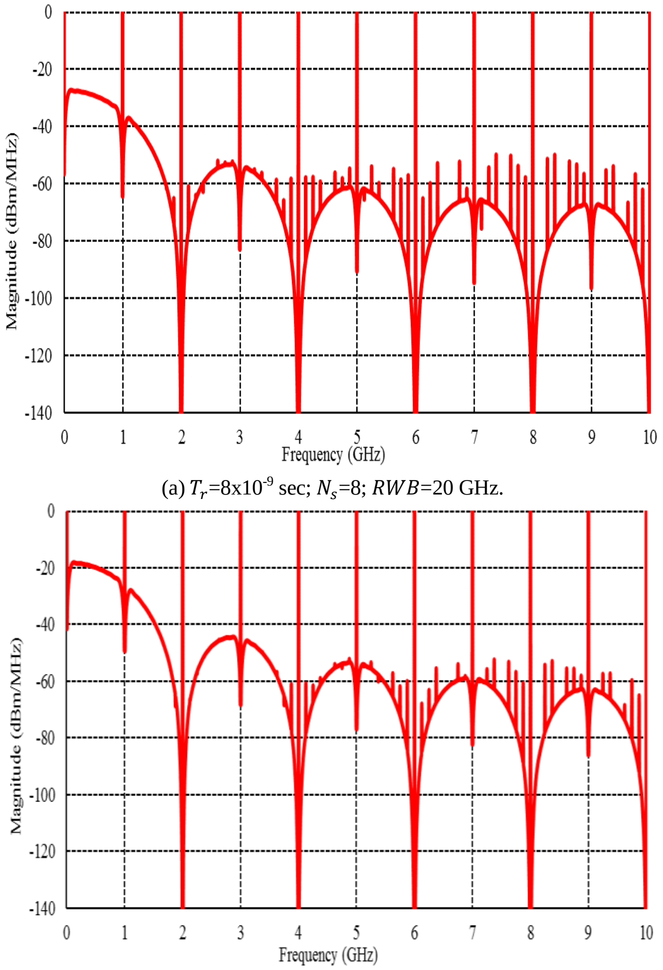
A. Analytical Results

In this section, we compare a conventional DLPPM TH-IR with the new proposed DLPPM TH-IR, where T_r=8×10^{-9} sec, N_s=8, and RWB=20, with 40 GHz applied and tested for both cases. Different levels of N_s have been selected for the DLPPM TH-IR system to detect the effects on PSD.

Fig. 4 shows the theoretical PSD obtained by using the conventional PPM TH-IR system described in [6] for Resolution Bandwidth (RBW) equal to 20 and 40 GHz. It is evident that the tested signal had both random-like and harmonic-like components. Fig. 4 shows that the PSD section consists of 10 spectral lines. These spectral lines do not comply with regulatory spectral masks such as the one used by the FCC. When the RBW increases from 20 to 40 GHz, as shown in Fig. 4(b), it is clearly seen that the displayed power of the spectral lines does not change. However, the level of the continuous-like component increases by about 8 dBm/MHz.

Fig. 5 shows theoretical PSDs obtained by using the conventional and proposed DLPPM TH-IR systems described in Section 2 and depicted in Fig. 1 and 2. In Fig. 5(a), the DLPPM TH-IR system has 8 slots (N_s), and each slot has been divided into 4 sub-slots (N_{ss}). Fig. 5(a) consists of spectral lines repeated every 4 GHz, with the total number of spectral lines reduced and smooth harmonics components achieved compared to the conventional PPM TH-IR system, as shown in Fig. 4(a).

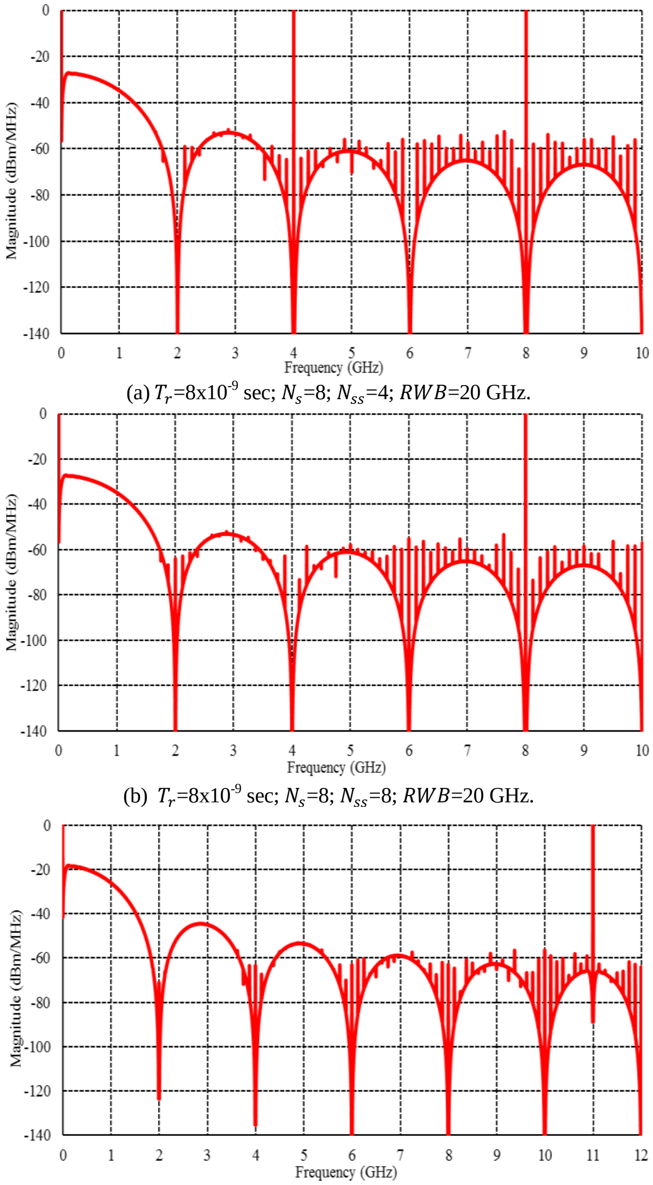
Fig. 5(b) shows the PSD section consisting of spectral lines repeated every 8 GHz. Fig. 5(c) shows the PSD section consisting of spectral lines repeated every 11 GHz, with the total number of spectral lines reduced and smoother harmonics components achieved. In the last case, Fig. 5(c), the spectral line is outside the UWB legalizing spectrum across 7.5 GHz, between 3.1 and 10.6 GHz [12], which provides an interference-free system. Identical responses have been achieved for both systems, as expected.



(a) T_r=8x10^{-9} sec; N_s=8; RWB=20 GHz.

(b) T_r=8x10^{-9} sec; N_s=8; RWB=40 GHz.

Fig. 4. Estimated PSDs Performed with different Resolution Bandwidths (RBWs) from [6].



(a) T_r=8x10^{-9} sec; N_s=8; N_{ss}=4; RWB=20 GHz.

(b) T_r=8x10^{-9} sec; N_s=8; N_{ss}=8; RWB=20 GHz.

(c) T_r=8x10^{-9} sec; N_s=8; N_{ss}=11; RWB=40 GHz.

Fig. 5. Estimated PSDs Obtained by an Analytical Model with different RBWs and N_{ss} (for Both Systems).

B. Simulation Model Results

The conventional and proposed DLPPM TH-IR systems**,** shown in Fig. 1 and 2, respectively, have been built inside Simulink/MATLAB in order to compare the two systems for bitrate, as well as the selected code for eliminating the spectral lines or increasing smoothness to reduce the interference problem on other users. In the simulation system, each frame has eight slots and each slot has eight sub-slots. Both systems have same bandwidth but the one depicted in Fig. 1 represents three bits and the one depicted in Fig. 2 represents six bits. As a result, a doubled bitrate is expected to be achieved by this technique (proposed system).

1) PSD: In this section, a [4 2 1] code set of systematic convolutional codes with rate 1/3 has been applied and tested with equal probability [13–15]. The same process has been repeated to detect the difference between the conventional and proposed DLPPM TH-IR systems.

The results for both systems are compared for the same metrics: Pseudo Noise (PN) length, convolutional encoder codes, symbol rate, pulse shape (0.5 ns), and with equal probability. In the DLPPM TH-IR system, each slot (N_s) is divided into 8 sub-slots (N_{ss}). The pulse position in both systems will be determined in the same way, with the only difference being the location of the pulse inside the slot. In both (conventional/proposed) DLPPM TH-IR systems, a convolutional encoder is used to decide the sub-slot pulse position. In the old system, the sub-slot position is randomly chosen by the (Sub-slot Technique) block, as shown in Fig. 1. In the new system, part of the data came from the source is used to choose the sub-slot position to represent three bits of data.

Fig. 6 shows an excellent PSD performance obtained by using conventional vs proposed DLPPM TH-IR systems with the same conditions. Excellent performance returns to the randomness between consecutive pulses. Furthermore, it can reduce the spectral lines or make them smoother. Thus, the proposed system will not affect the contribution achieved in [6], but it will enhance the bitrate as demonstrated in the next sub-section.

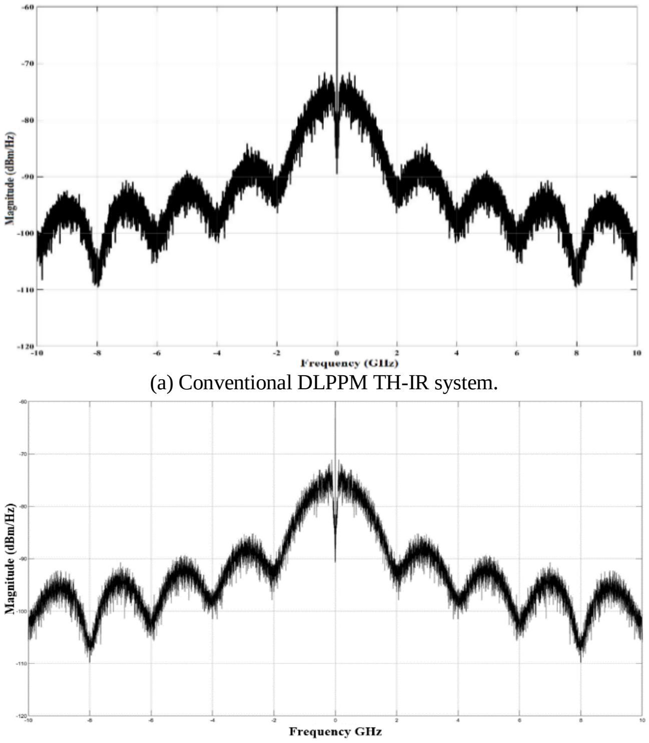
2) Bitrate: Fig. 7 shows the receiver side used to detect the data for both systems. In the old system receiver, as shown in Fig. 7(a), the (Slot Location Detector) block can detect which slot has a pulse regardless of which sub-slot is in. Then the signal will go to the PPM demodulator to convert these symbols to three bits encoded data. After that, Viterbi decoder is used for decoding and recovering the original data. In the new system receiver, as shown in Fig. 7(b), the (Slot/Sub-slot Location Detector) block is used to separate the information bits into two parts: the first part is comprising the bits that control the location of the slot and the second part is comprising the bits that specify the sub-slot. The slot location will be processed by the upper side of the receiver exactly as the old system, and the sub-slot by the lower side. The data from both sides are combined by (Data Combiner) block. The overall result is six bits of data received per one PPM frame instead of three bits per frame in the old system.

The easiest way to demodulate a PPM signal is to use a decoder. This is one of the best techniques for digital communications when computational complexity dominates in importance. It permits major equipment simplification while obtaining the full performance benefits of maximum likelihood decoding. The decoder structure is relatively simple for a short constraint length N, making decoding feasible at relatively high rates of up to 10 Gbit/s [16].

Simulation has been run for 20 μs for each system, where the data source (PN) sampling rate is 4 ns for the old system and 2 ns for the new one. The channel assumed to be noiseless. The number of bits received in the old system is 4801, as shown in Fig. 8(a) whereas the proposed system has received 9801 as shown in Fig. 8(b). The number of delayed bits is 200. By adding 200 to each system, the total bits received for the old and the proposed systems are 5001 and 10001, respectively, where both systems have a bandwidth of 2 GHz. Obviously, we can see that the bitrate has been duplicated exactly. The noiseless channel capacity is shown in (3) [17]:

$$Maximum\ Bitrate\ =\ 2 * BW * log_2(L) \qquad (3)$$

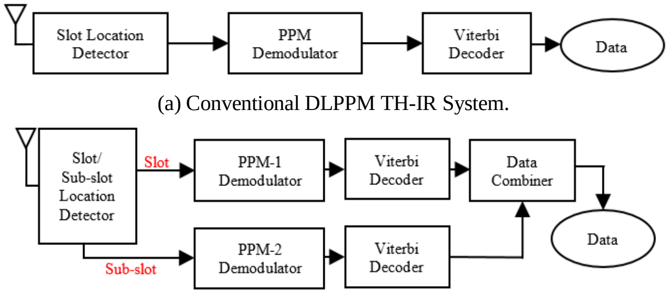
where BW is the bandwidth and L is the number of levels, which is eight for the conventional system and 64 for the proposed system.



(a) Conventional DLPPM TH-IR system.

(b) Proposed DLPPM TH-IR system.

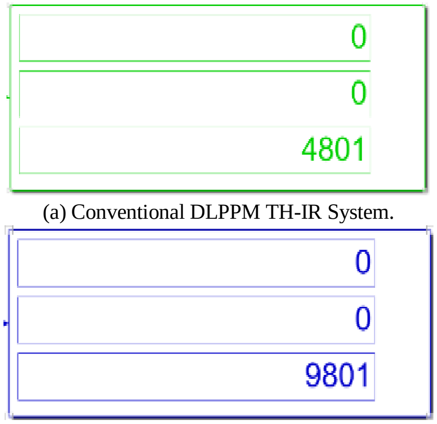
Fig. 6. PSD with Equal Probability for a Systematic Code [4 2 1].



(a) Conventional DLPPM TH-IR System.

(b) Proposed DLPPM TH-IR System.

Fig. 7. Block Diagram of the Receiver for both Systems.



0

0

4801

(a) Conventional DLPPM TH-IR System.

0

0

9801

(b) Proposed DLPPM TH-IR System.

Fig. 8. Bitrate Counter based on Receiver Built Inside Simulink/MATLAB.

IV. Conclusion

The present method provides a novel technique which meets the requirements described earlier with the help of a DLPPM TH-IR scheme that allows for a greater throughput than the conventional DLPPM TH-IR scheme. This can be achieved by maximizing the pulse separation and randomizing the time offset between pulses in a time-efficient manner to ensure the period is maximized.

A mathematical representation of the PSD of a DLPPM TH-IR UWB signal was derived. The analytical result was used to investigate the effect of the variable position of pulses in the DLPPM TH-IR system in the PSD of the signal, and it was found that it can be effectively used to eliminate some spectral lines or to diminish the peak value of the PSD. It has been observed that DLPPM TH-IR significantly outperforms conventional PPM TH-IR with respect to spectral efficiency when the location of a pulse is variable within each slot. The hardware complexity at the receiver side does not need to be increased, which makes the DLPPM scheme very attractive for TH IR-UWB communication systems.

Theoretically, the bitrate can be increased without limit by increasing the number of sub-slots. As a result, the Inter Symbol Interference (ISI) will be increased. This can be clearly noted when the phase error is added to the old and proposed systems. The old system is less affected than the new one. This can be explained as follows: in the old system, any change in the sub-slot position of the pulse will not affect the symbol data for one frame of the PPM signal. It is sufficient to detect the pulse inside the same slot to avoid ISI. In contrast, in the proposed system, a small phase can lead to an error of three bits of six bits for one frame of data.

Finally, the simulated results within Simulink/MATLAB were compared to the conventional DLPPM TH-IR UWB. The results showed that when testing with a systematic code [4 2 1], a smoother PSD without spectral lines and a doubled bitrate were achieved.

REFERENCES

[1] Villarreal-Reyes, R. M. Edwards and B. Al-haj Moh'd, "On for the Comparison of Measurement and Simulation of the Power Spectral Density of PPM TH-IR UWB Signals," Loughborough Antennas and Propagation Conference Proceedings, LAPC 2006, March 2006, pp.118-122.

[2] Win, M. Z., Scholtz, R. A.: Ultra-Wide Bandwidth Time Hopping Spread-Spectrum Impulse Radio Wireless Multiple Access Comm. IEEE Trans On Comm, vol. 48, pp. 679–691, (2000).

[3] Scholtz, R. A.: "Multiple Access with Time-Hopping Impulse Modulation," In: Proc. MILCOM 1993, Bedford, MA, pp. 447–450, (1993).

[4] T. Ballal and T.Y. Al-Naffouri, "Low-sampling-rate ultra-wideband channel estimation using equivalent time sampling," Signal Processing, IEEE Transactions on, vol. 62, no. 18, pp. 4882–4895, Sep. 2014.

[5] Das, B., G. K. Mishra, S. Kanungo, and B. K. kumar Sahu. "Performance improvement through interference mitigation techniques in Transmitted Reference UWB system used in WPAN overlay systems." (2015).

[6] Qasem, Nidal, and Bashar Al-Haj Moh'd. "Enhancing the Power Spectral Density of PPM TH-IR UWB Signals Using Sub-Slots Technique." International Journal of Computer Science and Network Security (IJCSNS) 17, no. 1 (2017): 124.

[7] S. Villarreal-Reyes and R. M. Edwards, "On the Behaviour of Simulation-DFT Based Analysis for Spectral Estimation of PPM TH-IR UWB Signals," Loughborough Antennas and Propagation Conference Proceedings, LAPC 2005, April 2005.

[8] Yajnanarayana, Vijaya, Satyam Dwivedi, Alessio De Angelis, and Peter Händel. "Spectral efficient IR-UWB communication design for low complexity transceivers." EURASIP Journal on Wireless Communications and Networking 2014, no. 1 (2014): 1-13.

[9] Marshall, Alan G., and Francis R. Verdun. Fourier transforms in NMR, optical, and mass spectrometry: a user's handbook. Elsevier, 2016.

[10] M. Z. Win, "Spectral density of random UWB signals", IEEE Communications Letters, Vol. 6, No. 12, pp. 526-528, Dec 2002.

[11] S. Villarreal-Reyes and R. M. Edwards, "Spectral Line Suppression in TH-IR Ultra Wideband Systems," 5th IEE International Conference on Mobile Communications Technologies (3G 2004) , Oct 2004.

[12] Federal Communications Commission. "Revision of part 15 of the commission's rules regarding ultra-wideband transmission systems, first report and order (ET Docket 98-153)." Adopted Feb 14 (2002): 2002.

[13] R. Johannesson ."Some rate 1/3 and 1/4 binary convolutional codes with an optimum distance profile". IEEE Transactions on Information Theory, 23 no.2(1977): 281-283.

[14] L. Hanzo, TH.Liew, and B. L. Yeap. Turbo coding, turbo equalisation and space-time coding". John Wiley & Sons, 2002.

[15] Nail, Rana, and Nidal Qasem. "Enhancing the Power Spectral Density of PPM-IR for Ultra-Wide Band signals by using a convolutional encoder." In Internet Technologies and Applications (ITA), 2015, pp. 14-17. IEEE, 2015.

[16] Chase, David. "Code combining-a maximum-likelihood decoding approach for combining an arbitrary number of noisy packets." IEEE transactions on communications 33, no. 5 (1985): 385-393.

[17] Bagad, V. S., and I. A. Dhotre. Computer Networks-II. Technical Publications, 2009.

Three-Dimensional Shape Reconstruction from a Single Image by Deep Learning

Kentaro Sakai[1], Yoshiaki Yasumura[2]
Graduate School of Engineering and Science
Shibaura Institute of Technology
Saitama, Japan

Abstract—**Reconstructing a three-dimensional (3D) shape from a single image is one of the main topics in the field of computer vision. Some of the methods for 3D reconstruction adopt machine learning. These methods use machine learning for acquiring the relationship between 3D shape and 2D image, and reconstruct 3D shapes by using the learned relationship. However, since only predefined features (pixels in the image) are used, it is not possible to obtain the desired features of the 2D image for 3D reconstruction. Therefore, this paper presents a method for reconstructing 3D shapes by learning features of 2D images using deep learning. This method uses Convolutional Neural Network (CNN) for feature learning to reconstruct a 3D shape. Pooling layers and convolutional layers of the CNN capture spatial information about images and automatically select valuable image features. This paper presents two types of the reconstruction methods. The first one is to first estimate the normal vector of the object, and then reconstruct the 3D shape from the normal vector by deep learning. The second one is direct reconstruction of the 3D shape from an image by a deep neural network. The experimental results using human face images showed that the proposed method can reconstruct 3D shapes with higher accuracy than the previous methods.**

Keywords—Computer vision; 3D reconstruction; deep learning; convolutional neural network; feature learning; normal vector

I. Introduction

Reconstructing three-dimensional (3D) shapes of objects from two-dimensional (2D) images is one of the most attractive areas of computer vision. Shape-from-shading is a traditional method of 3D reconstruction that uses object shadow images, reflection characteristics, and light source information [1, 2, 3, 4, 5]. Although 3D shapes can be reconstructed in a relatively short time by the methods, it is not practical due to strong constraints such as object texture. Tal Hassner et al. proposed a method of 3D shape reconstruction by the nearest neighbor method using 2D image and 3D shape database [6]. This method searches the database for the most similar image patches of the input image and integrates them to estimate the shape. However, it takes a lot of time to find the most similar patch. Mori et al. proposed a reconstruction method using bagging [7]. Bagging is a kind of ensemble learning method. This method learns the relationship between the pixel values of the patch in the image and the normal vector at the center of the patch. The patch is the window of the k*k size in the image. The learning method is bagging whose weak learner is a regression tree. By using the learned relationship, the method reconstructs 3D shape from an unknown image. From the experimental results, this method could reconstruct 3D shapes more accurately than the previous methods. However, this method uses predefined feature of the image, the pixel value of the patch. If the method can use the desirable features in images, it reconstructs 3D shape more accurately. The desirable features can be acquired by learning. For acquiring the features of the image, deep learning is the most suitable method because the deep learning has ability for feature learning [8, 9, 10].

Therefore, this paper presents a method for 3D reconstruction from a single image by deep learning. This method adopts Convolutional Neural Network (CNN) [11, 12, 13] for feature learning because it is successful in various fields such as object recognition and semantic segmentation. This method outputs the 3D coordinates of an object from the pixel values of 2D image. This method acquires spatial information of an image using the pooling layers and the convolutional layers of the CNN. From the acquired features, this method reconstructs the 3D shape from a 2D image. For 3D reconstruction, two types of the reconstruction methods are proposed in this paper. The first one is first estimating the normal vector of the object, then reconstructing the 3D shape from the normal vector by deep learning. The second one is direct reconstruction of 3D shape from an image by a CNN.

This paper is organized as follows. In Section 2, a method to reconstruct the 3D shape from a single image by using CNN is described. First, a method for 3D reconstruction by estimating normal vector of the object is presented. Second, this paper presents a method for directly reconstructing 3D shape from a single image. In Section 3, the experimental settings such as hyper parameters and experimental results are presented. Finally, conclusion of this paper is described in Section 4.

II. Three Dimensional Reconstruction by Deep Learning

This section presents a method for reconstructing 3D shape from 2D image by the CNN.

A. Previous Works for 3D Reconstruction

Fig. 1 shows the overview of the previous method for reconstructing a 3D shape by bagging [7]. This method learns the relationship between 2D images and normal maps of the 3D shapes. First, a normal map is created from the 3D coordinates of the object. Next, the method acquires the relationship between the pixel values in the patch and the normal vector of the center of the patch. The patch is a window of the arbitrary k*k size. The relationship is acquired by bagging with a regression tree as a weak learner. By using the acquired relationship, the normal map of the shape is estimated from a new 2D image. Finally, this method reconstructs the 3D shape from the estimated normal map. Since this method uses predefined features, it cannot reconstruct 3D shape from an image accurately because it does not use the valuable features for 3D shape reconstruction.

B. 3D Reconstruction Method by Estimating Normal Vector

Here, this paper proposes a method for 3D reconstruction by estimating normal vectors of the object. This method basically consists of same procedures of the previous work in Fig. 1. The differences between the proposed method and the previous method are that (1) normal map estimator is created with CNN, and (2) 3D coordinates of the object are estimated from the normal map by deep learning. Since the previous method adopts traditional learning methods for estimating normal vectors, the estimated normal vectors are not sufficiently accurate. This is because the previous method cannot utilize valuable features of the image for 3D reconstruction. On the other hand, Convolutional Neural Network (CNN) enables to acquire valuable features of the image automatically. Therefore, the proposed method can reconstruct 3D shape more accurately by using CNN. The input of the CNN is pixel values of an image and the output is normal vectors of the correspondence pixels.

Next, from the estimated normal vector, the proposed method reconstructs 3D shape of the object. Since 3D coordinate estimation from the normal vectors is hard problem, some previous works tackle this problem. These works uses some constrain of the surface such as smoothness. To solve this problem, the proposed method create a deep neural network to learn the relationship between the normal vectors and 3D coordinates of the object, and the network estimates 3D coordinates from the normal vectors. The merit of this method is that the 3D coordinates can be estimated more accurately by using deep learning in spite of the noisy normal vectors. The input of the deep learning is the normal vectors of the pixel, and the output is the 3D coordinates of the correspondence pixels.

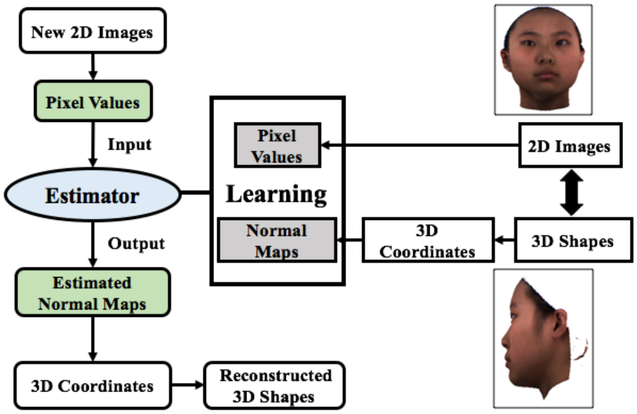


Fig. 1. Overview of the Previous Method.

C. Direct 3D Reconstruction Method

The overview of the direct reconstruction method is shown in Fig. 2. This method directly estimates the 3D coordinates of an object from a 2D image without estimating normal vectors.

The input of this method is pixel values of a 2D image and the output is the corresponding 3D coordinates. CNN learns the relationship between the pixel values and the 3D coordinates of the shape.

Fig. 3 shows the structure of the CNN. The convolution layer contains a set of filters. The filter extracts features in the image. Since some filters detect edge information in the image, they can extract unsmooth shape such as eyes and mouth. Similarly since some filters detect gradation information, they can extract smooth shape such as cheek and forehead. In the learning process of CNN, it can acquire the valuable filters by learning. This characteristic of the convolution layer enables to learn features for 3D shape reconstruction.

The pooling layer summarizes the features in patches, and reduces the spatial dimensions. By using pooling layers, the CNN can be robust to the position shift and rotation in the image. The convolution layers and pooling layers enable the CNN to learn features in the image for 3D reconstruction.

Finally fully connected layers outputs 3D coordinates by combining the features from the convolution layers and pooling layers.

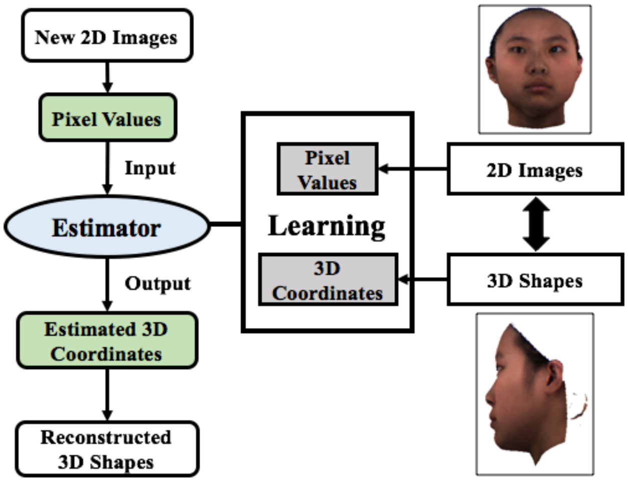


Fig. 2. Overview of the Direct 3D Reconstruction.

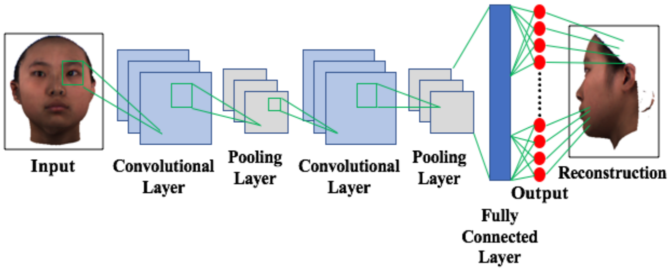


Fig. 3. Convolutional Neural Network for 3D Reconstruction.

III. Experiment

This section presents experimental evaluation of the proposed method. This experiment is conducted with 2D images and 3D coordinates data of actual faces.

A. Experimental Setting

The data of this experiments is Beijing University of Technology's "BJUT-3D Face Database" [14]. Fig. 4 shows an example of the data. The dataset consists of 463 face data that are 2D face images and the corresponding 3D coordinates. The dataset is split into training set (313 data), validation set (75 data), and test set (75 data). The training set is used for creating an estimator to reconstruct 3D shape from an image. The validation set is used for preventing overfitting and tuning hyper parameters. The test set is used for accuracy evaluation of 3D shape reconstruction from an unknown image.

To experiment by using the CNN, the proposed method needs to adjust hyper parameters. Epoch is the number of iteration that training data is repeatedly learned. In mini-batch learning, training set is divided into small subsets (mini-batches). The mini-batch is used for reducing the variance of the gradient. The number of filters indicates the number of filters that are used for features extraction in a convolutional layer. Increasing the number of filters enables to obtain various features. Dropout means disabling some of the nodes in the layers with an update process. Dropout is a way to prevent CNN from overfitting.

To evaluate the robustness of the proposed method, this paper experiments with 2D images of five types of light conditions such as light source direction and light color. The light conditions are: (1) light from upper right (2) darker light (3) light from right side (4) blue light (5) light from upper side. The examples of the images of each light condition are presented in Fig. 5.

B. Experimental Results of Reconstructed Normal Vector

Table I shows experimental results compared with the previous methods in the various types of the light conditions in Fig. 5. The evaluation of the results uses cosine similarity between the estimated normal vectors and the true normal vectors of the object. If the cosine similarity is higher, the result is better. The proposed method exceeded 0.92 under all condition. On the other hand, the previous method has lowered the value under some conditions. These results showed that the proposed method is more robust for various light conditions than the previous methods. Averagingly the proposed method achieved the best result.

C. Experimental Results of Reconstructed 3D Shape

Table II shows experimental results for each hyper parameter by the direct reconstruction method. This experiment adopts Mean Square Error (MSE) for evaluating the results of the proposed method. The MSE means difference between the real shape and the reconstructed shape. From the table, the error is the lowest when the convolution layer has 64 filters. In general, it is assumed that if the number of filters is large, various characteristics can be extracted. However, the error is increasing when the number of filters is the largest. This is because unimportant features are extracted by using too many filters.

The proposed method reduced test error by using dropout. Without dropout, the error in the training set decreases, but the error in the test set increases. From the results, dropout prevents the CNN from overfitting.

Fig. 6 shows the transition of the errors of training and validation data. Since the errors decreases as learning progresses, learning for reconstruction is performed well. Table III shows the errors of the 3D reconstruction in the light condition (1). The table shows that the error by the direct reconstruction method is the lowest. Also the MSE of the reconstruction method by estimating normal vector is less than those of the previous methods. This is due to the reconstruction method from the normal vectors by deep learning. From this result, the proposed two methods can reconstruct more accurately than the previous methods. Table IV shows the MSE by the direct reconstruction method under the different light conditions. The MSEs by the direct reconstruction method in all conditions are higher than the best MSEs of the previous methods as shown in Table III and Table IV. The direct reconstruction method did not lower the MSE values much even under adverse conditions.

Fig. 4. Face Data.

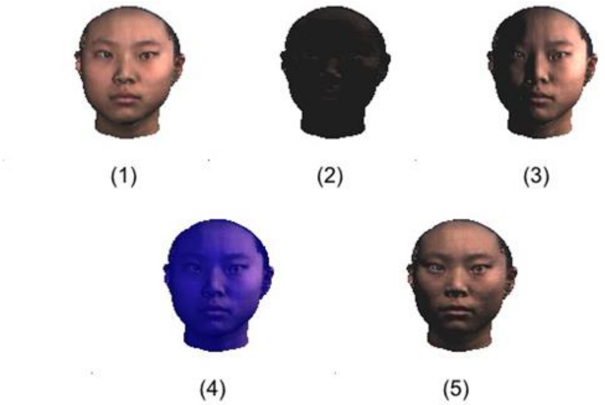


Fig. 5. Examples of Face Images under different Light Condition.

TABLE. I. Results of the Estimated Normal Vectors

	(1)	(2)	(3)	(4)	(5)
Bagging [7]	0.931	0.753	0.882	0.857	0.800
Bagging by optimal learners [7]	0.980	0.884	0.956	0.950	0.908
The proposed method	0.952	0.951	0.950	0.944	0.922

TABLE. II. EXPERIMENTAL RESULTS FOR EACH HYPER PARAMETER

Epoch	50			
No. of mini-batches	8			
No. of filters	32	64	128	64
Dropout	Yes	Yes	Yes	No
Training error	25.51	12.85	17.87	11.07
Test error	16.90	11.39	23.74	14.29

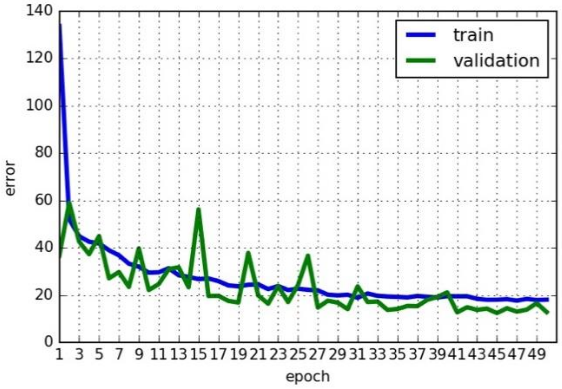


Fig. 6. Sum of the Error in Learning Process.

TABLE. III. THE ACCURACY OF THE RECONSTRUCTED SHAPE

	MSE
Nearest Neighbor method [2]	84.37
Bagging method [3]	80.59
Reconstruction method by estimating normal vector	29.53
Direct reconstruction method	24.87

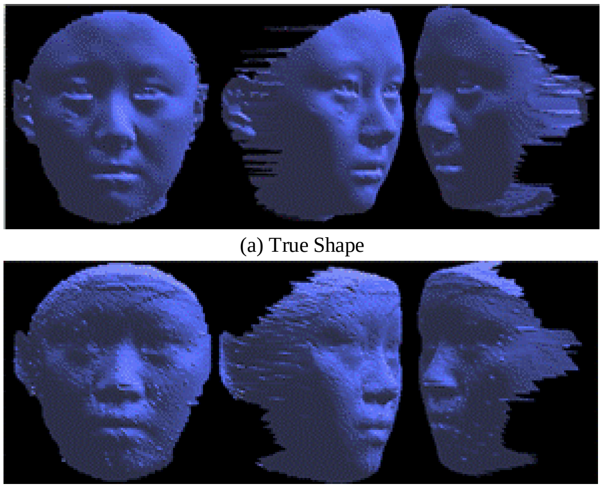
TABLE. IV. MSE BY DIRECT RECONSTRUCTION METHOD UNDER DIFFERENT LIGHT CONDITION

	(1)	(2)	(3)	(4)	(5)
MSE	24.87	30.93	34.25	37.06	27.81

Fig. 7 shows the comparison between the true shape and the reconstructed shape by the proposed method. The true 3D shape is shown in Fig. 7(a), and the reconstructed 3D shapes by the direct reconstruction are shown in Fig. 7(b). From the reconstruct result, the face outline is noisy, but the mouth and nose are reconstructed well. Fig. 8 shows the low accuracy result by direct reconstruction method. Fig. 8(a) shows the true shape of the face, and Fig. 8(b) shows the reconstructed face shape. From the result, the shape is roughly well reconstructed, but there is a lot of noise overall. For more accurate reconstruction, reducing such noise is future task.

Fig. 9 shows the error area that is painted black. Fig. 9(a) shows higher error area in the shape reconstructed with high accuracy, Fig. 9(b) shows the error area in the shape reconstructed with low accuracy. The black area in the figure indicates that the error of the area is higher than 3.0. From Fig. 9(a), the shape can be well reconstructed even for a part such as the nose and the eye which is difficult to reconstruct. However, the shape in Fig. 9(b) has higher errors throughout the face. This result indicates that the reconstruction result varies in accuracy depending on the image.

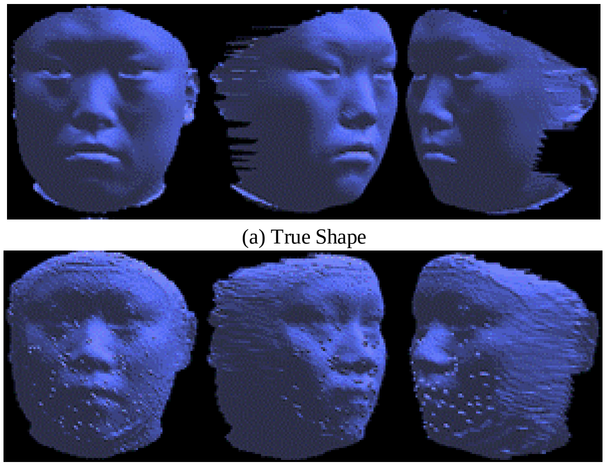
The experimental results show that the proposed method can reconstruct better than the previous methods. However, there are the problems that the reconstruction shape contains noise and the reconstruction error depends on the image. To resolve the problem, the number of the layer in the CNN needs to be increased.



(a) True Shape

(b) Reconstructed Shape

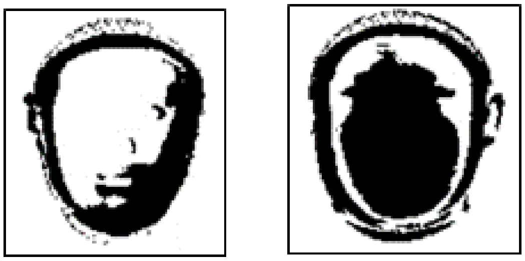
Fig. 7. Results of High Accuracy Reconstruction.



(a) True Shape

(b) Reconstructed Shape

Fig. 8. Results of Low Accuracy Reconstruction.



(a) High Accuracy

(b) Low Accuracy

Fig. 9. Higher Error Area.

IV. Conclusion

This paper presented a method for reconstructing a 3D shape from an image by deep learning. The proposed method outputs 3D coordinates of the shape from the pixel values of an image. The CNN can acquire the valuable features of the image. This paper proposed two types of the reconstruction methods. The first one is first estimating the normal vector of the object, then reconstructing the 3D shape from the normal vector by deep learning. The second one is directly reconstruction of 3D shape from an image by a deep neural network.

From the experimental results, the proposed two methods can reconstruct 3D shape better than the previous method. From the experimental results under the various types of light conditions, the proposed methods are robust to the light conditions. The direct reconstruction method achieved the better results than the previous methods and the reconstruction method by estimating normal vectors by deep learning.

For future work, Generative Adversarial Networks (GAN) [15, 16, 17] is one of the most promising methods for more accurate and smooth shape reconstruction. GAN consists of two neural networks, generative network and discriminative network. For 3D reconstruction, generative network creates a 3D shape from a 2D image, and then discriminative network distinguishes shapes created by the generative network from the true shapes. More accurate shape will be generated by contesting between the two networks.

References

[1] R. Zhang, P.S. Tsai, and J.E. Cryer, M. Shah, "Shape-from-Shading: a survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.21, No. 8, pp. 690-706, (1999).

[2] F. Srtori and E.R. Hancock, "Victor transport for shape-from-shading", Pattern Recognition, Vol. 38, No. 8, pp. 1239-1260, (2005).

[3] Abdelrehim H. Ahmed, Aly A. Farag, "A New Formulation for Shape from Shading for Non-Lamberitian Surfaces", 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR06), Volume 2, pp. 1817 - 1824, (2006).

[4] D. Yang, J. Deng, "Shape From Shading Through Shape Evolution", The IEEE Conference on Computer Vision and Pattern Recognition, pp.3781-3790, (2018).

[5] M. W. Tao, P. P. Srinivasan, S. Hadap, S. Rusinkiewicz, J. Malik, R. Ramamoorthi, "Shape Estimation from Shading, Defocus, and Correspondence Using Light-Field Angular Coherence", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, Issue 3, pp.546-560, (2016).

[6] T. Hassner and R. Basri, "Example Based 3D Reconstruction from Single 2D Images", IEEE Conference on Computer Vision and Pattern Recognition Workshop, pp. 15-15, (2006).

[7] Y. Mori, Y. Yasumura, and K. Uehara:"3D Face Reconstruction from a Single Image Using Machine Learning Methodology", Proceedings of the 2009 ICMITA, pp.29-32 (2009).

[8] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3415-3424 (2017).

[9] Z. Han et al., "3D2SeqViews: Aggregating Sequential Views for 3D Global Feature Learning by CNN With Hierarchical Attention Aggregation," IEEE Transactions on Image Processing, vol. 28, no. 8, pp. 3986-3999, (2019).

[10] Wen Y., Zhang K., Li Z., Qiao Y. , A Discriminative Feature Learning Approach for Deep Face Recognition, Proc. of ECCV 2016, pp. 499-515 (2016).

[11] Andrew G. Howard, Menglong Zhu, Bo Chen,Dmitry Kalenichenko, Weijun Wang, TobiasWeyand, Marco Andreetto, and Hartwig Adam.Mobilenets: Efficient convolutional neural net-works for mobile vision applications.CoRR,abs/1704.04861, (2017).

[12] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick; Mask R-CNN, The IEEE International Conference on Computer Vision (ICCV), pp. 2961-2969 (2017).

[13] K. Kamnitsas, C. Ledig, V. F. J. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, B. Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate nrain lesion segmentation, Medical Image Analysis, Vol;. 36, pp.61-78 (2017).

[14] "BJUT-3D Face Database", Multimedia & Intelligent Software Technology Beijing Municipal Key Laboratory, Beijing University of Technology.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative Adversarial Nets" , Advances in Neural Information Processing Systems 27, (2014).

[16] M. Mirza, S. Osindero, "Conditional Generative Adversarial Nets", arXiv preprint arXiv:1411.1784, (2014).

[17] P. Isola, J. Zhu, T. Zhou, A. A. Efros, "Image-To-Image Translation With Conditional Adversarial Networks", The IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125-1134, (2017).

An Attribution of Cyberattack using Association Rule Mining (ARM)

Md Sahrom Abu[1] , Aswami Ariffin[4]
Malaysian Computer Emergency Response Team
Cybersecurity Malaysia, Cyberjaya
Selangor DE, Malaysia

Siti Rahayu Selamat[2], Robiah Yusof[3]
Faculty of Information Technology and Communication
Universiti Teknikal Malaysia Melaka
Durian Tunggal, Melaka, Malaysia

***Abstract*—With the rapid development of computer networks and information technology, an attacker has taken advantage to manipulate the situation to launch a complicated cyberattack. This complicated cyberattack causes a lot of problems among the organization because it requires an effective cyberattack attribution to mitigate and reduce the infection rate. Cyber Threat Intelligence (CTI) has gain wide coverage from the media due to its capability to provide CTI feeds from various data sources that can be used for cyberattack attribution. In this paper, we study the relationship of basic Indicator of Compromise (IOC) based on a network traffic dataset from a data mining approach. This dataset is obtained using a crawler that is deployed to pull security feed from Shadowserver. Then an association analysis method using Apriori Algorithm is implemented to extract rules that can discover interesting relationship between large sets of data items. Finally, the extracted rules are evaluated over the factor of interestingness measure of support, confidence and lift to quantify the value of association rules generated with Apriori Algorithm. By implementing the Apriori Algorithm in Shadowserver dataset, we discover some association rules among several IOC which can help attribute the cyberattack.**

Keywords*—*CTI; association rule mining; Apriori Algorithm; attribution; interestingness measures

I. Introduction

With rapid development of computer networks and information technology such as internet connectivity, cloud storage and social media, various devices can easily connect to the internet. While this improvement has help internet users to access the latest information quickly, it also has bad consequences where an attacker can improve their tactic, technique and procedure (TTP) to launch a more complicated cyberattack. According to the statistic released by Malaysian Computer Emergency Response Team (MyCERT) as shown in Fig. 1, the number of malicious network activity, specifically on botnet in Malaysia had averagely surpassed 1 million unique IP infections per year [1].

This infection rate had caused a growing concern toward internet users in Malaysia because cybercriminals can manipulate the infected device for illegal activities. The infected machines can be used to deploy malware, initiate attacks on websites, steal personal information and mining cryptocurrencies. The number of infections rate is very alarming, and it causes a lot of problems among the organization because it requires an effective cyberattack attribution to mitigate and reduce the infection rate. Besides, this growing concern among internet users in Malaysia, Cyber Threat Intelligence (CTI) has gain wide coverage from the media due to its capability to provide CTI feeds from various data sources that can be used for cyberattack attribution. However, a proper process of voluminous data available in Cyber Threat Intelligence (CTI) is needed to achieve an effective cyberattack attribution.

Hence, the objective of this paper is to learn more about the relationship of basic Indicator of Compromise (IOC) using network traffic dataset from data mining approach. The network traffic dataset is obtain from Shadow server feed using a crawler. After that the extraction of rules to discover the interesting relationship between large sets of data items is conducted using an association analysis method. As a result, the implementation of association analysis method using Apriori Algorithm on Shadow server dataset can help to attribute the cyberattack based on useful information behind the association rules among several IOC.

The remaining of the paper is organized as follows: Section II presents the research background and related work based on association rules mining in CTI. Section III describes the proposed methodology that includes data collection using CTI feeds, data preprocessing and association analysis using the Apriori algorithm. While Section IV elaborates the rules extraction methods and represents the outcome of using interestingness measures to evaluate the rules generated. Finally, Section V provides a brief conclusion for this paper.

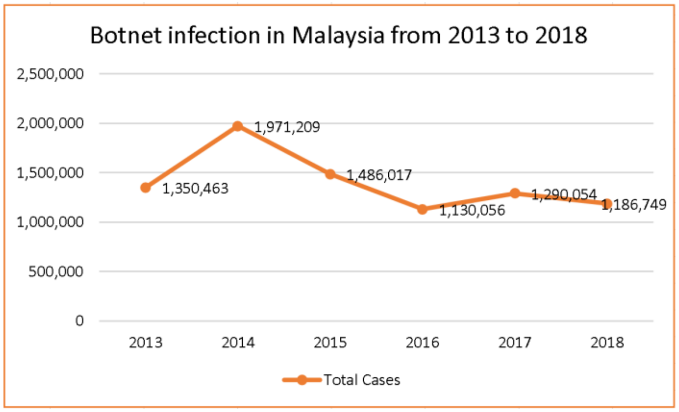


Fig. 1. Statistic of Botnet Infection in Malaysia.

II. Research Background and Related Works

A. Cyber Threat Intelligence (CTI) for Threat Attribution

There is no concrete definition to explain Cyber threat Intelligence (CTI) and it tends to change based on the working environment and business nature [2]. According to Gartner, CTI is evidence-based knowledge, including context, mechanisms, indicators, implications and actionable device, about an existing or emerging menace or hazard to asset that can be used to inform decisions regarding the subject's response to that menace or hazard [3]. While Pokorny et.al. [4] generally defines CTI as data that are collected from the operational environment, then it is processed and refined to produce information as shown in Fig. 2.

This information is then analyzed and transformed into an actionable format that provides intelligence for threat attribution. Although defenses mechanism has evolved, attackers learned and upped their tactic, technique and procedure (TTP) by using fileless malware, code obfuscation, polymorphic designs and dynamic attack infrastructure that made basic IOC useless. So, Threat attribution is a demanding task, complicated and require a comprehensive intelligence or context [5].Threat attribution can be divided into four levels [6][7]. (1) Attribution to the specific hosts involved in the attack, (2) Attribution to the primary controlling host, (3) Attribution to the actual human actor, (4) Attribution to an organization with the specific intent to attack. These attribution levels are achievable through multiple techniques.

Wheeler [8] in his study, has described several techniques for cyberattack attribution that include tracing back based on log records, intrusion detection system, malware analysis and honeypots [9] as a guideline for a security analyst to identify the origin and threat actor behind the cyberattack. However, these traditional cyberattack attribution techniques have a limitation on discovering hidden knowledge beyond an IP address. The knowledge discovery beyond an IP address through an in-depth analysis of the problems from data sources especially focusing on association analysis is significant in helping security analysts to attribute the cyberattack effectively. Hence there has been a lot of research studies in the area of data mining to discover the useful and hidden knowledge among large groups of items or objects in transaction databases, relational databases, or other information repositories using Association Rule Mining (ARM) method.

B. Association Rule Mining (ARM)

Association Rule Mining (ARM) method has attracted many data mining researchers due to its capability to discover useful and interesting patterns from extensive, noisy, fuzzy and stochastic data. This method used to discover the relationship between variables in voluminous data. The strong relationship among variables is called association rules. These association rules contain two steps which are:

Frequent itemset identification (Support as the threshold): Find all frequent itemsets in a database that have transaction support above a predefined minimum threshold.

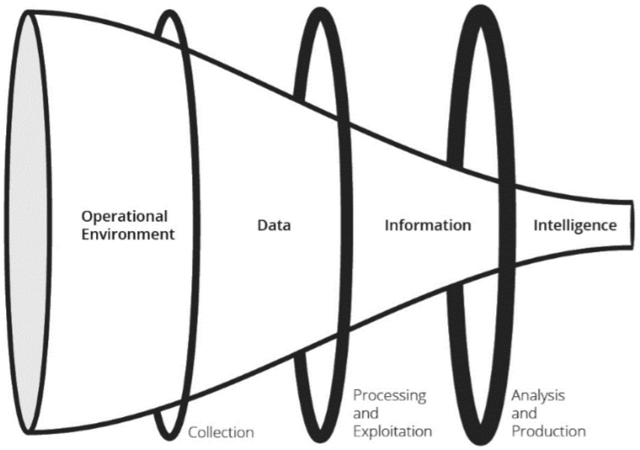


Fig. 2. Relationship of Data, Information and Intelligence [4].

Rule generation (Confidence as the basic): These frequent itemsets use to generate the association rules that have confidence above a predefined minimum threshold.

Finding frequent itemsets in the database require more attention because of the difficulties involves searching for all the possible itemset combination. While rule generation in the second part is a straightforward task, these two parts can be represented as Equation 1.

$$P \rightarrow Q[s, c] \quad (1)$$

Where:

The association between P and Q is whenever P appears Q also be likely to appear. P and Q may be a single condition or set of conditions. P is called the rule's antecedent part and Q is called a consequent part.

s; support is the probability that P and Q found together in a transaction.

c; confidence is the conditional probability that Q found in a transaction when P is present.

Currently, the most widely used algorithms in Association Rule Mining (ARM) is Apriori Algorithm. Agrawal [9] developed this algorithm to study customers' purchasing behavior in supermarkets where goods are often purchased together by customers. Besides, Apriori Algorithm also has been used in many areas of daily life successfully, including energy, recruitment, communication protocol, monitoring and network traffic behavior [10]. Hence the implementation of Apriori Algorithm in determining network traffic behavior can help security analysts to study attacker behavior in conducting cyberattack. Furthermore, the result regarding the attacker behavior from association rules generated using Apriori Algorithm can facilitate the security analyst on cyberattack attribution process.

1) Apriori algorithm: Apriori is an algorithm introduced by R. Agrawal and R Srikant in 1994 [9] and structured to focus on databases that consist of a large amount of transaction data. The basic Apriori algorithm utilised the bottom-up technique that makes it possible to extend the frequent subsets one item at a time which is termed as the candidate generation step. Further, clusters of candidates are evaluated against the data. The algorithm dismisses the process when no further relevant extensions are possible. The Association Rule Mining for Apriori Algorithm is defined as a two-step technique, namely, 1) generating the frequent itemset and 2) generating the rule. In the first step, it involves discovering all frequent item-sets that have support greater than or equal to a pre-determined minimum support count. While in the second step, it involves producing all the relevant Association Rules from frequent item-sets. In this step, it further involves on evaluating the Support and Confidence for all the rules and pruning the rules that do not reach the minimum support and minimum confidence threshold values.

This two-step technique can be elaborated using Table I.

TABLE. I. TRANSACTIONAL DATA

TID	TR-1	TR-2	TR-3	TR-4	TR-5
X	1	1	0	1	1
Y	1	1	1	0	1

Legend:
TID – Transaction ID
TR-Transactional
X-Itemset X
Y-Itemset Y

Table I containing the transactional data set that can be used to understand the basic terminology in Apriori Algorithm as the following:

1) Itemset: This is the collection of one or more items or products from the transactions. The term K-item-set denotes a set of k items or products. As illustrate in Table I, the K itemset represent by itemset $X = (x_1,x_2,\ldots x_k)$ and itemset $Y=(y_1,y_2,..,y_k)$.

2) Support Count: The total number of occurrences of an itemset *X* and/or itemset *Y* is defined as support count

3) Support: Support [11] measure the usefulness of association rules. It is defined as a proportion of transactions in a dataset that contains the itemset. It measures the frequency of association. How many times *X* and *Y* involved in association rules occur together in the dataset. When the frequency of *X* and *Y* occurring at the same time is equal to or greater than the designated minimum support threshold, *X* and *Y* meet frequent itemsets. Support can be represented as Equation 2.

$$support\ (X \rightarrow Y) = \frac{Transactions\ containing\ both\ X\ and\ Y\ items}{Total\ number\ of\ transactions} \quad (2)$$

$$support\ (X \rightarrow Y) = \frac{3}{5} = 60\%$$

4) Confidence: Confidence is importance because it can indicate the strength or the reliability of an association rules [11]. It is defined as the ratio of the number of transactions that include all items in a frequent itemset to the number of transactions that include all items in the subset. It determines how frequently item *Y* occurs in the transaction that contains *X*. Confidence represented the conditional probability of an item as shown in Equation 3.

$$confidence\ (X \rightarrow Y) = \frac{Total\ number\ of\ transactions\ containing\ X\ and\ Y}{Total\ number\ of\ transactions\ containing\ item\ X} \quad (3)$$

$$confidence\ (X \rightarrow Y) = \frac{support\ (X,Y)}{support\ (X)}$$

$$confidence\ (X \rightarrow Y) = \frac{3}{4} = 75\%$$

5) Lift: The lift value is a measure of the importance of a rule [11]. The lift measures how frequent *X* and *Y* occur together than expected if they were statistically independent. Lift value 1 indicates *X* and *Y* are independent. The lift can be represented as Equation 4.

$$lift\ (X \rightarrow Y) = \frac{support\ (X{\rightarrow}Y)}{support\ (X) * support\ (Y)} \quad (4)$$

$$lift\ (X \rightarrow Y) = \frac{(\frac{3}{5})}{(\frac{4}{5}) * (\frac{4}{5})} = 0.384$$

The lift is a value between 0 and infinity:

a) If Lift (I) < 1, then X and Y are said to be interdependent on each other negatively.

b) If Lift (I) = 1, then X and Y did not find themselves dependent on each other and said they were independent.

c) If Lift (I) > 1, then X and Y appear together more often in the data and are said to depend on each other positively.

6) Frequent Itemset: The value of Support and Confidence determines the interestingness of the generated rule. This achieved by setting the minimum support and minimum confidence thresholds. The Item-sets whose support is greater than or equal to the specified minimum support threshold is defined as the frequent itemset.

Regardless of how the association rule is defined, it requires a suitable measure to achieve relevant and effective association rules by measuring the strongest dependencies between variables. For example, interestingness measures such as support, confidence, lift, correlation, and entropy have been used extensively to evaluate the interestingness of association rule.

2) Interestingness measure in ARM: Piatetsky-Shapiro introduced rule interestingness (RI) measures to evaluate the values of patterns [12] objectively. This measure can effectively quantify the correlation between the antecedents and the consequent for enormous association rules generated by ARM that can meet the aims of the researcher. The relevant and effective association rules are achievable through interestingness measure that includes objective measure and subjective measure [11]. The objective measures based on the statistical strengths or properties of the generated rules and subjective measures that are obtained from the user deduction or interest of their problem domain. Most of the research in the data mining field has use support and confidence as the de-facto "interestingness measures" for discovering relevant association rules [12]. However, support and confidence do not capture the correlation that exists between the antecedent and consequent of an association rule. There are several interestingness measures such as Laplace, Conviction, or Lift that can be used to fix this shortcoming. Among the three measures, Lift is the simplest yet the most powerful in capturing and representing the type of correlation exists between antecedent and consequent in association rule [13]. So, this paper will be used support and confidence as the de-factor interestingness measure and complement it with lift to evaluate the relevant association rule in facilitating cyber attack attribution.

III. Proposed Methodology

In this research, Association Rule Mining in CTI framework is performed using Apriori Algorithm as shown in Fig. 3. Fig. 3 illustrates the entire process of association rule mining in CTI framework that consists i) Preprocessing network traffic data, ii) Generating logical rules using Apriori algorithm and iii) Apply the generated rule to facilitate cyberattack attribution. The Apriori Algorithm can discover groups of items occurring frequently together in lots of transactions and such groups of items are called frequent itemsets. The association rule generated from this process is measured using support, confidence, and lift. Given a set of transaction, the problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support (called minsup) and minimum confidence (called minconf), respectively.

The implementation of Apriori Algorithm on Shadowserver dataset was done using R language. The capability of R language in performing statistical computing, data mining and graphics was optimize in this paper to process the filtered data and visualization.

A. *Threat Intelligence Feeds*

Data collection for this paper is limited to CTI feeds from OSINT that related to network intrusion activities. For this paper, OSINT CTI feeds from Shadowserver as shown in Fig. 4 has been chosen because it can provide various types of useful information and Indicators of Compromise (IoC) for cyberattack attribution [14]–[16]. The focus of this research is to gather CTI data that contain network resources from existing cyberattack.

Fig. 4 shows data collection process for Shadowserver dataset. Data collection started with collecting popular network resources such as the domain of search engines or government website, IP address of common DNS server and MD5 hash value of notorious malware. This network resource collected using crawler and APIs provide by Shadowserver. Then this data stored in excel storage before going through data preprocessing phase for data integration and data cleaning. The list of network resources and features from CTI data using APIs provide by Shadowserver is shown in Table II.

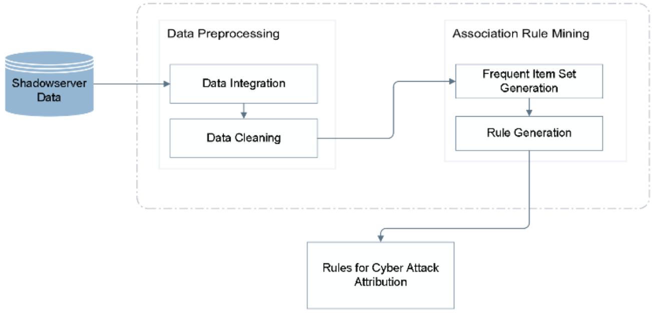


Fig. 3. Data Processing and Association Rule Analysis in CTI Framework.

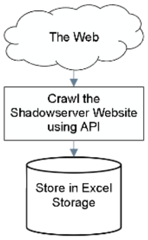


Fig. 4. Data Collection Process.

TABLE. II. The Detail Fields in Shadowserver feed

Data Source [Shadowserver]	Number of records	Features	Description
Dataset 1	334848	Timestamp	The start time of infection or attack occur
		Destination IP	malicious IPs used in the attack among the IPs reported in the threat report associated with a particular network resource.
		Source IP	
		Infection type	Malware type detect by antivirus

As far as the Shadowserver dataset is concerned, a record for network resources from this dataset is prepared with certain practical values correlated to precise network attributes [17]. Hence, in this research, the selected network resource for Shadowserver data include the following: the timestamp, the destination IP address, the source IP address, and the infection type detect by antivirus software. This dataset is obtained to discover significant knowledge behind raw data using ARM to facilitate cyberattack attribution process.

B. *Data Preprocessing for CTI Feeds*

Data scientists spent 80% of their effort in data preprocessing to produce intelligence from raw data that come in various formats and data types [18]. Data preprocessing can ensure the data in our possession are all fit, applicable and clean. It also plays a crucial role in increasing the accuracy of decision making by providing quality data. Data preprocessing phase consists of data integration and data cleaning that can be used to produce a clean and useful data before it can be used for ARM. This process is very important to make sure ARM can produce an effective association ruleset for cyberattack attribution. Data cleaning must be performed to identify potential issues with CTI feeds. With dirty, incomplete, noisy or otherwise "garbage in garbage out" data, the CTI framework unable to produce an effective cyberattack attribution. There are two steps are taken on performing a data preprocessing namely Data Integration and Data Cleaning. In Data Integration, all Shadowserver data merged into a single dataset; doing this results in duplicate IOCs such as IP addresses, hash value, domain name, URL and GeoIP. These duplicate IOCs are vital because it shows the correlation between different feeds. However, storing duplicate IOCs create redundant data. So, we need to perform data cleaning for this data set. While, in Data Cleaning, the cleaning is performed on the incomplete data set by filling in missing values or removing them altogether, along with eliminating noisy data and outliers. Other than that, identifying and repairing issues with the text that may cause data to become misaligned, such as embedded special characters, tabs or line breaks also perform. Once we get the clean data, we use these to do association rule mining to get the rules about network intrusion attack and analyze the meaning of the rules to facilitate cyberattack attribution.

C. *Association Rule Mining Algorithm in CTI Famework*

After the CTI feeds have been preprocessed for producing clean and useful data, the results will be used for association analysis to formulate an association ruleset. This association ruleset is to facilitate a cyber-attack attribution process in the CTI framework to produce an effective threat attribution. The association ruleset can assist security analysts in identifying the origin of the cyberattack and cyberattack attribution level.

As for the experimental setup using R, the configuration of threshold for minimum support value is 0.001 and the threshold of minimum confidence value is 0.5. There are 43 association rules meet this threshold configuration. Therefore to have a practical overview on the result generated, we use scatter plot as shown in Fig. 5 to visualize the association rules while Fig. 6 provides more information about these 43 rules using matrix visualization of grouped antecedents. The top-left corner plot of this matrix represents the most interesting rules based on lift measure.

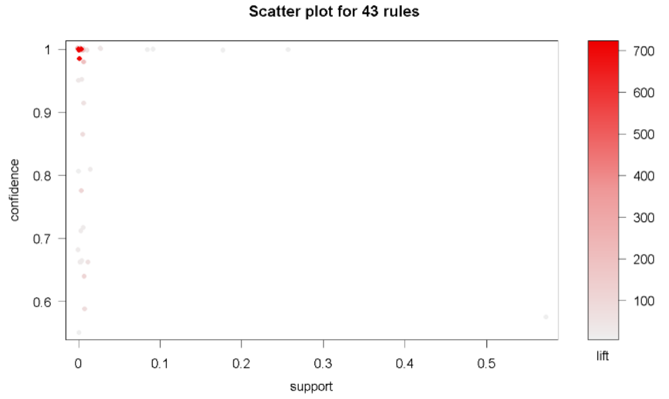


Fig. 5. The Illustration of 43 Rules in Scatter Plot with Minsup = 0.001 an Mincof = 0.5.

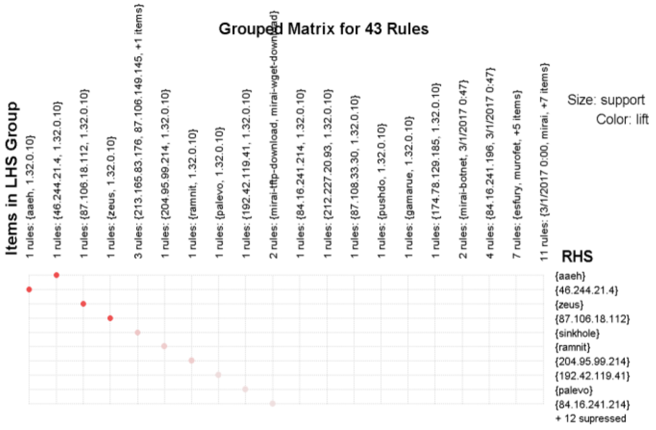


Fig. 6. The Visualization of 43 Rules in Group Matrix with Minsup = 0.001 and Mincof = 0.5.

Generally the experimental setup of this proposed methodology limit the value of support to 0.1 due to the large scale of dataset that we obtained and the confidence is configured between 0 to 1 which between the range allowed for confidence [19]. While the value of minimum support and the minimum confidence is adjusted manually to discover some specific and interesting rules from a large number of random rules [20]. Finally, the result from this experiment is presented in a table and sorted based on different measurement (support, confidence and lift) to analyze the relationship between IOC for this association rules.

IV. Association Rule Evaluation using Interestingness Measure

The number of association ruleset generated using ARM can be massive and even tricky for domain specialists to study and summarize the meaning behind the ruleset. Moreover, it is also impractical to sift through a broad set of rules containing noise and irrelevant rules. As a solution to this issue, interestingness measure can be used for filtering or ranking association rules. As discussed in Section 2, this paper will only focus on objective interestingness measure specifically using support, confidence and lift. While, the thresholds for minimum support (minsup) and minimum confidence (minconf) are manually defined by user [20][21][10].

A. *Evaluation by Lift*

Table III depicts the top five association rules with respect to lift measure. There are three categories to interpret the relationship of X / Y in lift measurement. If the lift is equal to 1, it means that X and Y are independent. If the lift is higher than 1, it means that X and Y are positively correlated. If the lift is lower than 1, it means that X and Y are negatively correlated. Based on Table III, we can see that the itemset 46.244.21.4, aaeh, 87.106.18.112 and zeus respectively have a positive correlation relationship with each other. While the item set 213.165.83.176 also has a positive correlation relationship with sinkhole. It shows that this IP is malicious and being sinkhole by an organization due to it malicious activity.

B. Evaluation by Support

Table IV shows top five association rules based on support with threshold configured as minsup = 0.15 and minconf = 0.15. Meanwhile, the visualization in Fig. 7 illustrates ten rules that satisfy this configuration.

Basically the top 10 rules in Fig. 7 sum up the combination of rules among Mirai, 212.61.180.100, 195.38.137.100 and Dorkbot which indicate there is a strong association among these four items that frequently occur together. Mirai is a malware that turns poorly conFig.d networked devices running Linux into botnets that can be used to launch a large-scale cyberattack that specifically targeted Internet of Thing (IOT) devices such as home routers, DVRs and webcams [22]. While, Dorkbot is a family of malware worms that spreads through instant messaging, USB drives, websites or social media channels like Facebook. Based on these two variants we can deduce that IP 212.61.180.100 belongs to an IOT device such as IP cameras or home router that has been infected through USB drives while IP 195.38.137.100 has been infected by dorkbot variant through malicious link that spread in social media such as Facebook and WhatsApp.

C. Evaluation by Confidence

Confidence measure can provide the most reliable association rule generated in experimental setup. Table V presented the top 5 most reliable rules with a threshold for minsup = 0.001 and minconf = 0.2 while Fig. 8 visualize the top 10 rules for this measurement.

The top 10 rules based on confidence measurement shows that high confidence rules usually related to sinkhole, zeus and aaeh.There is three IP that being sinkhole because being compromise and use by an attacker to launch malicious activity. IP 46.244.21.4 that associates with aaeh variant being used as a dropper to download other malicious code. While IP 87.106.18.112 that strongly associates with zeus being used to steal sensitive information that related to financial data.

TABLE. III. TOP 5 RULES BASED ON LIFT MEASURE WITH MINSUP = 0.001 AND MINCONF=0.5

Antecedent (*X*)	Consequent (*Y*)	Support	Confidence	Lift
46.244.21.4	aaeh	0	1	719.2
aaeh	46.244.21.4	0	1	719.2
87.106.18.112	zeus	0	1	702.2
zeus	87.106.18.112	0	0.98	702.2
213.165.83.176	sinkhole	0	1	231.7

TABLE. IV. TOP 5 RULES BASED ON SUPPORT MEASURE WITH MINSUP = 0.15 AND MINCONF=0.15

Antecedent (*X*)	Consequent (*Y*)	Support	Confidence	Lift
mirai	212.61.180.100	0.26	1.00	1.74
212.61.180.100	mirai	0.26	0.45	1.74
dorkbot	212.61.180.100	0.23	0.45	0.79
212.61.180.100	dorkbot	0.23	0.39	0.79
195.38.137.100	dorkbot	0.18	1.00	2.00

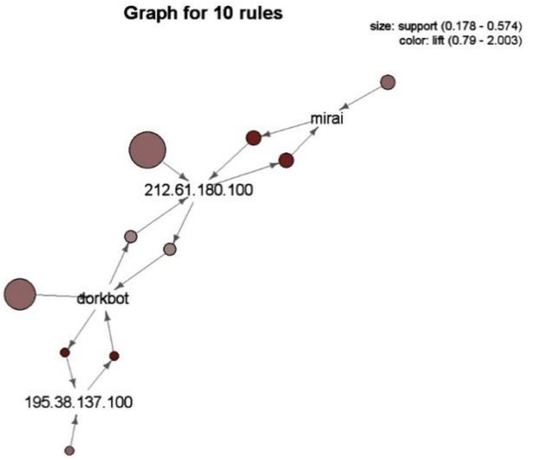


Fig. 7. Top 10 Rules based on Support Measurement with Minsup=0.15 and Minconf=0.15.

TABLE. V. TOP 5 RULES BASED ON CONFIDENCE MEASURE WITH MINSUP = 0.001 AND MINCONF=0.2

Antecedent (*X*)	Consequent (*Y*)	Support	Confidence	Lift
213.165.83.176	sinkhole	0	1	231.74
87.106.149.145	sinkhole	0	1	231.74
46.244.21.4	aaeh	0	1	719.19
aaeh	46.244.21.4	0	1	719.19
87.106.18.112	zeus	0	1	702.24

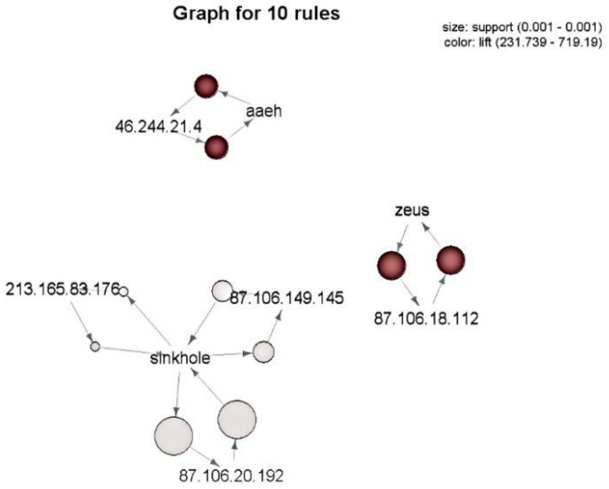


Fig. 8. Top 10 Rules based on Confidence Measurement with Minsup=0.001 and Minconf=0.2.

In this research, the same process is done using three different datasets from Shadowserver in order to prove the proposed work as shown in Table VI.

TABLE. VI. THE DETAIL FIELDS IN LEBAHNET FEED

Data Source [Shadowserver]	Number of records
Dataset 2	332874
Dataset 3	325730

Based on the results obtained, it discovers the strongest association rule indicates several malicious IP being targeted by cybercriminal is used to steal sensitive information that related to financial data. This association rule can help security analyst to focus on attack campaign and threat actor that related to financial attack. Apart from that, there is also an association rule that involved IOT devices such as IP camera and home router. This device most probably been compromised as a botnet to launch DDoS activity. From there, security analyst can focus on this IP for attack campaign and threat actor that actively involved in DDoS attacks. Security analyst also can further attribute this IP through pivoting and enrichment using third-party tools such as Passive DNS, Domain Tools IRIS and Maltego.

V. Conclusion and Future Works

Cyber threat intelligence provides a massive amount of raw data that contained useful information behind it. Association rule mining can help to discover significant knowledge behind this raw data to facilitate cyberattack attribution process in CTI framework.

In this paper, we employ Apriori algorithm to process CTI feed from Shadowserver dataset. Firstly, we explain the structure of Shadowserver dataset before going through data preprocessing process (data integration and data cleaning) using R language. Secondly, the Apriori Algorithm was explained in implementing the step of generating association rules. Finally, we evaluate the association rule using support and confidence as the de-factor in interestingness measure and complement it with lift to obtain the strongest association rules that reflect attacker characteristics when launching cyberattack. The finding of the experiment showed that the useful information about cyber-attack and attacker by association analysis can be discovered based on threat intelligence. In addition, the output data of association analysis can provide the information of cyber-attack relationship in cyber-attack attribution. For future work, more association rule algorithm and other statistical measures can be implemented to improve association ruleset effectiveness and accuracy in facilitating cyberattack attribution.

Acknowledgment

This study was kindly supported by The Ministry of Communications and Multimedia (KKMM), Cybersecurity Malaysia and Universiti Teknikal Malaysia Melaka (UTeM).

References

[1] MyCERT, “Malaysia Incident Statistic Report,” 2019. [Online]. Available: https://www.mycert.org.my/portal/statistics?id=b75e037d-6ee3-4d11-8169-66677d694932. [Accessed: 05-Jun-2019].

[2] Md Sahrom Abu and R. Y. , Siti Rahayu Selamat, Aswami Ariffin, “Cyber threat intelligence – Issue and challenges,” Indones. J. Electr. Eng. Comput. Sci., 2018.

[3] Gartner, “Definition: Threat Intelligence,” 2017. [Online]. Available: https://www.gartner.com/doc/2487216/definition-threat-intelligence. [Accessed: 10-Nov-2017].

[4] Z. Pokorny et al., The Threat Intelligence Handbook, Second Edition. 2019.

[5] L. Perry, B. Shapira, and R. Puzis, “NO-DOUBT: Attack attribution based on threat intelligence reports,” 2019 IEEE Int. Conf. Intell. Secur. Informatics, ISI 2019, pp. 80–85, 2019.

[6] J. Ryu and J. Na, “Security Requirement for Cyber Attack Traceback,” in 2008 Fourth International Conference on Networked Computing and Advanced Information Management, 2008, vol. 2, pp. 653–658.

[7] J. Hunker, B. Hutchinson, and J. Margulies, “Role and Challenges for Sufficient Cyber-Attack Attribution,” Inst. Inf. Infrastruct. Prot., pp. 5–10, 2008.

[8] D. A. Wheeler and G. N. Larsen, “Techniques for Cyber Attack Attribution,” Inst. Def. Anal. Rep., no. October, 2003.

[9] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules (expanded version). Research Report IBM RJ 9839,” Proc. 20th Intl. Conf. VLDB, pp. 487--499, 1994.

[10] Y. Liu, K. Yu, X. Wu, Y. Shi, and Y. Tan, “Association rules mining analysis of app usage based on mobile traffic flow data,” 2018 IEEE 3rd Int. Conf. Big Data Anal. ICBDA 2018, pp. 55–60, 2018.

[11] C. Ju, F. Bao, C. Xu, and X. Fu, “A Novel Method of Interestingness Measures for Association Rules Mining Based on Profit,” vol. 2015, no. 2, 2015.

[12] M. Jalali-Heravi and O. R. Zaïane, “A study on interestingness measures for associative classifiers,” in Proceedings of the 2010 ACM Symposium on Applied Computing - SAC ’10, 2010, no. June, p. 1039.

[13] N. Hussein, A. Alashqur, and B. Sowan, “Using the interestingness measure lift to generate association rules,” J. Adv. Comput. Sci. Technol., vol. 4, no. 1, p. 156, 2015.

[14] X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah, “Acing the IOC Game : Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence,” pp. 755–766, 2016.

[15] Z. Zhu and T. Dumitras, “FeatureSmith: Automatically Engineering Features for Malware Detection by Mining the Security Literature,” Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Secur. - CCS’16, pp. 767–778, 2016.

[16] C. Sabottke, O. Suciu, T. Dumitraş, C. Sabottke, and T. Dumitras, “Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits,” Proc. 24th USENIX Secur. Symp., 2015.

[17] S. Prakash and M. Vijayakumar, “An Effective Network Traffic Data Control Using Improved Apriori Rule Mining,” Circuits Syst., vol. 07, no. 10, pp. 3162–3173, 2016.

[18] J. Pérez, E. Iturbide, V. Olivares, M. Hidalgo, N. Almanza, and A. Martínez, “A data preparation methodology in data mining applied to mortality population databases,” Adv. Intell. Syst. Comput., vol. 353, pp. 1173–1182, 2015.

[19] A. Shah, “Association rule mining with modified apriori algorithm using top down approach,” Proc. 2016 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. iCATccT 2016, pp. 747–752, 2017.

[20] S. Mahmood, M. Shahbaz, and A. Guergachi, “Negative and positive association rules mining from text using frequent and infrequent itemsets,” Sci. World J., vol. 2014, 2014.

[21] L. Yan, Y. Ke, and W. Xiaofei, “Association Analysis Based on Mobile Traffic,” 2014 4th IEEE Int. Conf. Netw. Infrastruct. Digit. Content, 2014.

[22] P. Asghari, A. M. Rahmani, and H. H. S. Javadi, “Internet of Things applications: A systematic review,” Comput. Networks, vol. 148, no. 4, pp. 241–261, 2019.

Extending Tangible Interactive Interfaces for Education: A System for Learning Arabic Braille using an Interactive Braille Keypad

Hind Taleb Bintaleb[1], Duaa Al Saaed[2]
Information Technology Department, College of Computer and Information Sciences
King Saud University, Riyadh, KSA

***Abstract*—Learning Braille for visual impairments means being able to read, write and communicate with others. There exist several educational tools for learning Braille. Unfortunately, for Arabic Braille, there is a lack of interactive educational tools and what is mostly used is the traditional learning tools, such as the Braille block. Replacing those tools with some more effective and interactive e-learning tools would help to improve the learning process. This paper introduces a new educational system with a tangible and interactive interface. This system aims to help blind children to learn Arabic Braille letters and numbers using an interactive tactile Braille keypad together with the educational website. The interactive tactile Braille keypad was built using an Arduino connected with the educational website. A usability test was conducted and results showed that the system is easy to use and suggested that using the interactive Braille keypad with the educational website will improve the learning outcomes for blind children.**

Keywords—Braille; tangible interface; e-learning; Arduino; accessibility; usability; visually impaired; blind

I. Introduction

In our life, literacy which is the ability to read and write is substantial in different aspects such: education and communication. It helps one to lead their own life without depending on others. Learning to read and write is considered an important step especially in at childhood. It is important to gain basic knowledge about the language and enhance skills in reading and writing.

People who are blind or have low vision need to practice their lives normally. They have the right to learn everything such: reading and writing. The Braille system helps them to understand and to study non-visual modes of communication.

Braille is a tactile system for reading and writing used by blind people or those who have poor vision. Braille was named after its inventor Louis Braille (1809 – 1852) who was blind. In 1824, at age 15 Braille presented this system to help blindness and people with poor vision in reading and writing quickly and more easily. Braille is not a language, but it is a representation of any language character using 6 dots ordered in a matrix of 2 x 3 “two columns with three rows”, see Fig. 2. A certain set of dots when raised represent a certain character, sixty-four combinations are possible using one or more of these six dots [1].

A basic Braille template (cell) is a tactile configuration of six raised\embossed dots. It is upright rectangular shapes made of two vertical columns made of three dot positions. The cell is organized as a matrix of 2 × 3 dots. Those dots are numerically identified by the numbers 1 to 6, see Fig. 2. There are different combinations of raised\embossed dots, each unique configuration represents an alphabetical letter or a number or a symbol. a consonant, a vowel, a number, a diacritical mark or an abbreviated suffix [2][3]. Through the combination of dot positions and their distribution on the two vertical columns, the symbol takes a distinctive tactile shape. Empty dot positions help the reader identify the embossed positions forming the letters. Between dot cells there is a barrier. The direction of embossing symbols is right to left, and reading goes from left to right, even in Arabic and in top-to-bottom scripts [3].

Learning Braille is challenging, it takes time and practice. It needs the support and encouragement of family and teachers; thus, it is important to provide students, their families and teachers with digital tools that could enhance this process and make it easy and fun. There are several educational tools for learning English Braille. Unfortunately, for Arabic Braille, we still use the old traditional learning tools, such as the Braille block, see Fig. 2. Replacing those tools with some more effective and interactive tools would help to improve the learning process. Thus, in this study present an Arabic Braille learning system with an interactive tactile Braille keypad.

The paper is structured as follows: Section 2 shows the related work. A description of software accessibility will be presented in Section 3. In Section 4, we will present our developed Arabic Braille educational system. Section 5 covers system and usability testing. We conclude with Section 6 where we discuss conclusions and future work.

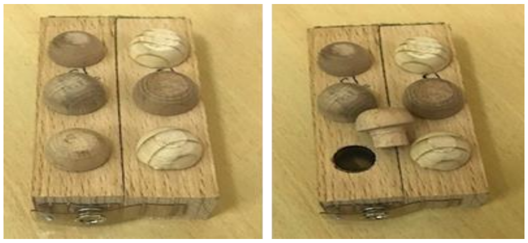


Fig. 1. Traditional Braille Alphabetic Learning Tool [4].

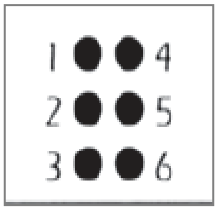


Fig. 2. The Structure of the Basic Braille Cell [2].

II. Related Work

Learning Braille is very important for the blind and people with low vision to be able to read and write.

Many studies have been conducted presenting different solutions to enhance the process of learning Braille. In our review of literature, we found that some of the research studies focuses on presenting a Braille learning system using self-contained hardware. One of the most recent studies [19] conducted a BraillePad to help the children learning English and Tagalog Braille, another study [20] is proposed to learn Spanish Braille for visually impaired users. In [5] they used US English QWERTY keyboard with 6 solenoids and microcontroller and an alphabet input will be entered by a non-visually impaired person to help the visually impaired user in learning the Arabic Braille character.

On the other hand, other studies presented systems with assistive integrated hardware together with an educational software for learning Braille. Various studies have conducted an incorporate environment of hardware and software to help in Braille learning process. One of the most recent studies in this domain is [6] which introduced an innovative system to learn Taiwanese Braille using Braille learning gloves. Similar to the previous study in [7] built a Vibrotactile Braille-Reading Device called UbiBraille. It is inspired by the standard writing system of the Perkins Brailler. In [8] study the (OBR) system was designed. It is an innovative use of RFID and developing new uses for RFID technologies. The OntoBraille@RFID (OBR) it is a Braille learning platform consisting of identified RFID tags, RFID reader and the learning software.

Generally, each of study provides a set of different hardware combination and innovative for Braille learning. Table I summarizes the studies and shows additional information about the used technologies. . It can be clearly seen, that only two out of the four studies developed systems supporting two languages for support Braille system, while the remaining supported only one language. None of these systems is supporting the Arabic Braille. Not all systems proposed in those studies are providing three levels of learning (i.e. character, word and sentence), four out of six have covered learning characters and words, and one, system covered only characters. One study has presented a system that covers all three levels and it supported Chinese and English languages for Braille system.

As far as we know, no studies available have proposed or developed a Braille learning system (hardware integrated with software) in Arabic language. Thus, in this study aims to presenting an Arabic Braille learning system consisting of both hardware and software components. It integrates a specially designed Braille interactive keypad with Arduino technology, together with a web-application for learning Arabic Braille (numbers and characters) which none of the previous systems have provided.

III. Software Accessibility

Accessibility is a term which have a range of definitions. It describes how is it easy to use E-Systems by people with special needs [9]. It means to remove any barriers that face people with special needs from performing life activities and other services [10]. Any software should be accessible by normal users and those with special needs. Software accessibility is extremely important and will always be quite challenging for software designers, they need to ensure that people with disabilities can get full access to a system so they can interact, navigate, perceive, and understand it [11]. In order to achieve accessibility, software designers and developers take into consideration the fact that the ability to see, hear, interact with the system, read text or process information varies from user to user. In fact, some users require special assistive input and output technologies to help them carry out these activities. Several key legal issues, accessibility guidelines, and resources are available to help in making software accessible to individuals with disabilities. Here we try to summaries some of the main issues to be considered, when designing a software, to improve the accessibility for the visually impaired [12][9]:

- Text must be resizable and high color contrast is important. The visual information should be represented in a descriptive text associated with images and other multimedia and vice versa[12][12][12][12][12][12][12].
- All visual content must be supported with an alternative auditory and readable content. In the case of images, this mean adding an alt attribute to describe its content.
- The language of the page contents should be declared with markup to facilitate the pronunciation for screen readers.
- User interface should follow principles of accessible design, such as: device-independent access to functionality, keyboard operability, self-voicing.
- Enabling the feature of multimodality of input such: keyboard, speech, mouse or other pointing device.
- Grouping elements and providing contextual information about the relationships between elements can facilitate access to those elements.
- Documents should be clear and simple and easily understood.

IV. Developed Braille Educational Sysetm

In this section we present our design process and discuss system architecture, hardware and software system components.

A. System Architecture

The architecture of the proposed system is a Client-Server architect. The rationale behind this design decision is as follows:

- The data on a remote server can be accessed by any device (PCs, mobiles, tablets and laptops) via the internet.
- No need for installation on the user device.
- Web-based systems are supported by NVDA screen reader in contrast to a built-in PC program.
- Web-based systems support the accessibility standards and every element can be accessed by the keyboard commands such: tabindex, alt and access key properties, whereas the PC programs are not.
- Web-based systems can be converted to the mobile view so, the users can open the sites by their mobiles easily and efficiently.

An illustration of the system architecture is presented in Fig. 3.

B. Database Architecture

MySQL BD has been used to store all the letters, numbers, sounds and images. Database is used instead of files or arrays cause the large data we have. Storing data in a DB is more manageable than storing in files or arrays. ER of the system is illustrated in Fig. 4.

TABLE. I. RELATED BRAILLE LEANING SYSTEMS SUMMARY

Ref	Braille Language	Learning Level	Assistive Technique Type	Technology
[13]	English and Tagalog	characters and words	Auditory feeding	Arduino Mega and LCD screen
[6]	Taiwanese	characters and words	Auditory feeding and vibrotactile capabilities.	a small-sized microcontroller (MTK7697) with built-in Bluetooth and battery module
[14]	Spanish	characters and words	Auditory feeding and tactile material	Portable device with Bluetooth adopted in, and USB inputs
[5]	Arabic	characters	Tactile material	Microcontroller PIC16F877A, LCD display
[7]	English	characters and words	vibrotactile capabilities	Arduino Mega ADK board, vibration motor
[8]	Chinese and English	Characters, words and sentences	Auditory feeding using MS Speech and tactile material	RFID

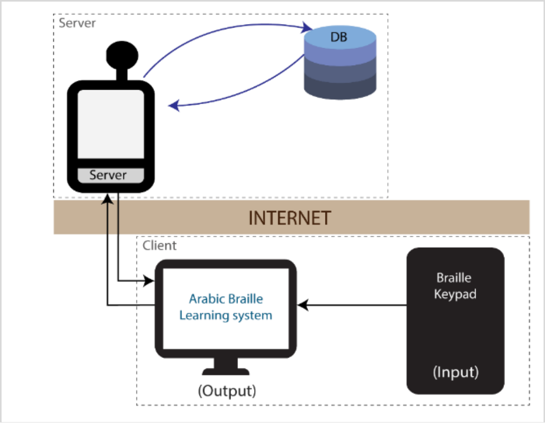


Fig. 3. The System Architecture.

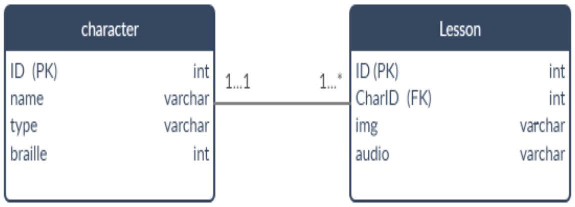


Fig. 4. Entity Relation Diagram.

C. Interactive Braille Keypad

It is an input Braille keypad device designed to implement the traditional Braille block to be used to navigate throughout the system and interact with it.

It has eight buttons: six for the Braille dots representation, an Enter button and a Tap button to navigate through the system elements. It is the main input device in the system. The Braille keypad device has two parts, first part consists of six numbered buttons dots forming two columns, three buttons each. Second part consists of two buttons for Enter and Tap commands. Fig. 5 illustrates the Braille keypad sketch. The keypad was designed to be more accessible and easier to use one hand.

D. Hardware

Mini Arduino Leonardo was used to control the keyboard commands. Arduino controller will be integrated inside the keypad. It consists of the following:

- Mini Arduino Leonardo
- Buttons (2 pieces)
- Toggle-switches (6 pieces)
- USB connector
- Other connectors

Arduino Leonardo: according to the official Arduino web site: “Arduino is easy to use hardware and software electronic platform, which is an open source” [15]. Arduino has many different versions and types, Arduino Leonardo Fig. 6 was used because it has a built-in USB communication and it controls the keyboard commands.

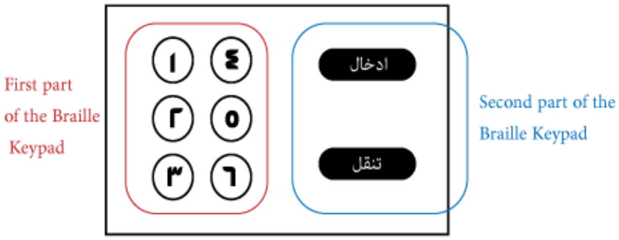

First part of the Braille Keypad
١
٤
٢
٥
٣
٦
ادخال
تنقل
Second part of the Braille Keypad

Fig. 5. Braille Keypad Sketch.

Fig. 6. Arduino Leonardo[15].

Breadboard: a breadboard is a widely used tool for designing and testing circuit [16] as shown in Fig. 7. For this project, the breadboard will be used to connect the ground (GND) pin from Arduino to its row.

Jumper cable: cable will be used to connect Arduino with sensors and relay. Latching push button switch: this button acts as Off/On maintained switch. We have chosen this button with a flashed light when pressed to provide more accessibility for visual impairments users. Button: normal button Off/On.

E. Educational Website

In order to provide an effective learning system, we had to understand the problem space, learn about Braille system and get more insight in order to be able to gain a better understanding of the system's requirements which basically targets the visual impaired children. For that, we visited Al-Nour Institute for the blind. During this visit, we conducted interviews with elementary school teachers teaching the three foundation courses. These courses teach the basis of Arabic Braille for students in their first study year. The main purpose of these interviews was to learn about the teaching strategies, in addition to getting their suggestions and recommendations on teaching Braille.

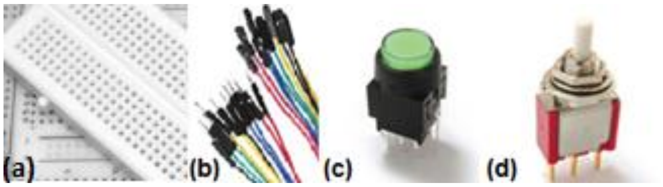


Fig. 7. Images of Some of the Hardware Parts used in Designing the Interactive Keypad : (a) Breadboard, (b) Jumper Caples (c) Latching Push Button Switch (d) Button [16].

Based on a well-established set of requirements, the educational website was designed to consist of 5 sub-modules:

a) About Braille cell: this module is an introductory lesson about the Braille cell providing an audible explanation of the structure and coding of Braille cell.

b) How to use the Braille Keypad: before going into learning Braille, learner needs to be introduced to the main interface of the system which is the speciallly design keypad to simulate a Braille cell and interact with the system. The system provides the learner, while sensing the keypad, with an audible explanation of the shape and structure of the keypad and how to use it, this includes the different buttons and the action of each.

c) Arabic Braille learning: This is the core module of the system with the main functions. The main user interface of the learning module includes three main functions:

- Lesson: the explanation lesson of the Arabic Braille character (letter or number).
- Practice: a practicing option that enables the child to try and write the character with Braille keypad without any constraints such as times or score.
- Re-learn: this is an option which gives the child the choice of replaying the lesson.

First, the learner should choose the letter/number lesson either by selecting a letter/number from the list or continue from the lesson he stopped on. The lesson will start showing a word that starts with the letter to be learned and saying the word and the letter, e.g. the word "أسد" starts with the letter "أ", the letter will be represented on screen by the Braille cell and will be explained using texts, images and audio materials. The same thing for numbers' lessons, e.g. three flowers will appear with the number "3". It will be represented on screen by the Braille cell and will be explained. After finishing the lesson, the practicing mode will appear to make sure that the learner has been understood the lesson or he/she can repeat the lesson again. In practice, the system will show and pronounce the letter/number. Then, the learner will be asked to answer using the Braille keypad. If the answer is correct, the system will ask the learner either to go to the next lesson or go back to the previous page. Else the learner has the choice to re-answer again or chooses re-learn option to repeat the lesson from the beginning. An illustration of this scenario is shown in Fig. 8.

d) Testing: an important part in any eduacational system is knowlwdge testing. In this module, the testing mechanism for letters and numbers is the same, where ten random letters or numbers will be chosen; depending on the selected learning (letters or numbers). The system will ask the learner to enter the matched Braille code of the displayed character using the Braille keypad as it is illustrated in Fig. 9. The learner has three attempts for incorrect answers, after the third incorrect answer, the system will show him the correct answer. Finally, total scores with the feedback will be displayed after finishing the test.

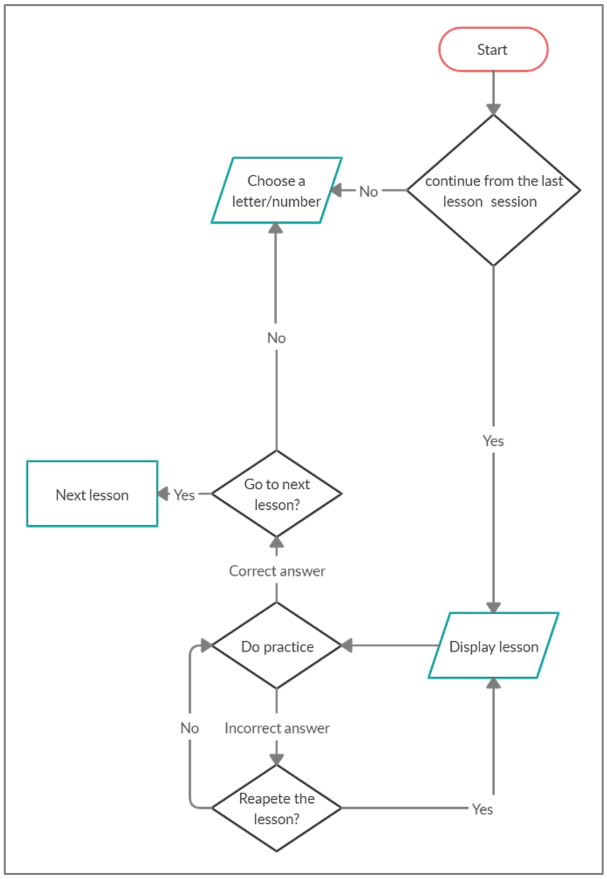
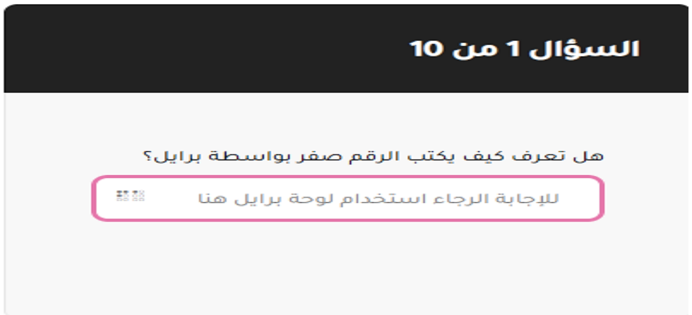


Fig. 8. Flow Chart of the Lesson.



السؤال 1 من 10

هل تعرف كيف يكتب الرقم صفر بواسطة برايل؟

للإجابة الرجاء استخدام لوحة برايل هنا

Fig. 9. A Testing Example (Testing Number 0).

e) Reporting: providing a progress repor is very usefull for the parents/teachers to monitor and guide the learner during the learning process. Once the learner exits from the system a complete readable report will show the progress, lessons history, practice and test results with the feedback. Those reports are compatible with screen readers.

V. System and Usability Testings

We conducted two separate evaluations for the developed system. First a system testing was conducted to ensure that the system meets its functional and nonfunctional requirements. The main objective of is to evaluate the sufficiency of functionality. This includes: unit testing, system testing, and quality testing.

Usability and accessibility tests are very important especially when developing a system for users with special needs. Thus, the second evaluation is a usability test conducted with blind children, and the accessibility test was performed as part of system testing. In the following subsections we discuss all tests performed, starting with unit testing followed by system testing and quality testing. We conclude with the usability test.

A. Unit Testing

Unit testing is verifying a particular component of the system and checking if it is working without any errors [17]. It tests a unit of the system individually, so it verifies that a specific piece of code performs as expected. Unit testing of the system functions was passed successfully.

B. System Testing

System testing is the testing to evaluate the whole web application's compliance with specified requirements such: number of broken links/URLs in the site, web accessibility standards and the compatibility with most used browsers [18]. In order to test our website, we used an automatic evaluation tool named SortSite [19]. The tool was released in 2007 and widely used by many since then.

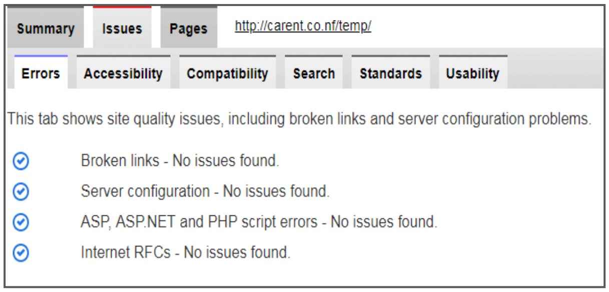
a) Links/URL Testing: the tool shows that the website has no missing images or broken links see Fig. 10.

b) Web Accessibility Testing: we test the website to ensure that it complies with accessibility best practice guidelines that make the web application accessible to people with disabilities. As a result, the website meets WCAG 2.0 [20] Level A conformance and Level AA with two issues.

c) Browser Compatibility Testing: the tool shows that the website is compatible with the major browsers, i.e. Internet Edge, Firefox, Safari, Opera, Chrome, and mobile browsers. However, users using old versions of Safari might face minor issue that will not prevent them to benefit from the website.

C. System Quality Testing

We evaluate the system to test whether our system meets the specified non-functional requirements. We used SQuORE analytics tool to perform the quality testing. SQuORE is an intelligence tool that has been founded in 2010 by a group of software engineering experts. SQuORE supports a lot of different programming languages so, it has the ability to assess a project that has been developed under multi-languages. This tool has a friendly interface that can be easily deal with it and understand the analysis result [21].



This tab shows site quality issues, including broken links and server configuration problems.

✓ Broken links - No issues found.

✓ Server configuration - No issues found.

✓ ASP, ASP.NET and PHP script errors - No issues found.

✓ Internet RFCs - No issues found.

Fig. 10. SortSite Results - Links/URL Testing.

"SQuORE business intelligence tool introduces a novel decision-based approach to software projects quality assessment by providing a more reliable, more intuitive, and more context-aware view on quality [22]." There are huge of software quality measurements such: cyclomatic complexity for control flow, Comment Density, etc. In SQuORE tool proposes several models and standards such: ISO 9126 and the Automotive HIS [22]. HIS (Hersteller Initiative Software) specifies a fundamental set of Metrics to be used in the evaluation of software and quality assessment. Each of these defined metrics is studied and reported. All violations of the agreed boundary limits at the function level have to be justified [23]. The results show that all the specified strategies are within the range. Moreover, the fault tolerance has the highest value that may affect the system reliability. The lowest page quality is the test page which has an "E" key performance whereas five pages of "D" key performance and three pages of "A" key performance which is the highest quality indicator.

D. Usability Testing

We conducted a usability study to determine how usable, satisfactory, efficient, and effective the web site is for the users, identify any usability problems and finding ways of improving the usability of the system. A total of three blind children participated in this study. To demonstrate this study, this testing is divided into the following: 1) test environment set-up. 2) participants. 3) evaluation methods of. 4) assessment measures. 5) tools 6) results. 7) findings and recommendations.

a) Enviroment setup: the evaluation was conducted in a quiet and closed meeting room at Kafeef Organization. Our laptop as a local host for our system is placed on a meeting desk. The visually impaired child is setting on a chair facing the laptop while his hand is on the Braille keypad. The child mother was present in the room during the evaluation for assistance. Fig. 11 shows the environment set-up with a visually impaired child.

b) Participants: a sample of three children were recruited in the study as shown in Table II. Two participants were male and one female. All were (4-9) aged. 2 out of 3 were totally blind and only one was partial visually impaired. Moreover, none of them has advanced background of the Arabic Braille letters and numbers. In addition, none of the participants is familiar with a screen reader or have been used educational sites or have used technology and Internet before except for participant (P2) who used to access Internet and educational site using iPad. All participants looking to use an Arabic Braille learning system and think that it will help them to learn.

This study was approved by association of the blind (Kafeef) Ethics Committee. Parents of the participants were informed that the test outcomes will be used for research purposes only and were assured that privacy and personal identity information of all participants will be protected.

c) Methods: two of the industry-standard usability methods were applied in this evaluation: subject evaluation via pre-test and post-test interviews and the observation method during the testing session [24].

Once the parents of the children participants accept and sign the consent form. A pre-test interview questions are conducted. The goal of the pre-test interview is to gather children's demographic information such as gender and age, their background of Arabic Braille, technical experience such as using the Internet and educational web sites. During the test session, we have selected the observation usability evaluation method to watch the participants while they are using the system and take notes. Participants were asked to complete several tasks in the system. These tasks include:

- Task 1: Choose the Arabic letters learning option.
- Task 2: Go to the "ب" lesson.
- Task 3: Perform practice of the chosen lesson and answer the question.
- Task 4: Trying to go to the Arabic letters' test and answer the first question using the Braille keypad.
- Task 5: Trying to exit the program and see the report of current session.

TABLE. II. PRE-TEST INTERVIEW COLLECTED INFORMATION

Participant No.	P1	P2	P3
Gender	Male	Male	Female
Age	4	9	8
Visually impaired kind	Total (Blind)	Partially	Total (Blind)
Arabic Braille background	No	Junior	junior
Familiar with Screen reader	No	No	No
Educational sites used before	No	Yes	No
Technology and Internet usage	Not using	Using iPad weakly	Not using
looking to use an Arabic Braille learning system	Yes	Yes	Yes
If there is an educational site for Arab Braille, do you think that it will help you to learn	Yes	Yes	Yes
Would like to use educational sites for the visually impaired	Yes	Yes	Yes

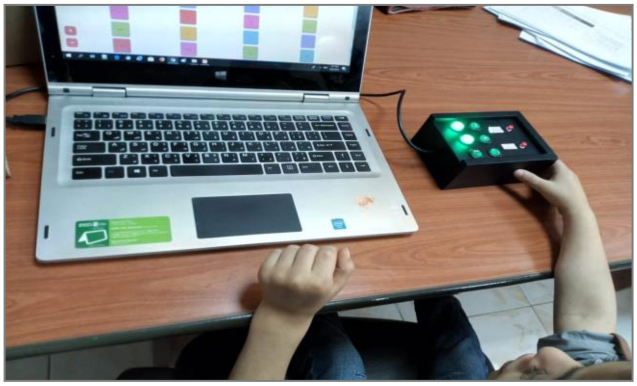


Fig. 11. The Environment Set-up with a Visually Impaired Child.

Once the participants complete the tasks, they interviewed with help of their mothers to get their overall impression of the system after using it. Most of the questions contained a Likert scale to assess the user experience of the system, satisfaction level of the system and if issues were countered. Moreover, we asked questions about whether the children would like to use the web site in the future and if they have any recommendations to enhance the web site or the Braille keypad.

d) Assesment Measures: In order to measure the usability of our system, we concerned in conducting the usability test in terms of efficiency, effectiveness and satisfaction. Table III describes each term and its measurement data.

e) Tools: Two software tools were used to assist the completion of this evaluation. These tools are:

- Morae: a usability software tool that allows recording the user's interactions on the screen. The software was installed on the laptop, which allows us to have a portable usability lab [25].
- Excel: was used to analyze the observed data and draw charts.

f) Results: All the participants completed the given tasks successfully, except one participant (P1) who could not complete Tasks 2 and 4. From our observation, we found that this participant doesn't know what laptop means or even website and he doesn't know any Arabic letters or numbers. He may need extra preparation sessions to introduce him to the technology and to be familiar with it before performing the test. We calculated the efficiency of the web site by measuring the task completion time see Table IV. The participants were varying in the time spent to complete these tasks. Fig. 12 shows the time spent on each task in seconds for each participant. It reveals that the best result was achieved by the participant (P2) who spent the shortest time in all tasks compared to the other participants. His progress performance was due to the fact that he used to use the technology in addition to his experience in using educational sites. Moreover, we can notice that as demonstrated in Fig. 13 Task 4 took the longest time. We think that this is due to its order of other tasks, it needs to go back twice from the previous task page to achieve it.

Other performance metrics to measure is the effectiveness of the system which was calculated by the number of user errors and number of assistances required to complete each task. Regard the number of user errors for each task only participant (P1) commit an error in two tasks (Task 2 and 4). In fact, the same error was because the participant entered the url link settings by mistake and ended the task. However, Task 2 needed the greatest number of assistances to complete than other tasks by average of 56%. all the children found the system is motivative and have enjoyed the learning by the web site. Moreover, they agreed that the Braille keypad was easy. Two of them stated that the web site was easy to use, and they felt successful but with tired, and we think that feeling "tired" maybe due to the new experience and being anxious to do it successfully. Additionally, children were asked to give their suggestions and recommendations to improve the system. Most of them gave the same suggestion, which is disable the url to be unreadable by the screen reader.

g) Findings and recommendations: the recommendations we get are driven by the participant behaviors, verbal feedbacks, comments and suggestions that we ask participants after using the system. These recommendations are either for the web site or for the Braille keypad and range from highly important to less important, depend on the number of users that recommend or faced the same issues. The following TABLE I summarizes the recommendations.

TABLE. III. THE MEASUREMENT DATA FOR THE USABILITY TESTING

	Metrics	Measurement
Quantitative measures	Efficiency	-Task Completion Time
	Effectiveness	-Number of user errors -Number of assistances required to complete each task
Qualitative measures	Satisfaction	-Participant's opinion and their satisfaction of the system

TABLE. IV. PERFORMANCE RESULTS

User	Task 1	Task 2	Task 3	Task 4	Task 5
P 1	46.00	112.00	30.00	323.00	38.00
P 2	19.00	22.00	12.00	27.00	10.00
P 3	33.00	51.00	18.00	65.00	18.00
Average	32.67	61.67	20.00	138.33	22.00

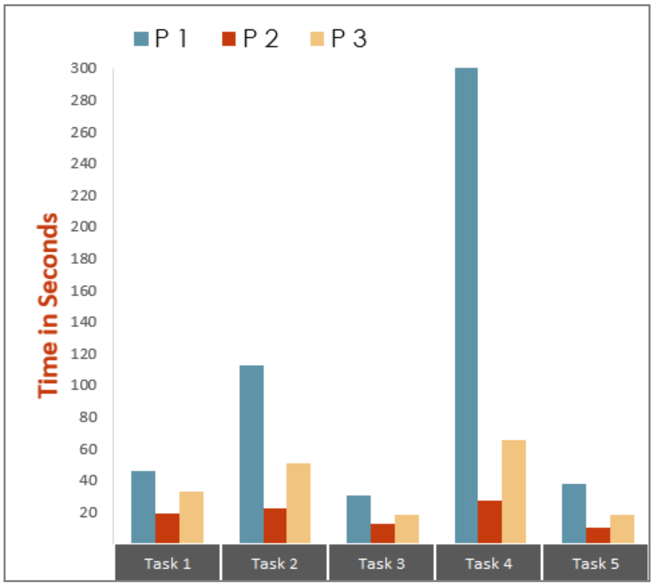


Fig. 12. Time Spent on Each Task by Seconds.

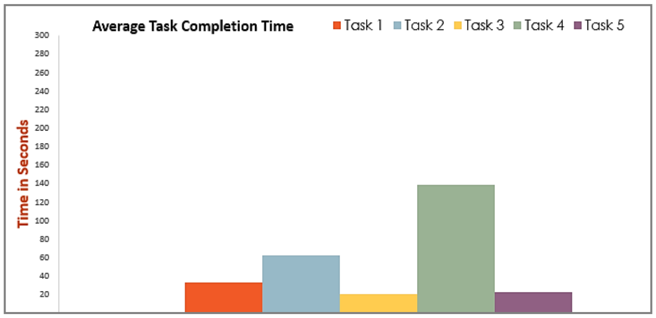


Fig. 13. Average Task Completion Time.

TABLE. V. TABLE I. FINDING AND RECOMMENDATIONS

#	Recommendation and Suggestions	Location	Priority
1	The URL of the page should be disabled or not readable by the screen reader because it interrupts the movements between site elements. Moreover, its in English languages so its not understandable by users.	website	High
2	The user should practice many times to be familiar with the screen reader and learn himself without any intervention from his parents	website	High
3	Should control the buttons lighting in the Braille keypad because this is annoying for visually impaired users. on other hand, some visually impaired found that the high light helps them to place the Braille dots.	Braille Keypad	High
4	Providing a button to repeat what the screen reader said instead of going through elements again	website	Moderate
5	Providing a choice to change the buttons color	Braille Keypad	Low
6	Putting two Braille cells instead of one which make it more feasible in learning the Arabic numbers	Braille Keypad	Low
7	Provide a chargeable Braille keypad that connected wireless with any portable device	Braille Keypad	Low

VI. FUTURE WORK

Future work efforts will focus on adding new features and higher level lessons such as including diacritics marks and Arabic words. We also plan to implement the recommendations concluded from the results of usability testing. We hope to enhance this system by improving the designed keypad and incorporate more than one cell to be able to spell words not only code letters and numbers.

VII. CONCLUSION

The proposed system facilitates the Braille learning process. It is designed to help the blind and visually impaired in learning Arabic Braille (letters and numbers) using a specially designed interactive tactile Braille keypad. The proposed system is not limited to the blind or visually impaired, but it can also be used by sighted people to learn Braille which is important in families having visually impaired children. A usability test was conducted and results showed improvements in the learning process and the use of interactive tactile Braille keypad was highly accepted by blind children who participated in the test. We hope that this work will be a valuable contribution to our community and especially to the visually impaired.

ACKNOWLEDGMENT

Authors would like to thank Al-Nour Institute for the Blinds for their support and cooperation in this study. This research project was supported by a grant from the "Research Center of the Female Scientific and Medical Colleges", Deanship of Scientific Research, King Saud University.

REFERENCES

[1] J. Jiménez, J. Olea, J. Torres, I. Alonso, D. Harder, and K. Fischer, "Biography of Louis Braille and Invention of the Braille Alphabet," Surv. Ophthalmol., vol. 54, no. 1, pp. 142–149, 2009.

[2] E. S.-H. and R. M. Joshi, Handbook of Arabic Literacy, 2014th ed., vol. 9. Dordrecht: Springer Netherlands, 2014.

[3] E. Britan-, " Bleaching → Semantic Bleaching Bornu Arabic → Subsaharan Arabic," vol. 1960, pp. 35 – 36, 2003.

[4] Kafeef, "ك ف يف," 2018. [Online]. Available: http://www.kafeef.org. [Accessed: 27-Sep-2018].

[5] D. S. Awang Damit, A. I. Che Ani, A. I. Muhamad, M. H. Abbas, and F. Z. Ali, "Dual braille code translator: Basic education tool for visually impaired children," I4CT 2014 - 1st Int. Conf. Comput. Commun. Control Technol. Proc., no. I4ct, pp. 399–402, 2014.

[6] T.-J. Yang, W.-A. Chen, Y.-L. Chu, Z.-X. You, and C.-H. Chou, "Tactile Braille learning system to assist visual impaired users to learn Taiwanese Braille," SIGGRAPH Asia 2017 Posters - SA '17, pp. 1–2, 2017.

[7] H. Nicolau, J. Guerreiro, T. Guerreiro, and L. Carriço, "UbiBraille: designing and evaluating a vibrotactile Braille-reading device," in Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, 2013, p. 23.

[8] J. Tang, "Using ontology and RFID to develop a new Chinese Braille learning platform for blind students," Expert Syst. Appl., vol. 40, no. 8, pp. 2817–2827, 2013.

[9] H. Petrie and N. Bevan, The evaluation of accessibility, usability, and user experience, no. June 2009. 2009.

[10] E. Bergman and E. Johnson, "Towards {Accessible} {Human}-{Computer} {Interaction}," Adv. human-computer Interact., vol. 5, no. 1993, pp. 87–114, 1995.

[11] S. Schmutz, A. Sonderegger, and J. Sauer, "Implementing Recommendations from Web Accessibility Guidelines," Hum. Factors, vol. 58, no. 4, pp. 611–629, 2016.

[12] A. Kavcic, "Software Accessibility: Recommendations and Guidelines," EUROCON 2005 - Int. Conf. Comput. as a Tool"", vol. 2, pp. 1024–1027, 2005.

[13] R. Lopez, "Matuto, Magbasa, Maglaro: Learning to Read Braille Through Play," 2018.

[14] N. L. C. Gomez et al., "SBK: Smart Braille Keyboard for Learning Braille Literacy in Blind or Visually Impaired People," Proc. 8th Lat. Am. Conf. Human-Computer Interact., no. Figure 1, pp. 26:1--26:4, 2017.

[15] A. Kurniawan, Getting Started with Matlab Simulink and Arduino. PE Press, 2013.

[16] J. Boxall, Arduino workshop: A Hands-On introduction with 65 projects. No Starch Press, 2013.

[17] G. A. Di Lucca, "Testing web-based applications: The state of the art and future trends," Proc. - Int. Comput. Softw. Appl. Conf., vol. 2, no. May, p. 65, 2005.

[18] Paul C. Jorgensen, Software Testing A Craftsman's Approach, vol. 47, no. (2). 2014.

[19] PowerMapper, "One Click Website Testing," 2019. [Online]. Available: https://www.powermapper.com/. [Accessed: 15-Apr-2019].

[20] w3c, “Web Content Accessibility Guidelines (WCAG) 2.0. W3.org.,” 2019. [Online]. Available: https://www.w3.org/TR/WCAG20/. [Accessed: 15-Apr-2019].

[21] SQUORING, “squoring Providing the best Business Intelligence tool for software and systems projects monitoring,” 2019. [Online]. Available: https://www.squoring.com/en/societe/a-propos-de/. [Accessed: 21-Apr-2019].

[22] B. Hervé, “Software Qualimetry at Schneider Electric: a field background,” pp. 1–10, 2010.

[23] A. Korn, E. Haunschild, B. Kalusche, E. Sax, and M. Gmbh, “HIS AK Softwaretest Inhaltverzeichnis,” pp. 1–8, 2008.

[24] G. B. Segura, Densidad de masa ósea en la espondilitis anquilosante, vol. 17, no. 2. 2001.

[25] Morae, “usability testing with Morae,” 2019. [Online]. Available: https://www.techsmith.com/morae.html. [Accessed: 25-Apr-2019].

Very High-Performance Echo Canceller for Digital Terrestrial Television in Single Frequency Network

El Miloud Ar-Reyouchi[1]
Telecommunications Computer Science, Abdelmalek Essaadi University and SNRT
Tetouan, Rabat, Morocco

Yousra Lamrani[2]
Kamal Ghoumid[3]
Electronic informatics Telecommunications, ENSAO
Mohammed I University
Oujda, Morocco

Salma Rattal[4]
Electrical engineering
FSTM Hassan II University
Casablanca, Morocco

***Abstract*—The principal aim of this paper is to cancel out the natural and man-made echoes in single-frequency networks (SFN). The challenge is to detect and remove feedback echoes and enhance the intelligibility of the essential parameters in SFN of digital terrestrial television broadcasting (DTTB) transmitter systems, especially the Modulation Error Ratio (MER), with optimizing coverage areas. We suggest a Digital Video Broadcasting (DVB) gap filler (GF) with two types of echo cancelling: Digital Adaptive Equalizer (DAE) and Doppler Enhanced Echo Canceller (DEEC). The proposed GF outperforms standard GF (SGF), finite impulse response filter (FIR GF), and adaptive GF (AGF) techniques by 33%, 17%, and 13%, respectively. Furthermore, the obtained MER makes the proposed GF (PGF) ideal for operating in SFN using Coded Orthogonal Frequency Division Multiplex (COFDM) technique.**

Keywords*—Gap filler; Digital Adaptive Equalizer (DAE); Doppler Enhanced Echo Canceller (DEEC); Single-Frequency Networks (SFN); Coded Orthogonal Frequency Division Multiplex (COFDM)

I. Introduction

In DTTB, the terrestrial digital TV, and gap-filler transmitters broadcast the same TV service using the single frequency. The Coded Orthogonal Frequency Division Multiplex (COFDM) techniques [1],[2] allow TV transmitters to create an SFN, which is widely used for economizing frequency [3]. The presence of a GI (guard interval) [4], in COFDM, gives it excellent robustness to lute against echoes derived from multipath interference. The coincidence of several TV broadcasts produces different types of echoes, whether artificial or natural. The echoes may appear when receiving simultaneously [5] the same signal from several transmitters. Another reason that can cause echoes is the reflection of the signal on large objects such as mountains.

To improve wide-area performance applying SFN synchronization, the studies in [6],[7] have treated various network architectures depending on different parameters like COFDM technology, the selected scattering power, and a perfect adaptation measurement[8] with a suitable antenna[9] for the reception of DVB.

In SFN, the practical network planning approach of the transmitter's location and antenna orientation are always necessary to reduce the interference. Echoes can also be reduced by decreasing the number of transmitters per local area. The selection of a transmitter with multiple antennas to the transmission and reception MIMO (multiple inputs, multiple outputs) [10], [11] heights they can promote smooth operation of the self-interference cancellation technique.

In the traditional Multi-Frequency Networks (MFN) technology, each transmitter uses a different frequency to avoid undue interference between its neighbors. Interference can be assimilated to an echo [12] if the content of the signal is the same on both transmitters.

In SFN, the echoes and interferences are avoided by COFDM modulation on which they are based. Since GF transmits and receives on different frequencies, it can also operate on the identical frequency with low-medium power DVB transmitters. This research approach is chosen to provide additional converge on those blind spots not covered by the high power transmitter using efficient proposed GF (PGF) operating in SFN using the same channel. The strength of the PGF is reflected in the ability to annul the feedback echoes effectively, increasing the frequency efficiency. The PGF is based on two echo-canceling DAE and DEEC. The DEEC is based on the least-mean-square (LMS) adaptive filter [13] using the LMS algorithm [14], the DEEC provides a filter that models the channel between the transmitter (Tx) and receiver (Rx) antennas. With DAE [15], the performance of a DTTB for Quadrature amplitude modulation (QAM) can be ameliorated considerably. Many studies have explored different aspects of the behavioral feedback path, such as [16]. In [17], the authors analyze the required overhead to ensure a target average signal-to-interference-plus-noise ratio (SINR) using Time-division multiple access (TDMA) of each node. In [18], the interference cancellation (IC) reduces interference effect, and the overlap ratio could be enhanced while maintaining a high quality of service (QoS). The echo cancellation method has been studied in [19],[20], and analyzed in a deferent domain such as in [21] for voice over IP (VoIP) communication acoustic echo.

In this work, broadcast technology is adopted (one point to all other points, the authors present two echo cancellation methods as a critical technique to solve the echo and interference problem in the SFN environment using DAE and DEEC like an echo cancellation method. Furthermore, the ideal proposed technique should be compared with other research results, such as [22]. This paper does not only ensure the safe and robust operation of the repeater against the dynamics of the environment [23] but also can suppress Doppler Effect and avoid the interferences and echoes to the received signal from the transmitter in its overall coverage area. Overall, it has to be noted that the applied Echo cancellation [24] techniques not only help to improve isolation [25] but also gives perfect emission-reception isolation. Our goal is to get satisfaction levels of 80-85% of a parameter, which have a direct relationship with the echoes suppression, especially the MER. The MER is a paramount measure, and DVB widely requires it for OFDM signal performance in SFN, it is analyzed in [26] as received signal power in an actual digital television transmitter. MER is affected by the delay of received signals [27] from the SFN network, and the result of MER showed the quality of service (QoS) at the receiver.

The remaining of this paper is classified as follows. In the second section, we briefly introduce a general system Performance. The third section exposes and discusses the problem with the remedy. The fourth section presents and discovers the echo canceller utility and service, such as DAE & DEEC. The fifth section deals with the summary of the experimental results obtained. Finally, the sixth section presents conclusions.

II. General Performance

A. General Performance

The transmission path is known as “Rician channel,” which takes into consideration the effect of multipath signals, noise, and the dominating direct signal path between Tx and Rx.

Temporal delay shall be estimated according to the following formula: $\lambda = c / f = cT \Rightarrow \Delta d = c\tau$ where λ is the wavelength, c is the speed of light, and f is frequency. The offset delay is more important for remote transmitters (200 µs for 60 km). Consequently, at the receiver side, due to the presence of the multiple electromagnetic paths, more than one signal is received by a specific instance, and each one of them reaches at different times $\tau(t)$ with different energy strengths $\beta(t)$ and different angles $\alpha(t)$.

Let $S(t)$ be the transmitted complex signal having a carrier frequency f_0 modulated by a baseband complex signal $x(t)$, and it can be written as:

$$S(\mathrm{t}) = x(t).e^{j2\pi f_0 t} \tag{1}$$

When reflected signals are added to a direct signal, the quality of the reception becomes worse. Multipath-propagation interference varies the signal intensity and produces intersymbol interference (ISI). These cases are modeled in the channel, which is defined as a Rice channel. In this type of channel, the Gaussian channel and its characteristics also be present. The received signal suffers a multipath channel with distinct waves. With the additive white Gaussian noise omitted, it can be expressed as:

$$S^{'}(\mathrm{t}) = \sum_{i=1}^{m} a_i(t) S(t-\tau_i(t)) \tag{2}$$

which give the following equations:

$$S^{'}(\mathrm{t}) = \sum_{i=1}^{m} a_i(t).e^{-j2\pi f_0 \tau_i(t)}.x(t-\tau_i(t)).e^{j2\pi f_0 t} \tag{3}$$

where $a_i(\mathrm{t}) = \beta_i(\mathrm{t}).e^{\alpha_i(\mathrm{t})}$ and $\alpha_i(\mathrm{t})$ designate the attenuation and the delay of the i-th path. The received signal $S^{'}(\mathrm{t})$ can be rewritten as the baseband signal $y(t)$.

$$y(\mathrm{t}) = \sum_{i}^{m} c_i(t)\, x(t-\tau_i(t)) \tag{4}$$

Where

$$c_i(\mathrm{t}) = a_i(t) e^{-j2\pi f_0 \tau_i(t)} \tag{5}$$

The Doppler shift affects the attenuation $a_i(t)$ periodically and is relatively small when compared with the carrier frequency f_0 .

A simulation of the power requirements for the terrestrial DVB-T standard applied the following mathematical model to describe the channels with echoes. A certain number of echoes can be taken into consideration. In the Rician channel model, the Gaussian channel exists as well as its characteristics. The signal is completed with the reflection of signals in different ways. The received signals, $y(t)$ according to [28] becomes:

$$y(t) = \frac{\rho_0 x(t) + \sum_{i=1}^{m} \rho_i e^{-j2\pi\alpha_i} x(t-\tau_i)}{\sqrt{\sum_{i=0}^{m} \rho^2{}_i}} \tag{6}$$

The mathematical equation, which described the influence of Rayleigh channel on the signal:

$$y(t) = \frac{\sum_{i=1}^{m} \rho_i e^{-j2\pi\alpha_i} x(t-\tau_i)}{\sqrt{\sum_{i=0}^{m} \rho^2{}_i}} \tag{7}$$

where ρ_0 is the attenuation in the line of sight of the transmitter, ρ_i is the attenuation in echo path i , θ_i is the phase rotation in the echo part i , and τ_i is the relative delay time in the echo part. The Rice Factor K designates the signal ratio by way of the line-of-sight broadcasting path to the sum in all echo ways (8).

$$K = \frac{\rho_0}{\sum_{i=1}^{m} \rho^2{}_i} \tag{8}$$

B. MER Performance Analysis

Intermodulation (IMD), C/N, and RF Level that can be used to evaluate the quality of service (QoS) for DVB-H/T/T2 reception. The most useful one is the MER, which is used to quantify the performance of a digital radio (or digital TV) transmitter or receiver in communication, and it provides a clear and fast overview of the echoes and overlaps measurements; therefore, we highlight the MER.

In the OFDM system, a random bit sequence is generated, and then the bits are mapped into 64-QAM symbols. The I and Q values of this sequence are stored as the $\tilde{I}_j$ and $\tilde{Q}_j$ array. The $\tilde{I}_j$ and $\tilde{Q}_j$ array is next used to form the OFDM frequency domain signal, in which the resulting OFDM symbols I_j, Q_j are mapped.

MER measurement of OFDM in SFN is used to measure the modulation quality [29], and practically its value is enough (according to measurements made on SFN) to judge the absence or the presence of the echoes in SFN.

MER over several symbols N (is the number of points in a measurement) is defined as:

$$MER = \frac{\sum_{j=1}^{N}(\tilde{I}_j^2 + Q_J^2)}{\sum_{j=1}^{N}\left[(I_j - \tilde{I}_j)^2 + (\mathrm{Q}_j - Q_j)^2\right]} \tag{9}$$

where: $I_j, Q_j, \tilde{I}_j, \tilde{Q}_j$ are the ideal and quadrature components of the j-th measured/ referenced OFDM signal, respectively.

C. Intermodulation Performance Analysis

IMD is the most useful signal analyzer measurements and widely used and is quite an appealing metric of linearity, which is very important to prevent it, for radiofrequency. In a Digital Terrestrial TV, IMD is a crucial measurement caused by nonlinearities effects in the DVB system. IMD measurement can be particularly troublesome for high-frequency amplifiers for radio communications; also can automatically detect the fundamental and third-order distortion that forms a principal limit to the circuit linearity. The introduction of third-order intermodulation is due to the non-linear characteristic.

III. Problematic with the Proposed Remedy Solution

A. Problematic

The GF is usually installed in high and isolated stations, the transmitting antenna designed to cover the shadow area optimally. The exemplary GF installation problem is the unavoidable RF coupling between the broadcast signal from the antenna patterns of broadcasting stations and the signal picked up by the receiving antenna.

If the coupling level is too high, this turns into a risky and challenging situation that can easily destruct the equipment hardware due to the positive return path. In the case where the GF is located within an SFN, The problem becomes even more complicated. To remedy this problem, we use a Digital Echo Canceller in the digital processing module of a GF.

In the SFN overlap area, the weaker of the received signals from multipath directions is deemed considered as an echo. Thanks to the global positioning system (GPS), the synchronization of the transmitters of the broadcasting centers allows echoes to be placed precisely at guard intervals (GI) which length depends on the duration of the echoes.

The problem worsens considerably, the COFDM solution, however, does not readily apply in domains where the user density is low, and impediments such as distance and terrain create challenging obstacles to conventional approaches to the DTTB network. The overview of DVB standards, namely, DVB-T/H/T2, was not able to solve the echoes squarely and overlaps phenomenon using COFDM techniques. From a physical point of view, the Overlap effects (between the antennas) are inevitable.

B. Remedy

For good isolation between the Tx and Rx antennas, the solution is a digital echo cancellation based on a software solution of an echo canceller GFs that can resolve the most challenging echo conditions. Therefore, to limit isolation between the Rx and Tx antennas and suppress echoes with interference in SFN, the PGF can optionally incorporate two echoes cancelling solutions DAE and DEEC.

DAE is a powerful and useful tool to remove echoes whose gain margin (GM) level can reach up -10 dB Moreover, especially in SFN operation, and in severe reception conditions, DAE can to remove multipath propagation impacts by equalizing the amplitude distortions of the GF received distorted signal spectrum that is created by very near echoes.

The DEEC is also able to suppress echoes but with considerable GM where the echo levels exceed the input signal by 24 dB and can often remove Doppler Effect. DEEC is integrated to avoid multipath, as well as Doppler echoes. The GF cancels feedback echoes with a GM up to 24dB, and three gaps echo, giving a selective cancellation that would go until 37,6µs. Fig. 1 shows a correct example of the positioning of the RX and TX antennas.

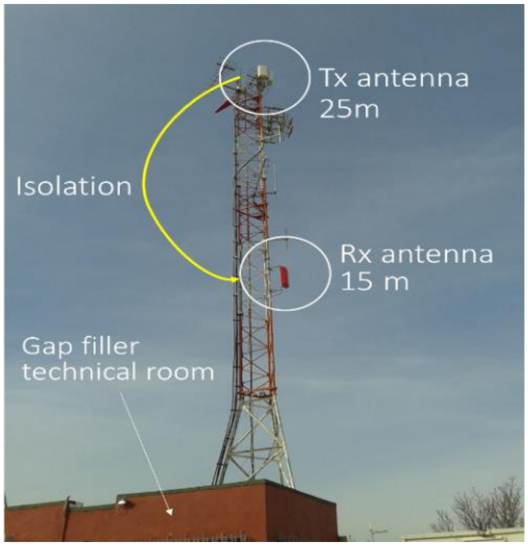


Fig. 1. Correct Positioning of the TX and RX Antennas Seeking Maximum Isolation.

We select the Rx antenna position to at least 15 meters away from the TX antenna (if possible). We try to target the Tx and Rx antennas for opposite sides.

Fig. 2 shows a simplified system Topology containing a high-power transmitter and GF as well as the Field Test Receiver Measurements: Covered by GF interfere with the main transmitter at which the outdoor measurements are performed.

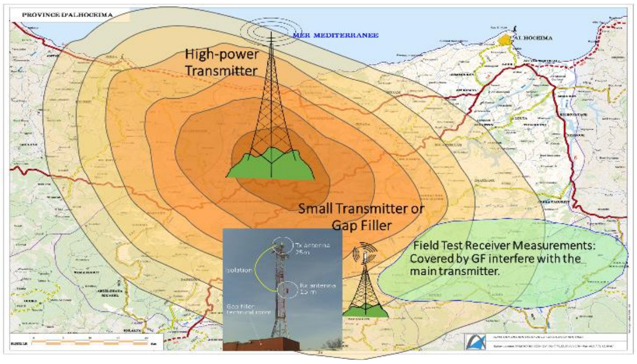


Fig. 2. Topology for SFN DVB-T/T2 Network Measurements.

IV. Echo Canceller: Functional Description

Fig. 3 shows the block diagram representation of the digital GF internal composition. The system carries out a down-conversion of the receiving RF signal to an intermediate frequency (IF), filters the resulting signal, and reconverts it back to RF that it is amplified before rebroadcasting starts.

When the echo signal level 10dB inferior to the principal signal, the standard GF can work without any echo canceller in SFN. In the case where reception conditions are particularly difficult in SFN, the need for DAE for an echo signal level up to 15dB or DEEC for echo signal level higher than the fundamental signal is extremely important.

A. DEEC: Technical Representation

An adaptive LMS filter is the indispensable content of DEE that can detect the eventual feedback echoes and echoes of the SFN then can usually delete them all or part of the output. The DEEC high-performance gives it great robustness against echoes in the most challenging reception conditions. DEEC can remove high return echo levels, suppress echoes with greater GM, and cancel Doppler Effect improving MER performance:

Output MER is elevate to 27dB for a 20dB GM.

Output MER is superior to 24dB for a 24dB GM.

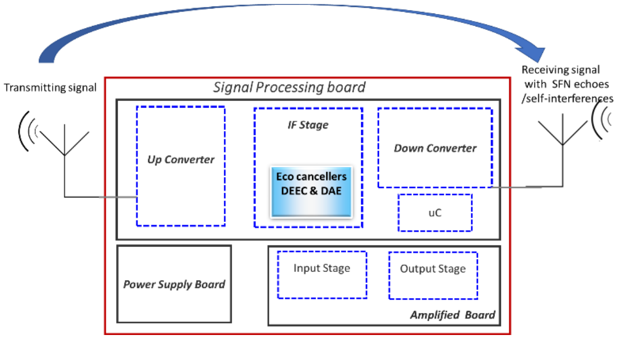


Fig. 3. Simplified Gap Filler Block Diagram.

The cancellation coverage depends on three temporal gaps, essentially 6µs large each, appropriate in continuous mode, and they entirely fully include a range from 0,5µs to 18,5µs. The three gaps can move individually, and they can be extended to 37,6µs. The settings and measuring values of Fig. 4 and 5 are shown in Table I.

The DEEC graphs, of the input-output signal selection, are shown in Fig. 4 and 5, respectively.

DEEC can also suppress any multipath and even Doppler echo in a mobile cancellation gap going up to 37, 6 µs. The echo cancellation level (20dB) with DEEC is presented in Fig. 4(a) and 4(b). The 20dB level is higher than the signal using DEEC and with the best output MER (27dB).

The return echoes can reach high levels at the input signal. In that case, to obtain an active suppression of this echo of return, its corresponding cancels gap use an external reference. This cancellation gap system should always be active, which is shown in Fig. 6.

TABLE. I. Setting and Measuring the Value of Fig. 4 and 5

Parameters	**Fig. 4**	**Fig. 5**
TV/Radio Standard:	OFDM DVB-T/H	OFDM DVB-T/H
Channel	32	32
UHF RF	562 Mhz	562 Mhz
Channel Bandwidth:	8 Mhz	8 Mhz
Signal level	-36.50 dBm	-6.50 dBm
Attenuation	0dB	31 dB
BER	0.0 e-8	0.0 e-8
MER	28.1 dB	28.7dB
DEMOD	MPEG	MPEG

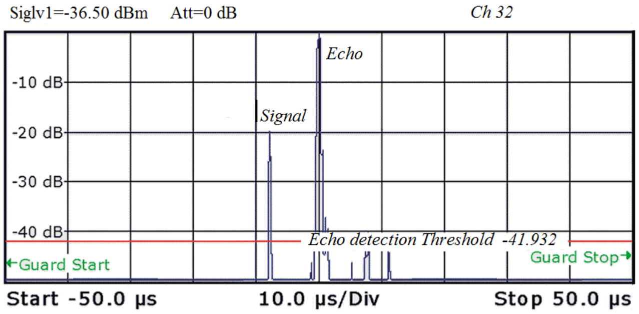


Fig. 4. DEEC Echo Canceller: Input Signal Spectrum.

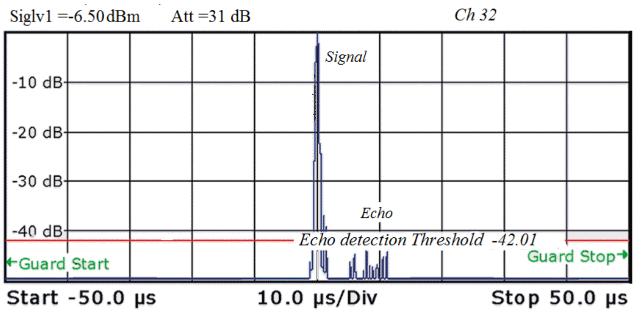


Fig. 5. DEEC Echo Canceller: Output Signal Spectrum.

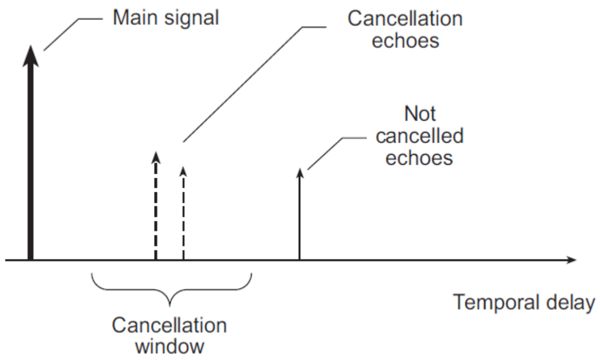


Fig. 6. Cancelation Gaps.

Fig. 7 shows the configurations of the most common DEEC gap positions, which cover a wide variety of cancellation scenarios.

There are three configurable cancellation gaps in the DEEC, namely:

Continuous cancellation: a continuous gap of 18µs

Selective cancellation: up to 37,6µs

Cancel any echo present in the cancellation gap (multipath propagation).

The following are the user-configurable parameters for the canceller:

Gap N° 1: Defines the cancellation velocity of gap N°1.

Gap N°2: Defines the cancellation velocity of gap N° 2.

Gap N°3: Defines the cancellation velocity of gap N° 3.

The cancellation velocity is directly related to the ability to track and cancel echoes with fast frequency and amplitude changes as present on Doppler, Rice, or Rayleigh channels.

However, increasing the speed of a cancellation gap creates a penalization on the output signal MER. The users should examine the different speeds to find a compromise between cancellation performance and output signal MER.

Gap N°1: Activates/deactivates the cancellation gap N° 1.

Gap N°3: Activates/deactivates the cancellation gap N° 3.

Gap N° 1 and Gap N° 3 are designed to suppress SFN echoes. These kinds of echoes may not appear/be relevant in every cancellation scenario, so these gaps can be switched off to improve the output signal MER, as explained above.

However, gap Nº2 is designed to challenge feedback echoes (which can be higher than the fundamental input signal), so it must continuously be activated to avoid feedback amplification that may harm and damage the equipment.

Gap Nº1: Defines the cancellation starting point of the gap Nº1. The available values are 0.5µs - 31.6µs; the default value is 12.0µs.

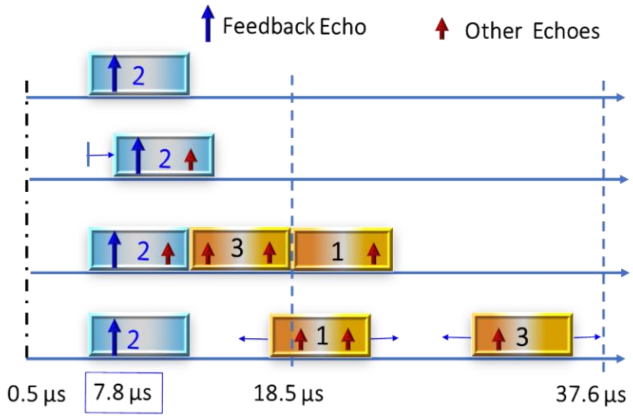


Fig. 7. Possible Cancellation.

Gap Nº2: Represents the cancellation starting point of the gap Nº 2. The available values for setting up the gap Nº2 position are 5.0µs - 36.1µs; the default value is 6.0µs.

Gap N°3: Defines the cancellation beginning point of the gap Nº 3. The available values are 0.5µs - 31.6µs; the default value is 18.0µs.

Each of these gaps has a time duration of almost 6µs in an 8MHz bandwidth operation.

B. DAE: Technical Representation

The DAE echo canceller is designed to effectively moderate echo conditions guaranteeing a GM level of support 15dB GM and equalizing spectrum for robust feature extraction for the more robust fight against echoes. Furthermore, DEA is capable of removing any echo caused by the presence of coupling between the broadcasting, receiving antenna, more precisely, the input signal whose delay varies between zero and 8µs. DAE circuit can also rectify the distortions in the GF input signal amplitude, which are within the cancellation gap, caused by multipath propagation. The settings and measuring values of Fig. 8 and 9 are shown in Table II.

The DAE graphs of the input-output signal selection are shown in Fig. 8 and 9, respectively.

TABLE. II. SETTING AND MEASURING THE VALUE OF FIG. 8 AND 9

Parameters	**Fig. 8**	**Fig. 9**
TV/Radio Standard	OFDM DVB-T/H	OFDM DVB-T/H
Ch	32	32
UHF RF	562 Mhz	562 Mhz
Channel Bandwidth:	8 Mhz	8 Mhz
Signal level	-40 dBm	16 dBm
Attenuation	0dB	20 dB
BER	0.0 e-8	0.0 e-8
MER	28.1 dB	28.7dB
DEMOD	MPEG	MPEG

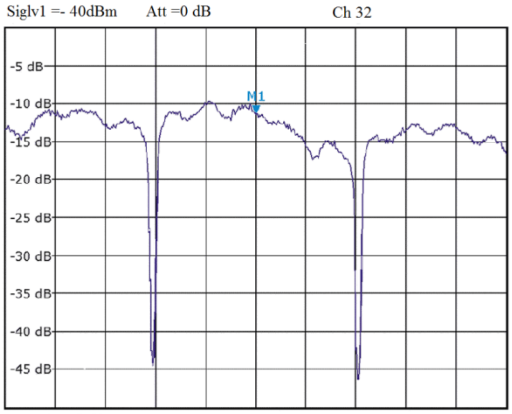


Fig. 8. Cancellation Example of an Echo up to 10 dB with DEA Echo Canceller. Input Signal Spectrum Equalization.

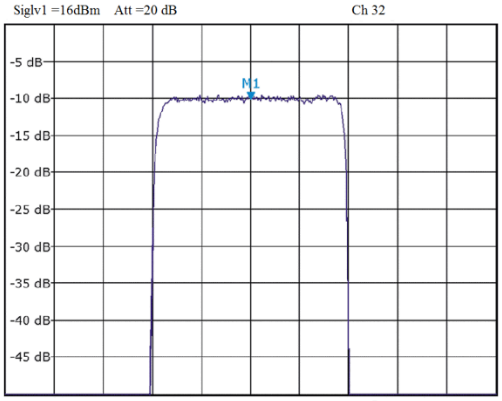


Fig. 9. Cancellation Example of an Echo up to 10 dB with DEA Echo Canceller. Output Signal Spectrum Equalization.

V. Results and Discussion

A. Measurement Results

The experimental measurements have been conducted mainly to evaluate the MER performances of the PGF scheme with a DEEC &DAE, compared with SGF, over a Rician multipath fading channel. The materials listed below are required during the installation of a GF, which are used in the results of indoor measurements.

- Computer with Ethernet cable.
- Signal analyzer ANRITSU model MS2712E or equivalent.
- Power Meter and Dummy load.
- RX antenna must have high directivity and high Ratio Front/Back.
- The professional Yagi antenna is suitable for this kind of installation.

The Tx antenna should not have a very open radiation diagram; ideally, the main lobe has 90° to 180°. It is also essential to pay attention to the side lobes, as they can affect the Rx antenna and cause strong coupling (echo) between receiving and transmitting signals.

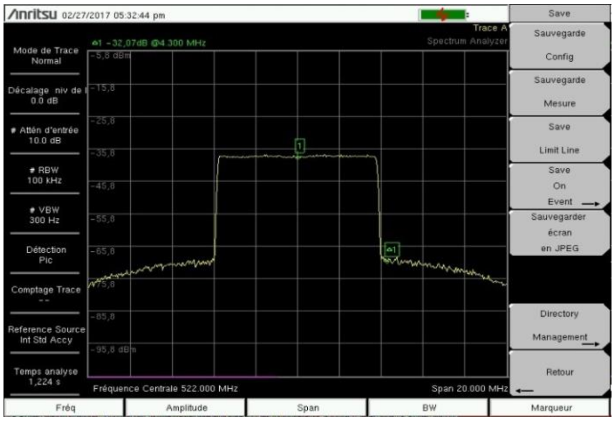
To perform the measurement equipment ROMAX TV Explorer HD + was used as a level meter. It is used to analyze the DVB-H/T/T2 signal providing several quality measurements (Power, C/N, MER, BER, constellation diagram, etc.). The results of the study could be useful for the DVB-T/H/T2 broadcast improvement, such as Tx adjustment and GF installation, to optimize the DTB efficiency.

Measurement points can be selected 1 Km away from the transmitter location, which means that the measurement points that will possibly be on each radiation path should not be more than 20°.

B. Intermodulation Measurements

We consider passing the signal through a GF, which is measured with Signal analyzer ANRITSU. Fig. 8(a) and 8(b) show the IMD Shoulder performance measurements of frequency relative to the center frequency of 8 MHz DVB-T versus Level measured, with 100 kHz resolution bandwidth (RBW). Fig. 10 represents the Standard GF (SGF), without DEEC &DAE, while Fig. 11 represents the PGF.

The nonlinear distortion in Fig. 10 caused by echoes phenomenon further brings an IMD Shoulder degradation performance of about -16.08 dB compared to the result proposed method Fig. 11. Therefore, the proposed Echo Cancellers, incorporated in GF, can still be used to improve the IMD performance of the DTT system significantly.


Anritsu 02/27/2017 05:32:44 pm

Trace A
Spectrum Analyzer

⌂1 -32,07dB @4.300 MHz

Mode de Trace
Normal

Décalage niv de l
0.0 dB

Attén d'entrée
10.0 dB

RBW
100 kHz

VBW
300 Hz

Détection
Pic

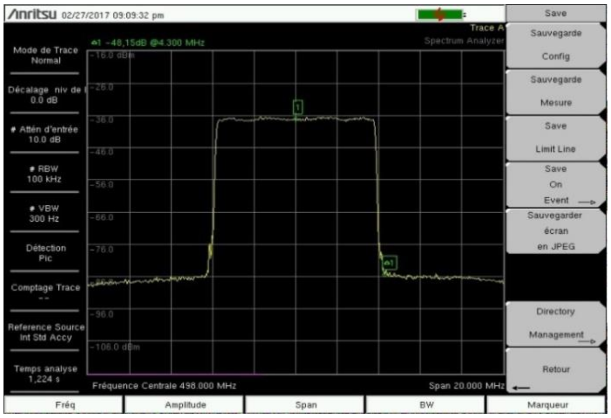
Comptage Trace
--

Reference Source
Int Std Accy

Temps analyse
1,224 s

Fréquence Centrale 522.000 MHz
Span 20.000 MHz

Fig. 10. IMD Measurement: -5.8 dBm Corresponds to the Average Output Power Test Interface of SGF.


Anritsu 02/27/2017 09:09:32 pm

Trace A
Spectrum Analyzer

⌂1 -48,15dB @4.300 MHz

Mode de Trace
Normal

Décalage niv de l
0.0 dB

Attén d'entrée
10.0 dB

RBW
100 kHz

VBW
300 Hz

Détection
Pic

Comptage Trace
--

Reference Source
Int Std Accy

Temps analyse
1,224 s

Fréquence Centrale 498.000 MHz
Span 20.000 MHz

Fig. 11. IMD Measurement: PGF.

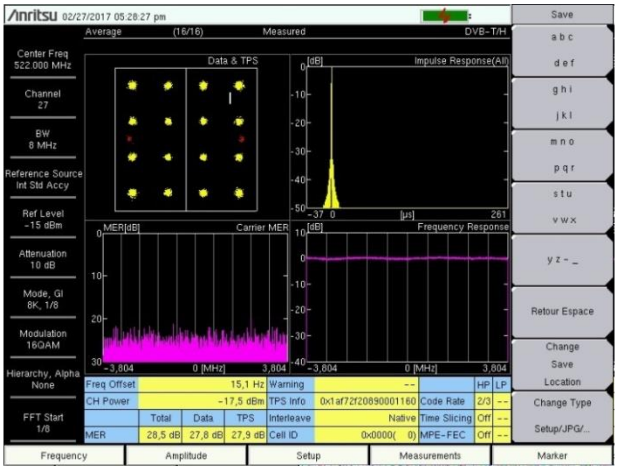
It should be noted that, based on experience, the increasing attenuation can surely minimize the IMD contribution of the RF signal analyzer. The measure of the third-order distortion with a significant attenuation is not detectable because it corners with the measurement values located outside tolerance.

C. MER Measurements

Fig. 12 and 13 present MER/Constellation measurement of the signal received from the test interface of the SGF and PGF, respectively, giving their performance comparison, 0 dB corresponds to average output power. Fig. 12 shows the proven MER measurement for a DVB-T signal that was mainly affected by unwanted modulator phase noise. It shows that the transmission quality of DVB-T is not satisfactory.

Fig. 13 illustrates the PGF result, and the constellation diagram also depicts a high-quality DVB-T signal (the points are not scattered).

From Fig. 12 and 13, it can be observed that the MER performance of the PGF is better than SGF; an improvement can reach up to 15 dB. We also observe that the DEEC & DAE can improve the robustness of the constellation, free of noise and interferences that lead to the better average. Therefore, a higher MER value promotes wide broad coverage area

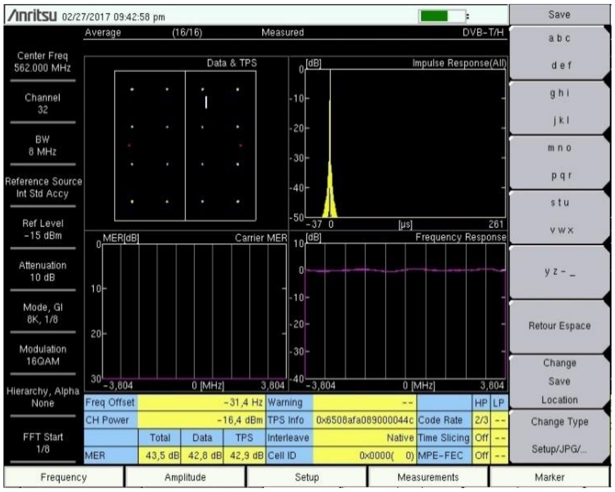

Anritsu 02/27/2017 05:28:27 pm

Average (16/16) Measured DVB-T/H

Center Freq 522.000 MHz
Channel 27
BW 8 MHz
Reference Source Int Std Accy
Ref Level -15 dBm
Attenuation 10 dB
Mode, GI 8K, 1/8
Modulation 16QAM
Hierarchy, Alpha None
FFT Start 1/8

Freq Offset	15,1 Hz			Warning	--		HP	LP
CH Power	-17,5 dBm			TPS Info	0x1af72f20890001160	Code Rate	2/3	--
	Total	Data	TPS	Interleave	Native	Time Slicing	Off	--
MER	28,5 dB	27,8 dB	27,9 dB	Cell ID	0x0000(0)	MPE-FEC	Off	--

Fig. 12. MER and Constellation Measurement: SGF.


Anritsu 02/27/2017 09:42:58 pm

Average (16/16) Measured DVB-T/H

Center Freq 562.000 MHz
Channel 32
BW 8 MHz
Reference Source Int Std Accy
Ref Level -15 dBm
Attenuation 10 dB
Mode, GI 8K, 1/8
Modulation 16QAM
Hierarchy, Alpha None
FFT Start 1/8

Freq Offset	-31,4 Hz			Warning	--		HP	LP
CH Power	-16,4 dBm			TPS Info	0x6508afa089000044c	Code Rate	2/3	--
	Total	Data	TPS	Interleave	Native	Time Slicing	Off	--
MER	43,5 dB	42,8 dB	42,9 dB	Cell ID	0x0000(0)	MPE-FEC	Off	--

Fig. 13. MER and Constellation Measurement: PGF.

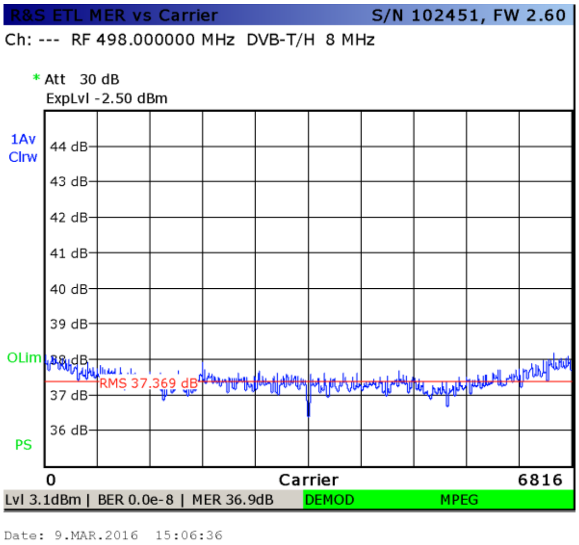
D. Comparison between the PGF and the Adaptive GF Technique

To further evaluate the performance of the proposed method, we compare it to another Echo Canceller method, more particularly, the adaptive GF technique. Fig. 14 and 15 ("MER vs. CARRIER") show the Echo Pattern of AGF technique compared to the PGF.

The average MER value is represented, in the screen, by a red line ted, which is measured in the whole channel (RMS) of about 37.4 dB (Adaptive GF) and 42.21 dB (PGF). It observed that the MER of PGF is better than the MER of AGF.

E. Comparison of an Existing Technique

Fig. 16 and 17 show the comparison simulation results (MER and receive signal levels) of different GF techniques obtained in four echo cancellation, namely standard GF(SGF), finite impulse response (FIR) filter, GF, adaptive GF, and the PGF.


R&S ETL MER vs Carrier S/N 102451, FW 2.60

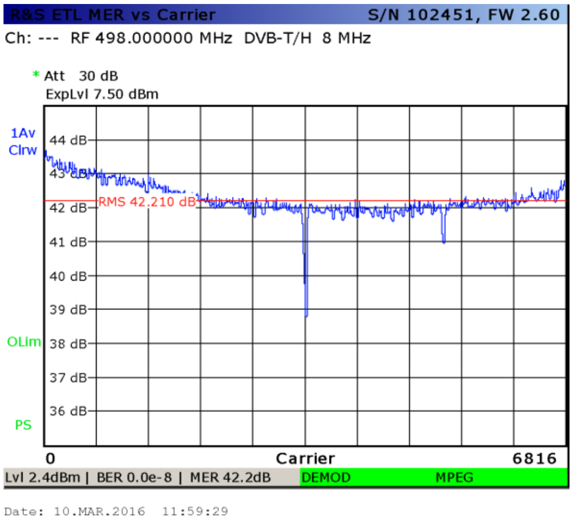
Ch: --- RF 498.000000 MHz DVB-T/H 8 MHz

* Att 30 dB
ExpLvl -2.50 dBm

Lvl 3.1dBm | BER 0.0e-8 | MER 36.9dB DEMOD MPEG

Date: 9.MAR.2016 15:06:36

Fig. 14. MER vs. Carrier of Adaptive GF.


R&S ETL MER vs Carrier S/N 102451, FW 2.60

Ch: --- RF 498.000000 MHz DVB-T/H 8 MHz

* Att 30 dB
ExpLvl 7.50 dBm

Lvl 2.4dBm | BER 0.0e-8 | MER 42.2dB DEMOD MPEG

Date: 10.MAR.2016 11:59:29

Fig. 15. MER vs. Carrier of PGF.

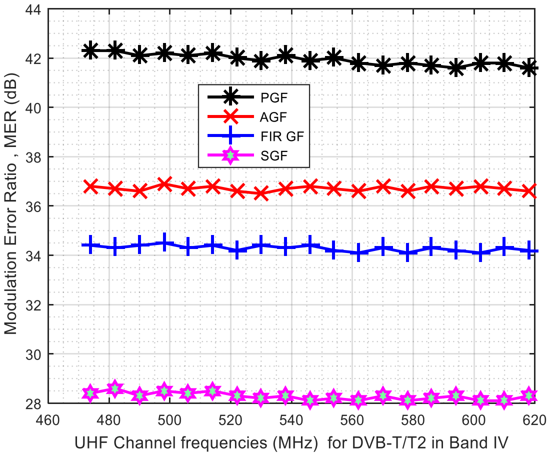


Fig. 16. MER vs. UHF Channel Frequencies Mhz.

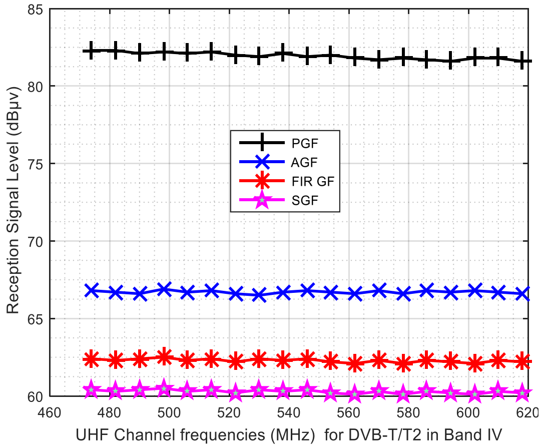


Fig. 17. Reception signal level vs. UHF channel frequencies Mhz.

Simulation results (Fig. 16 and 17) show perfect agreement with those obtained by measurements. Moreover, the MER and receive signal levels quality obtained by the PGF technique is much better than that obtained by SGF, FIR filter GF, and AGF.

F. *General Analysis and Discussion*

Table III summarizes the comparison between the different existing techniques obtained by four different Echo Cancellation techniques.

The Summary of analysis results, in Table III, illustrates that in four cases, the PGF outperforms the existing techniques. The proposed GF outperform standard GF(SGF), finite impulse response (FIR) filter (FIR GF)), GF and adaptive GF (AGF) techniques by 33%, 17%, and 13%, respectively, this outperformance is also characterized by:

Maximizing received signal quality: better than-60dBm level, MER.

Better than 37dB and lower incidence of multipath.

Less coupling between the TX and RX antennas: Echo Level always less than -20dB.

Can suppress high feedback echo levels.

TABLE. III. COMPARISON OF THE PROPOSED GF WITH OTHER GF TECHNIQUE

Echo cancellation Techniques	MER dB	IND dB	Echo levels dB	Signal quality dBm
SGF	28.5	-32.07	-10	-80
FIR GF	34.5	-37.8	-16	-71
AGF	36.9	-40.8	-18	-66
PGF	42.2	-48.15	-22	-58

DEEC can also suppress any multipath, and even Doppler echo in a free cancellation gap going up to 37.6 µs.

Also, in light of the findings, and achievement of the aims and objectives of the study, the results (with directive antennas on both the gap filler side and the principal transmitter side) are depicted in Fig. 16 and 17, in terms of MER/Reception signal level vs. UHF channel frequencies MHz respectively, between transmitter and receiver. It can be noted that the proposed solution provides the most extensive MER, and therefore seems to be the most suitable to provide coverage in SFN with a reduced number of GF.

The result is mainly related to the highly efficient air interface of the conventional GFs and, in particular, to its advanced coding technique. What is not taken into account in this study is the effect of severe multipath on the modulation schemes (ideal synchronization is assumed).

The overall research approach would suffer more than the many competitors, which are based on OFDM modulation, more robust to multipath. Such a drawback could be partially compensated by the antenna directivity, which provides a kind of spatial filtering of the echoes. Among the OFDM-based systems, the most extensive MER is provided by the DVB-H solution. The selection of the research method suffers for the weak coding scheme on the physical layer (i.e., error-correcting codes) in SFN, which requires precise synchronization and more considerable bandwidth.

VI. CONCLUSIONS

In this paper, the authors propose an efficient GF, including two echo cancellations, DEEC, and DAE, as a critical technique to solve the echo and interference problem in the SFN environment. Furthermore, it presents DAE and DEEC for ideally limit the isolation, in SFN, between the broadcasting and receiving antennas. The modified GF scheme (PGF) with DAE and DEEC echo canceller is proposed to cover shadow zones cancelling echoes with no self-interference from adjacent transmitters for DTT in SFN. Also, these GFs able to retransmit the RF signal under the most challenging echo and multipath propagation conditions removing high feedback echo levels. The PGF gives an excellent MER and can even remove the Doppler Effect. In our future work, we will use these areas as starting points for strengthening information protection for the new DVB generation in SFN.

ACKNOWLEDGMENT

The authors would like to acknowledge the moral and financial efforts of Dr. Wafaa Roky.

References

[1] H. Hamazumi, K. Imamura, N. Iai, K. Shibuya, and M. Sasaki, "A study of a loop interference canceller for the relay stations in an SFN for digital terrestrial broadcasting," IEEE. Global Telecommunications Conference (GLOBECOM'00) San Francisco, USA, pp. 167-171.2000

[2] M. Lanza, A.L. Gutiérrez, I. Barriuso, J.R. Pérez, M. Domingo, L. Valle, J. Basterrechea, J. Morgade, and P. Angueira , "Coverage optimization in single frequency networks using simulated annealing," IEEE International Symposium on Antennas and Propagation (APSURSI) , Spokane, USA, pp. 2789- 2792.2011.

[3] A. Mattsson, "Single-frequency networks in DTV," IEEE Transactions on Broadcasting, vol. 51, DOI: 10.1109/TBC.2005.858419 ,no.4, pp. 413-422, Dec, 2005.

[4] S.S. Das; F.H.P. Fitzek; E.D. Carvalho; R. Prasad, "Variabl guard Interval OFDM in presence of carrier frequency offset," IEEE Global Telecommunication Conference(GLOBECOM), MO, USA, DOI: 10.1109/GLOCOM.2005.1578296,pp. 2937-2941. Dec.2005

[5] Hsiao-Chun Wu, "Analysis and characterization of intercarrier and interblock interferences for wireless mobile OFDM systems," IEEE Transactions on Broadcasting, vol. 52, DOI: 10.1109/TBC.2006.872989, no. 2, pp. 203 –210. June. 2006.

[6] Vincent Savaux, Moise Djoko-Kouam and Alexandre Skrzypczak, "Study of cyclic delay diversity for single frequency network using DRM Standard," European Wireless19th European Wireless Conference, Guildford, UK, pp. 1-6. April .2013

[7] Marta Lanza,; Ángel L. Gutiérrez, Jesús R. Pérez, Javier Morgade , Marta Domingo, Luis Valle, Pablo Angueira and José Basterrechea, "Coverage optimization and power reduction in SFN using simulated annealing," IEEE Transactions on Broadcasting, vol. 60, DOI: 10.1109/TBC.2014.2333131,no. 3, pp. 474-485.Sep.2014

[8] El Miloud Ar-Reyouchi and Kamal Ghoumid,"Technical accuracy based on efficiency algorithm for measuring standing wave ratio in wireless sensor network," International Journal on Communications Antenna and Propagation vol. 9, nº. 2, pp.137-14, April 2019.

[9] J.-D. Kim, Y.-S. Byun," A new inter-carrier interference cancellation using CP-ICA scheme in OFDM Systems," IEEE 65th Vehicular Technology Conference - VTC2007-Spring, Dublin, Ireland, DOI:10.1109/vetecs.2007.489, pp. 2369- 2373. April .2007

[10] El Miloud Ar Reyouchi, Kamal Ghoumid, Amezian Koutaiba, Otman Mrabet, "MIMO-OFDM coded for digital terrestrial television broadcasting systems," World Academy of Science, Engineering and Technology 76, p. 617-621. Avril, 2013.

[11] El Miloud Ar Reyouchi, K. Ghoumid, K.Amezian, and O.Mrabet, "The powerful Alamouti code in MIMO-OFDM improvement for the next generation of terrestrial television broadcasting systems," International Journal of Engineering & Technology IJET-IJENS. vol. 14, no. 1, pp. 33-42, January 2014.

[12] Ar-Reyouchi, El Miloud. " Optimisation des performances des réseaux de communications sans fil: Performances des réseaux de communications sans fil pour la télégestion des stations de la Télédiffusion TV/FM''. https://www.amazon.com/Optimisation-performances-réseaux-communications-sans/dp/3841642861,Isbn: 978-3-8416-4286-8. Publisher: Paf ,June 2017.

[13] A. Gohn and J. Kim, "Implementation of LMS adaptive filter algorithm based on FPGA," 2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), Dallas, TX, USA, pp. 207-210,2019.

[14] Mahmood Farhan Mosleh, Aseel Hameed AL-Nakkash, "Combination of LMS and RLS adaptive equalizer for selective fading channel," European Journal of Scientific Research, vol.43.no.1, pp.127-137, Jun 2010,

[15] R. Meier, E. DeMan, T. G. Noll, U. Loibl and H. Klar, "A 2 pm CMOS digital adaptive equalizer chip for QAM digital radio modems," IEEE Journal of Solid-State Circuits, vol.23, DOI: 10.1109/4.5946,No.5,pp. 1212-1217, November.,1988.

[16] X. Hou, and C. Yang, "Feedback overhead analysis for base station cooperative transmission," IEEE Transactions on Wireless Communications, vol. 15, DOI: 10.1109/TWC.2013.013013.120534, no. 7, July 2016.

[17] Ar-Reyouchi El Miloud, Lichioui, Rattal Salma, "A group cooperative coding model for dense wireless networks," International Journal of Advanced Computer Science and Applications. 10. 10.14569/IJACSA.2019.0100750, 2019.

[18] H. Bawab, P. Mary, J. Hlard, Y. Nasser, and O. Bazzi, "Spectral overlap optimization for DVB-T2 and LTE Coexistence," IEEE Transactions on Broadcasting, vol. 64, no. 1, pp. 70-84, Mar. 2018.

[19] K. M. Nasr, J. P. Cosmas, M. Bard, and J. Gledhill, "Performance of an echo canceller and channel estimator for on-channel repeaters in DVB-T/H Networks," IEEE Transactions on Broadcasting, vol. 53, DOI: 10.1109/TBC.2007.903612, no. 3, pp. 609-618, September 2007.

[20] F. Lindqvist and A. Fertner, "Frequency domain echo canceller for DMT-based systems," IEEE Signal Processing Letters, vol. 18, DOI: 10.1109/LSP.2011.2170679, no.12, pp. 713-716, .December.2011.

[21] P. P. Kadam, Z. Saquib and A. Lahane, "Adaptive echo cancellation in VoIP network," 2016 IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, pp. 295-299,2016.

[22] C. Rocha, C. Akamine, G. Bedicks, E. L. Horta and G. Stolfi, "Adaptive gap filler for digital terrestrial television," IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Beijing, China, DOI: 10.1109/BMSB.2014.6873535 pp.1-6. June.2014.

[23] Michael Mao Wang, "Dynamic gain management for on-channel repeaters," IEEE Transactions on Broadcasting, vol.59, DOI: 10.1109/TBC.2013.2284417, no.4, pp. 685-692. Dec.2013

[24] M. M. Sondhi, "The history of echo cancellation", IEEE Signal Processing Magazine, vol. 23, DOI: 10.1109/MSP.2006.1708416, no. 5, pp. 95 –98. September .2006

[25] Fuyun Ling, "Achievable performance and limiting factors of echo cancellation in wireless communications," Information Theory and Applications Workshop (ITA), San Diego, USA, DOI: 10.1109/ITA.2014.6804210, pp.1-8.Feb.2014.

[26] Bundit Ruckveratham , Sathaporn Promwong'Empirical single frequency network threshold for DVB-T2 based on laboratory experiments'' Turkish Journal of Electrical Engineering & Computer Sciences, pp: 3342 – 3355, 2019

[27] Ruckveratham B, Promwong S. "Evaluation of SFN gain for DVB-T2. In: International Conference on Digital Arts, Media and Technology", Chiang Mai, Thailand. New York, NY, USA: IEEE. pp. 85-88,2017.

[28] ETSI EN 300 744 V1.4.1 (2001-01). European Standard. Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation for DTT. ETSI, 2001.

[29] Yin Wang, Zhaowu Chen and Ke Gong, "MER performance analysis of M-QAM OFDM with wiener phase noise," IEEE International Conference on Microwave and Millimeter Wave Technology, DOI: 10.1109/ICMMT.2007.381517 pp. 18-21, April 2007.

Assessing Advanced Machine Learning Techniques for Predicting Hospital Readmission

Samah Alajmani[1], Kamal Jambi[2]
Computer Science Department, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

Abstract—**Predicting the probability of hospital readmission is one of the most important healthcare problems for satisfactory, high-quality service in chronic diseases such as diabetes, in order to identify needful resources such as rooms, medical staff, beds, and specialists. Unfortunately, not many studies in the literature address this issue. Most studies involve forecasting the probability of diseases. For prediction, several machine learning methods can be implemented. Nonetheless, comparative studies that identify the most effective approaches for the method prediction are also insufficient. With this aim, our paper introduces a comparative study in the literature across five popular methods to predict the probability of hospital readmission in patients suffering from diabetes. The selected techniques include linear discriminant analysis, instance-based learning (K-nearest neighbors), and ensemble-based learning (random forest, AdaBoost, and gradient boosting) techniques. The study showed that the best performance was in random forest whereas the worst performance was shown by linear discriminant analysis.**

Keywords—Boosting; random forest; linear discriminant analysis; k-nearest neighbor; machine learning; hospital readmission; predictive analytics

I. Introduction

Healthcare is globally recognized as an important sector. It is one of the most rapidly growing divisions in the employment industries worldwide [1]. According to Russell Reynolds Associates, healthcare costs are predicted to reach a staggering $12 trillion within the next 7 years. Current costs are between $6–$7 trillion [2]. Looking at these figures, it is obvious that healthcare is at a crucial point in the growth and evolution of medicine. A perfect example of this can be seen in the United States, where the total expenditure on healthcare increased by up to 5.3%, and also topped $3 trillion nationally. In 2012, more than 1.5 million people died of diabetes, which is one of the most common and chronic illnesses of our times [3]. This serious disease continues to affect many people around the world. Diabetes affects more than twenty-three million person in the United States alone [4]. Furthermore, the main concern in diabetes care is hospital readmission, reasoned by spending of more than two hundred fifty million dollars on medications for diabetic patients who were readmitted in 2011 [4]. The Agency for Healthcare Research and Quality (AHRQ) revealed more than three million readmissions during a 30-day window in the US. Contribution of these hospital readmissions reached about forty-one billion dollars in hospital costs [5]. Additionally, the totals deaths in 2012 was estimated to be 3.7 million, which included 1.5 million deaths due to diabetes, and an additional 2.2 million deaths due to cardiovascular diseases, chronic kidney disease, and tuberculosis related to higher-than-optimal blood glucose. Research has shown that 43% of these 3.7 million deaths occur before individuals have reached the age of 70. Statistically, diabetes and high blood pressure are more prevalent in lower and middle-class people, as opposed to the higher-income group [3]. Hence, these diabetic patients tend to frequently visit healthcare facilities and hospitals, which guarantees them access to hospital resources and services, for example, the availability of sufficient services by care providers and other hospital staff, medical equipment, early detection and diagnosis of illnesses and diseases, medical treatments, and check-ups and plans developed by the medical staff for the patient. Preventing hospitalization is a distinguished aspect of limiting an affected person’s morbidity, improving their results, and constraining healthcare costs [6]. Accordingly, their ultimate purpose is to predict readmission possibility. In reality, readmission for 30 days, that is, within one month of discharge is an excellent indicator of high priority healthcare. The aim of this study is to discuss this issue as well [7].

Machine learning is one of the most popular and leading analytical techniques developed in modern times. It intends to solve highly complicated tasks because it is essential to concentrate on the most appropriate data from seemingly enormous amounts of data [8]. This type of learning includes gathering information from various fields to redefine issues beyond the normal limitations and to arrive at solutions based on a novel understanding of complicated attitudes. It expands over the fields of statistics, algebra and knowledge, data processing, analytics, etc. This type of learning is also influenced by artificial intelligence, control theory, biology, philosophy, information technology, cognitive science, and mathematical calculations. With machine learning, gathering accurate data—ranging from medical records, financial transactions, applications of loans, and supply maintenance—has become possible [9].

Machine learning can be categorized into three groups. Supervised learning is subject to learning as a data scientist attempts to teach a data training algorithm and its possible outcomes. Classification and regression are two different models in machine learning. The classification model aims to forecast unique classes such as blood groups, whereas the regression model predicts numerical values [10]. On the other hand, unsupervised learning tries to look for a pattern in hidden data, associations among variables, or trends in data

[10] [11]. The major goal of this type of learning is the capability to determine either the distribution of data or the hidden structure without the earlier categorization of training data or without being subject to supervision [12]. Lastly, reinforcement learning is learning by an individual where they can study behavior by means of both interactions (the trial and error method) and dynamic environment. Through this learning, a computer program can provide access to the dynamic environment for executing a special goal. It is essential to understand that the system did not have prior knowledge of the behavior of the environment and the only probable method will be through trial and error [10] ,[13]. It is worth noting that in this study, we extend our work in paper [14] by adding five other techniques to compare and detect the best accuracy among the following techniques regarding hospital readmission prediction: 1) Linear discriminant analysis (LDA); 2) Instance-based learning: K-nearest neighbor; and 3) Ensemble-based learning: AdaBoost, gradient boosting, and random forest.

The remainder of the paper is presented as follows: Section 2 includes a background discussion on machine learning algorithms used in this study. Section 3 presents related work on comparative study of machine learning techniques in the healthcare sector. Section 4 discusses the study methodology and presents the outcomes of the experiments. Section 5 summarizes and concludes the study.

II. Background

Developments in machine learning are clearly visible in different fields and industries in the past years. Hence, researchers have discussed the possibility of using machine learning technologies in healthcare, outlining different initiatives in the healthcare domain. Machine learning has many benefits for healthcare, such as predicting readmission in hospitals and diseases, among others. In addition, machine learning is capable of discovering a solution to form a strong relationship between the patient and doctor for reducing the increasing healthcare costs [1]. This section addresses different machine learning techniques used in this study.

A. Linear Discriminant Analysis

LDA is an essential data analysis approach, which has been widely applied in the past for recognition, dimensionality reduction, and supervised classification [15]. it is a mathematical classification technique which can search for a collection of predictors to distinguish between two targets. It is also correlated to regression analysis in that both try to express the relation between an independent variables group and a single dependent variable [16].

B. Instance-Based Learning

This algorithm can be called “Memory-based learning.” The most broadly utilized technique of instance-based learning for classification is K-nearest neighbor (KNN) [16]. This method is largely applied to sample classification. The KNN technique can measure the distance from training samples number N [17]. This technique does not attempt to build an internal model, and computations are not executed prior to the classification. In the features space, the training data instances are hardly retained. Next, depending upon the vote’s majority from the neighbors, a class of instance is determined. Moreover, an instance is determined for a class that is most common among the neighbors. For variables that are continuous, Murkowski, Euclidian, or Manhattan distance techniques are used, whereas the Hamming technique is used for variables that are categorical [16]. Depending upon the distance, the neighbors are determined using the KNNs. The determined distances are used to recognize and allocate labels to training instances’ (k) groups that are nearest to the new point. Despite its simplicity, the KNN has been utilized in a considerable number of applications.

C. Ensemble-Based Learning

Ensemble-based learning techniques lead to predictions that depend on a collection of several classifier outputs. Ensemble learners consist of boosting methods, for example, AdaBoost and gradient boosting, along with bagging methods such as random forest [16].“Boosting” signifies a general and effective strategy to yield highly precise prediction by gathering hard and slightly inexact thumb rules [18]. On the other hand, “bagging” depends on a bootstrapping strategy, in which different classification trees are improved by constantly choosing arbitrary training data subsets [19].

1) Boosting-Based techniques: AdaBoost and Stochastic Gradient Boosting rely on the concept of boosting [16]. AdaBoost (AB) is based on creating a prediction rule, which is extremely accurate, by joining a number of comparatively weak and inexact rules [20]. Furthermore, it is simple, swift, and easy to use with an iterative algorithm which requires only one parameter, iteration number. Moreover, it is not subject to over-fitting and simply determines outliners which are incorrectly classified or are difficult to classify [21]. However, misclassified and/or difficult instances are given significance by gradient boosting (GB), via the remaining errors—also known as pseudo-residuals—of a strong learner. With every iteration, errors are measured, and a weak learner adapts to them. Afterwards, the weak learner contributes to minimizing the total error of the strong learner [16].

2) Baging techniques: Random forests are a combination of tree predictors. It is an ensemble learning method (in addition to the thought that it could be a form of the nearest neighbor predictor). It creates a number of decision trees at the time of training and produces a class, which is the output of classes through individual trees. Furthermore, it attempts to reduce increased bias and variance issues through averaging to detect a balance between the two extremes [22].

III. Related Work

Apart from predicting the probability of hospital readmission, many studies have attempted to use machine learning techniques in healthcare problems. For example, AOA et al. [23] clarified the importance of machine learning techniques in identifying predictive and diagnostic indicators in a set of wide-scale data with extremely elevated geometric relation to genetics. They used SVM, logistic regression (LR), and NBs and proved that SVM had the best accuracy among others. Arun and Sittidech [24] used decision tree (DT), KNN, and NBs to build diabetes classification models. Then, boosting and bagging were executed using the base classifiers of KNN, NBs, and DT. Based on their tests, they concluded that the greatest accuracy was obtained when bagging was applied with DT. Sisodia and Sisodia [25] presented a model that could predict the probability of diabetes with high accuracy. In their experiment, they applied three techniques: NBs, DT, and SVM, to discover diabetes in its early stages. In accordance with their outcomes, NB achieved better accuracy than other algorithms. Singh [26] conducted an experiment to predict diabetes through the utilization of various machine learning techniques; the accuracy of the proposed technique was 87-95% better than others: DT (C4.5) at 81-85%, Bays classifier at 84-88%, and KNN at 80-82%. However, Shahon et al. had a different intention; they tried using AdaBoost to improve the overall accuracy of models. Consequent to these experiments, it was clear that AdaBoost had a better accuracy than standalone DTs, such as J48, and bagging [27]. Orabi et al.'s approach [28] to fuse regression included randomization. This method achieved an 84% accuracy rate when predicting diabetes according to age. Other investigators suggested a predictive model. They used three techniques of machine learning—SVMs, RFs, and LR—to predict diabetes in Indian women, as well as the factors that could cause the disease. Their comparative study proved that the RFs had the highest performance among other models [29].

In comparison, only a few studies have discussed the prediction of hospital readmission probability. For example, Strack et al. [30] utilized traditional statistical models toward this end. Some investigators concentrated on the comparison of various machine learning techniques to address the issue. For example, Alexander et al. suggested two methods. First, they merged unsupervised and supervised techniques of classification, and subsequently, merged DT and NB. They proved that the former method had better accuracy than the latter method with regard to readmission prediction [31]. Finally, Alajmani and Elazhary [14] used LR, multi-layer perceptron (MLP), NB, SVM, and DT to predict hospital readmission and evaluate accuracy among models. Based on their results, SVM achieved the highest accuracy of 95.22% among other techniques.

In general, very few studies in the healthcare sector are devoted to predicting the possibility of hospital readmission. In addition, there is a lack of research on comparison among different machine learning techniques for prediction. Consequently, this paper attempts to address and discuss both issues regarding the probability prediction of hospital readmission, which depends on real data with different algorithms.

IV. Methodology

It is imperative to clearly understand the data prior to commencing a comparative study, conduct pre-processing when needed, and choose appropriate features for the experiments. It is also important to mention that all experiments in this study were conducted using Python.

A. Dataset Explanation and Features

*1) Data comprehension:*This paper utilizes a sample dataset of diabetic patients from different hospitals across the US [32] , [30]. Such a dataset encompasses 13460 instances from age groups 30–50, with eighteen features. In Table I, the dataset variables and their associated descriptions are presented. Scientific interpretations of these features are beyond this article's scope. In addition, the distribution of features is depicted in Fig. 1.

2) Data pre-processing: This phase, which encompasses both data transformation and data cleaning, is considered to be a significant step. We tried to use an approach that is frequently used and more general in converting categorical variables into variables of real-value; this approach is called one-hot encoding [33]. First, with regard to data transformation, certain categorical variables such as Gender, Change, Age and DiabetesMed are converted into binary forms 0 or 1. Second, with regard to data cleaning, missing values of categorical data need to be accounting for. Toward this aim, the imputation is performed via the categorical data mode. This imputation method helps us with better prediction model performance in cases where missing data has already hidden helpful information [34]. After preprocessing the data became 3090 instances.

3) Feature selection: Here, feature selection is applied to reduce dimensionality, meaning we opt for features that are most relevant. In this research paper, the effect of variables on our target is evaluated. Moreover, this results in the elimination of low-importance variables. The most significant among them are features with high influence on accuracy [16]. The GB technique has been utilized [35] for categorical variables. The variables' average weights are demonstrated in Table II. Subsequently, a threshold of 0.014 is used to attain the variable set. Consequently, the features Age, Admission_source_id, and DiabetesMed are excluded as their weights are less than 0.014. However, the other features demonstrated in Fig. 2 are chosen and selected.

B. Constructing Models of Machine Learning

In this paper, the chosen models have 1 target/output with 2 values, which can be true or false regarding readmission to the hospital within a span of one month. This means the value of the readmission variable is TRUE if the patient is readmitted within a time span of one month. However, if there is no readmission, or if readmission has been carried out after the one-month period, then the value will be FALSE. As mentioned earlier, the driver set for forecasting consist of the selected features. The datasets for training and testing are selected randomly. Moreover, by choosing a 40% testing dataset and a 60% training dataset, a ten-fold cross-validation is applied.

1) Linear discriminant analysis: This model is built using the next parameters n_components, solver, and tol, where n_components is the number of components (< n_classes-1) for reducing dimensionality. Solver "svd" is the decomposition of a singular value. Finally, tol "1e-5" is the threshold to be utilized for estimation of rank in solver of svd. The accuracy of LDA is 0.6388515 and a 10-fold cross-validation is conducted for this model.

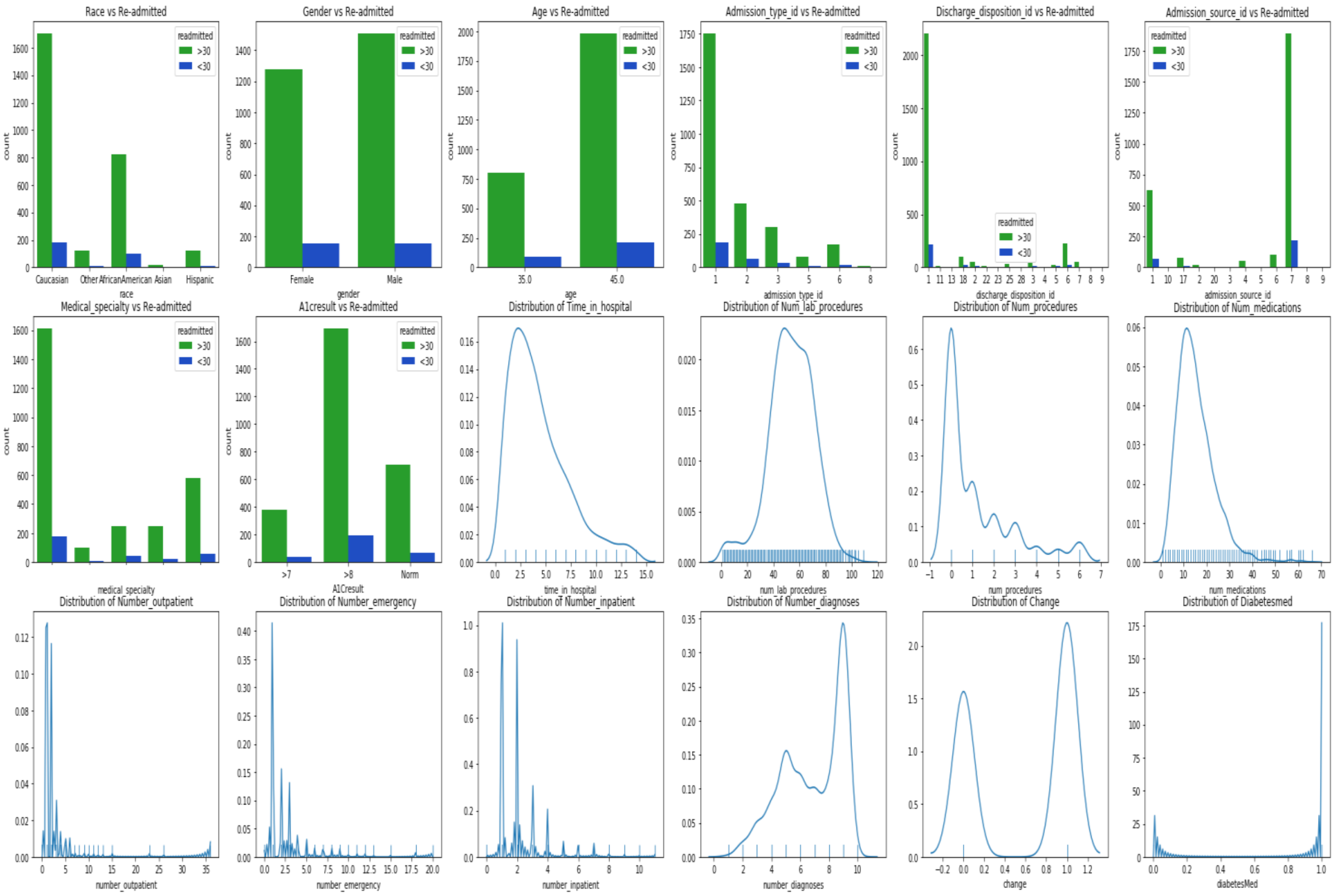


Fig. 1. Features Distribution.

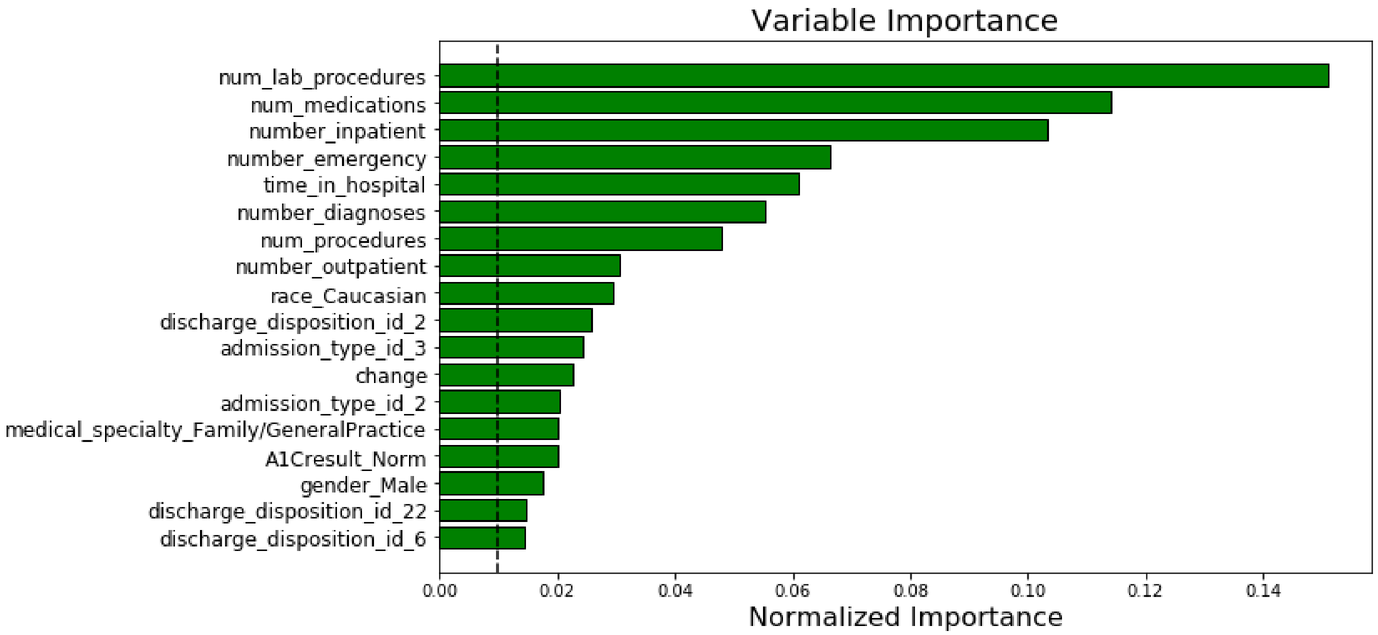


Fig. 2. Variables Importance.

2) K-Nearest neighbor: In this model, the most important parameter is n_neighbors, which represents the number of neighbors for use by default for k neighbors queries. Cross-validation is executed using different values of n_neighbors. Table III illustrates that the highest accuracy = 0.8847016 when n_neighbors = 5.

3) Adaboost: AdaBoost needs three important parameters: (1) n_estimators indicate the number of weak learners for repeat training, (2) learning_rate contributes to weak learners' weights, and (3) algorithm "SAMME" or "SAMME.R." This model uses grid search to evaluate the optimal accuracy and hyperparameters. The best parameters are n_estimators = 5000, algorithm = "SAMME," and learning rate=0.9; thus, the best accuracy of 0.9318079 is clarified in Table IV below.

4) Gradient boosting: This model is built using the following important parameters: (1) n_estimators, which can present number of boosting stages for execution, (2) learning rate indicates the shrinking of learning rate of every tree contribution, (3) creation is the function for measuring the split quality, and (4) max_depth refers to the maximum depth which can limit the number of nodes in the tree. Tuning the max_depth is important to get the best performance. Grid search is used to measure optimum accuracy and hyperparameters. Table V demonstrates that the best accuracy = 0.9362943 when n_ estimators = 200.

5) Random forest: We construct this model using 250 trees in the forest where 26 is the maximum depth of the tree and 10 is the lowest number of samples for dividing an inner node. In addition, grid search is used to find the best accuracy and the best parameters. The following Table VI presents the results.

TABLE. I. DIABETSE DATASET

Variable	**Data type**
Race	Categorical
Change	Categorical
DiabetesMed	Categorical
Age	Categorical
A1Cresult	Categorical
Gender	Categorical
Num_lab_procedures	Integer
Num_procedures	Integer
Num_inpatient	Integer
Num_oupatient	Integer
Num_medications	Integer
Num_diagnosis	Integer
Num_emergency	Integer
Medical_spacialty	Categorical
time_in_hospital	Integer
Admission_type_id	Integer
Admission_source_id	Integer
Discharge_disposition-id	Integer

TABLE. II. FEATURES IMPORTANCE

Variable	**Importance**	**Decision**
Race	0.029016	Acceptable
Change	0.023027	Acceptable
DiabetesMed	0.008867	Unacceptable
Age	0.010165	Unacceptable
A1Cresult	0.020177	Acceptable
Gender	0.020294	Acceptable
Num_lab_procedures	0.149317	Acceptable
Num_procedures	0.046521	Acceptable
Num_inpatient	0.104099	Acceptable
Num_outpatient	0.030696	Acceptable
Num_medications	0.111058	Acceptable
Num_diagnosis	0.055811	Acceptable
Num_emergency	0.066718	Acceptable
Medical_spacialty	0.019117	Acceptable
time_in_hospital	0.062025	Acceptable
Admission_type_id	0.023854	Acceptable
Admission_source_id	0.008961	Unacceptable
Discharge_disposition-id	0.027554	Acceptable

TABLE. III. KNN ACCURACY

n_neighbors	**Accuracy**
5	0.8847016
10	0.8501570
15	0.8205473

TABLE. IV. ADABOOST ACCURACY

n_neighbors	**Accuracy**
500	0.9255271
1000	0.9286675
5000	0.9318079

TABLE. V. GRADIENT BOOSTING ACCURACY

n_estimators	Accuracy
100	0.9344997
150	0.9358456
200	0.9362943

TABLE. VI. RANDOM FOREST ACCURACY

n_estimators	**Accuracy**
150	0.9349484
250	0.9358456
350	0.9344997

V. DISSCUSIOM AND RESULTS

In this study, different performance measures are utilized to compare the studied techniques [36]. Particularly, precision, accuracy, F1 scores, and recalls are relied upon for this reason. As presented in Equations 1, 2, 3, and 4, these parameters are described by true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Furthermore, TPs refer to cases where the prediction is YES, that is, patients will be readmitted in hospital within a duration of 30 days and when there is a match, meaning that the patients are indeed readmitted. Whereas, TNs refer to those cases where the prediction is a NO, and when the patients are NOT readmitted. On a different note, FPs refer to cases where the prediction is a YES, but patients are NOT readmitted, that is, a type I error. Lastly, FNs refer to cases where the prediction is a NO, but the patients are actually readmitted, that is, a type II error.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (1)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (2)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (3)$$

$$\text{F1\_score} = \frac{2*(Recall*Precision)}{(Recall\ +Precision)} \quad (4)$$

Accuracy refers to the frequency of the classifier being true. The recall is a sensitivity measure, for example, the proportion of TPs to the total number of TPs and FNs. It indicates the rate of cases where the model predicts patient readmission within a time span of 30 days, related to the number of events where the subject is actually readmitted. Alternatively, precision is a calculation of the rate of events when the model accurately predicts the patient's readmission during the 30-day time period, in contrast to sum of events when the model forecasts the patient's readmission. In Table VII, the performance measure values are illustrated.

As previously mentioned, we performed 10-fold cross-validation of the listed techniques. For each model, the training and testing accuracy with respect to 10-fold cross-validation is shown in Table VIII and Fig. 3.

Lastly, for the chosen models, the lowest, highest, and mean accuracies are illustrated in Table IX. It is obvious that ensemble-based learning (RF and AdaBoost) techniques accomplish the maximum accuracy of 0.9579 and 0.9550, respectively. Further, GB's accuracy is 0.9459 while KNN's accuracy is 0.9161. The least value of performance accuracy is 0.6835 in LDA. The complexity of each algorithm followed by each classification is the main reason behind the performance variation.

TABLE. VII. PERFORMANCE MEASURES FOR THE SELECTED MODELS

Models / Measures	**Accuracy**	**Precision**	**F1_score**	**Recall**
Random Forest	0.932705	0.988024	0.929577	0.877660
AdaBoost	0.931808	0.992929	0.928234	0.871454
Gradient Boosting	0.932705	0.970192	0.930812	0.894504
K-Nearest Neighbor	0.884702	0.857847	0.890405	0.925532
Linear Discriminant Analysis	0.638852	0.646952	0.638527	0.630319

TABLE. VIII. PERFORMANCE MEASURES FOR THE SELECTED MODELS

Random Forest	**AdaBoost**	**Gradient Boosting**	**K-Nearest Neighbor**	**Linear Discriminant Analysis**
0.937313	0.937313	0.925373	0.889552	0.623881
0.940299	0.952239	0.931343	0.904478	0.683582
0.937313	0.940299	0.943284	0.889552	0.614925
0.934328	0.934328	0.934328	0.889552	0.656716
0.931343	0.931343	0.916418	0.865672	0.600000
0.916168	0.913174	0.898204	0.838323	0.610778
0.931138	0.928144	0.934132	0.892216	0.679641
0.952096	0.955090	0.943114	0.916168	0.634731
0.957958	0.936937	0.945946	0.900901	0.630631
0.942943	0.927928	0.930931	0.909910	0.642643

TABLE. IX. SELECTED MODELS ACCURACY

Model	**Min**	**Max**	**Mean**
Random Forest	0.916168	0.957958	0.938090
AdaBoost	0.913174	0.955090	0.935679
Gradient Boosting	0.898204	0.945946	0.930307
K-Nearest Neighbor	0.838323	0.916168	0.890229
Linear Discriminant Analysis	0.600000	0.683582	0.637753

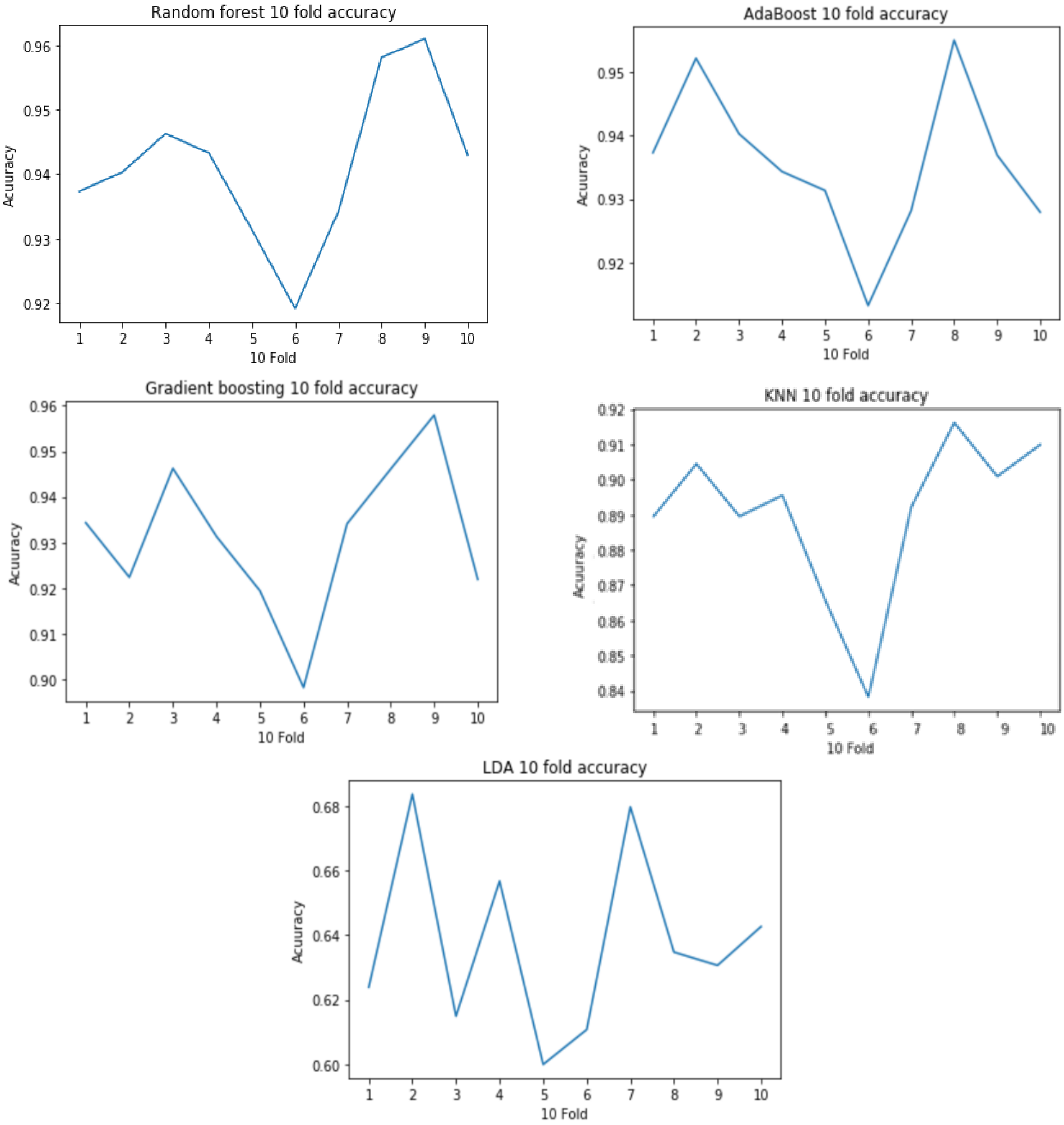


Fig. 3. 10-Folds Cross Validation for Models.

VI. Conclusion and Future Work

For the prediction of readmission, this research seeks to offer a standard for the most commonly applied modern features. The reliability of the chosen models is measured on a real dataset of diabetes from different hospitals in the USA. Based on the outcomes of this study, ensemble-based learning algorithms are suggested, and the highest accuracy is shown by the RF model, followed by the AdaBoost model. Nevertheless, the research will be expanded to a bigger dataset using different machine learning techniques.

References

[1] R. Bhardwaj, A. R. Nambiar, and D. Dutta, "A Study of Machine Learning in Healthcare," Proc. - Int. Comput. Softw. Appl. Conf., vol. 2, pp. 236–241, 2017.

[2] J. Bouwens and D. M. Krueger, "Embracing Change: The Healthcare Industry Focuses on New Growth Drivers and Leadership Requirements," Rusell Reynolds Associates. [Online]. Available: https://www.russellreynolds.com/insights/thought-leadership/embracing-change-the-healthcare-industry-focuses-on-new-growth-drivers-and-leadership-requirements.

[3] G. Roglic, "Global Report on Diabetes.," World Heal. Organ., vol. 58, no. 12, pp. 1–88, 2016.

[4] M. S. Bhuvan, A. Kumar, A. Zafar, and V. Kishore, "Identifying

Diabetic Patients with High Risk of Readmission," arXiv Prepr. arXiv1602.04257., 2016.

[5] "AHRQ: The Conditions that Cause the Most Readmissions," Advisory, 2014. [Online]. Available: https://www.advisory.com/daily-briefing /2014/04/22/ most-common-readmissions. [Accessed: 30-Oct-2019].

[6] K. Zolfaghar, N. Meadem, A. Teredesai, S. B. Roy, S. C. Chin, and B. Muckian, "Big Data Solutions for Predicting Risk-of-Readmission for Congestive Heart Failure Patients," in 2013 IEEE International Conference on Big Data.IEEE., 2013, pp. 64–71.

[7] D. J. Rubin, K. Donnell-Jackson, R. Jhingan, S. H. Golden, and A. Paranjape, "Early Readmission among Patients with Diabetes: A Qualitative Assessment of Contributing Factors," J. Diabetes its Complicat. Elsevier., vol. 28, no. 6, pp. 869–873, 2014.

[8] A. L. Bluma and P. Langley, "Artificial Intelligence Selection of relevant features and examples in machine," vol. 97, no. 97, pp. 245–271, 1997.

[9] T. Mitchell, "Machine Learning, McGraw-Hill Higher Education," New York, 1997.

[10] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017.

[11] P. Chowriappa, S. Dua, and Y. Todorov, "Introduction to Machine Learning in Healthcare Informatics," in Machine Learning in Healthcare Informatics.Springer, 2014, pp. 1–23.

[12] E. Bose and K. Radhakrishnan, "Using Unsupervised Machine Learning to Identify Subgroups among Home Health Patients with Heart Failure Using Telehealth," CIN - Comput. Informatics Nurs., vol. 36, no. 5, pp. 242–248, 2018.

[13] L. Kaelbling, A. Littman, and A. Moore, "Reinforcement learning: A survey," J. Artif. Intell. Res., vol. 4, pp. 237–285, 1996.

[14] S. Alajmani and H. Elazhary, "Hospital Readmission Prediction Using Machine Learning Techniques: A Comparative Study," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 4, pp. 212–220, 2019.

[15] P. Markopoulos, "Linear Discriminant Analysis with Few Training Data," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 4626–4630.

[16] S. F. Sabbeh, "Machine-Learning Techniques for Customer Retention: A Comparative Study," Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 2, pp. 273–281, 2018.

[17] K. Shailaja, B. Seetharamulu, and M. A. Jabbar, "Machine Learning in Healthcare: A Review," 2018 Second Int. Conf. Electron. Commun. Aerosp. Technol., no. Iceca, pp. 910–914, 2018.

[18] Y. Freund and R. Schapire, "A Short Introduction to Boosting," Journal-Japanese Soc. Artif. Intell., vol. 14, no. 771–780, p. 1612, 1999.

[19] G. Chirici et al., "Stochastic Gradient Boosting Classification Trees for Forest Fuel Types Mapping Through Airborne Laser Scanning and IRS LISS-III imagery," Int. J. Appl. Earth Obs. Geoinf., vol. 25, no. 1, pp. 87–97, 2013.

[20] R. E. Schapire, "Explaining Adaboost," in Empirical inference. Springer, 2013, pp. 37–52.

[21] N. Emanet, H. R. Öz, N. Bayram, and D. Delen, "A Comparative Analysis of Machine Learning Methods for Classification Type Decision Problems in Healthcare," Decis. Anal., vol. 1, no. 1, p. 6, 2014.

[22] L. Breiman, "Random Forests," Mach. Learn. Springer, vol. 45, no. 1, pp. 5–32, 2001.

[23] L. Cai, H. Wu, D. Li, K. Zhou, and F. Zou, "Type 2 Diabetes Biomarkers of Human Gut Microbiota Selected via Iterative Sure Independent Screening Method," PLoS One, vol. 10, no. 10, pp. 1–15, 2015.

[24] N. Nai-Arun and P. Sittidech, "Ensemble Learning Model for Diabetes Classification," Adv. Mater. Res., vol. 931–932, pp. 1427–1431, 2014.

[25] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes Using Classification Algorithms," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1578–1585, 2018.

[26] A. Singh, "Comparing Data Mining Algorithms for Diabetes Disease Prediction," Int. J. Contemp. Technol. Manag., 2018.

[27] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," Procedia Comput. Sci., vol. 82, no. March, pp. 115–121, 2016.

[28] K. M. Orabi, Y. M. Kamal, and T. M. Rabah, "Early Predictive System for Diabetes Mellitus Disease," vol. 1, pp. 420–427, 2016.

[29] D. Dutta, D. Paul, and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction Using Machine Learning Debadri," in 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, IEEE, 2018, pp. 924–928.

[30] S. B. et al., "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," Biomed Res. Int., vol. 2014, 2014.

[31] J. Kerexeta, A. Artetxe, V. Escolar, A. Lozano, and N. Larburu, "Predicting 30-day Readmission in Heart Failure Using Machine Learning Techniques," in Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies., 2018, vol. 5, no. Biostec, pp. 308–315.

[32] A. Asuncion and D. Newman, "UCI Machine Learning Repository," 2007. [Online]. Available: https://archive.ics.uci.edu/ml/index.php.

[33] K. Potdar, T. S., and C. D., "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," Int. J. Comput. Appl., vol. 175, no. 4, pp. 7–9, 2017.

[34] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of Missing Data in Industrial Databases," Appl. Intell., vol. 11, no. 3, pp. 259–275, 1999.

[35] Z. E. Xu, K. Q. Weinberger, and A. X. Zheng, "Gradient Boosted Feature Selection Categories and Subject Descriptors," Kdd, pp. 522–531, 2014.

[36] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks," Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, 2009.

3D Trilateration Localization using RSSI in Indoor Environment

Nur Diana Rohmat Rose[1], Low Tan Jung[2]
Department of Computer and Information Sciences
Universiti Teknologi PETRONAS, Perak, Malaysia

Muneer Ahmad[3]
Faculty of Computer Science and Information Technology
Universiti Malaya, Kuala Lumpur, Malaysia

***Abstract*—Received Signal Strength Indicator (RSSI) is one of the most popular technique for outdoor and indoor localization. There are many previous researches on RSSI-based indoor localization systems. However, most of them lack a solid classification method that reduce localization errors with better accuracy. This paper will focus on indoor localization methods to provide a technological perspective of indoor positioning systems. This paper proposes an indoor localization by using 3D trilateration method to locate target tags from RFID readers that used RSSI measurements for range determination. There will be six test cases for each reader. This system can track any target within the selected area with less localization error.**

***Keywords*—RSSI; RFID; Indoor Positioning System (IPS); trilateration; 3D localization**

I. Introduction

Localization systems are a significant permissive technology nowadays and becoming one of the crucial part of our daily life. It can be divided into outdoor and indoor environment. Global Navigation Satellite System (GNSS) or Global Positioning System (GPS) is central to the various types of localization technologies used by a wide range of people and groups in our society. It is a satellite navigation systems that provide autonomous geospatial positioning with global coverage. GNSS is almost similar to GPS, but there are a few key differences between the two. However, the precise functions of GPS has its limits which it does not work well in indoor environment due to the attenuated and scattered signals from the satellites by roofs, floors, walls, and other objects [1]. In order to overcome this issues, Indoor Positioning System (IPS), a revolutionary technology is used to locate people or objects within indoor environment such as at underground locations, parking garages, and multistory buildings.

There are a wide range of IPS technologies used to track and locate people or objects in indoor environment such as Wireless Local Area Network (WLAN), Radio-Frequency Identification (RFID), ultrasonic, Bluetooth, and Ultra-Wideband (UWB) [2]. However, IPS technologies cannot utilize GNSS for location detection unlike outdoor positioning systems that can achieve 1-3 meter accuracies through satellites [3].

The ability to locate people accurately and quickly in on-fire buildings is quite critical to establish effective fire emergency response operations [4]. In a recent study [5], a series of interviews with the first responders assessed the value of the indoor location information. It was noticed that information on indoor location was among the most needed items of information when reacting to building fire incidents. Fire statistics show that 369,500 fires or 76% of all structure fires occurred in home structures (family homes and apartments), a slight increase of 1.2% were registered in the US in 2013, resulting in 2,755 civilian fire deaths [6]. Recently, more attention has been given to indoor positioning systems for multistory buildings. Such systems need not only to describe the coordinates (x, y) but also the floor where the object is located.

II. Indoor Localization

A. Methods

Most technologies nowadays are focusing on accurate real time object tracking and localization within buildings [7]. Thus, the rapid growth of indoor localization technologies has become an important demand for some markets. Researches on IPS are still widely conducted in order to improve the performance of localization technologies. A large variety of methods, techniques, principles, and devices are used to provide indoor positioning data in the form of signal strength or range. A basic indoor localization should have a reference sensor node with a known position and a target to be located [8].

Lateration/Trilateration/Multilateration. All these terms refer to a position determined by distance measurement. Lateration is the most common method to define an object’s position by measuring its distance from multiple reference points. It uses the known distance of at least three fixed points in 2D space, or four fixed points in 3D space to determine an object's position [8]. Trilateration works by finding a series of circles that intersect with each other. Techniques based on the measurement of the signal propagation such as Time of Arrival (ToA), Time of Flight (ToF), Time Difference of Arrival (TDoA), and RSSI are defined as lateration techniques [9].

Time of Arrival (ToA) or Time of Flight (ToF) is the amount of time a signal takes to travel from a transmitter to a receiver. Because the rate of signal propagation is constant and known, a signal’s travel time can be used to measure distance directly [10]. However, the accuracy of the TOA-based methods also suffers from significant multipath conditions in indoor environment caused by diffraction of the radio frequency (RF) signal from objects such as walls and floors. The distance from the reference point can be calculated using this simple equation:

$$d = c * (t_{arrive} - t_{sent}) \quad (1)$$

where c is the speed of light. In 2D, this leads to the following equation:

$$d = \sqrt{(x_{ref} - x)^2 + (y_{ref} - y)^2} \quad (2)$$

where (x_{ref}, y_{ref}) is the known position of the reference point. Once this set is calculated for enough reference points (at least three points for 2D or at least four points for 3D), the exact position of the target can be calculated by finding the intersection.

Time Difference of Arrival (TDoA) does not require the time that the signal was sent from the target, only the time the signal was received and the speed that the signal travels [11]. Once the signal is received at two reference points, the difference in arrival time can be used to calculate the difference in distances between the target and the two reference points. This difference can be calculated using the equation:

$$\Delta d = c * (\Delta t) \quad (3)$$

where c represent the speed of light and Δt is the difference in arrival times at each reference point. In 2D, this leads to the following equation:

$$\Delta d = \sqrt{(x_2 - x)^2 - (y_2 - y)^2} - \sqrt{(x_1 - x)^2 - (y_1 - y)^2} \quad (4)$$

where (x_1, y_1) and (x_2, y_2) are the known positions of the beacons.

Received Signal Strength Indicator (RSSI) measures the amount of power present in a radio signal that a RF client device receives from an access point or router. One way to determine the efficiency of a communication link is by measuring the signal strength at the receiving antenna. If a transmitter is moved closer to a receiver, the strength of the transmitted signal increases at the receiving antenna. Otherwise, the signal strength at the receiving antenna decreases if a transmitter is moved further away. RSSI is measured in dBm, with a higher negative value (in dBm) indicating a weaker signal [12].

Angulation/Triangulation involves measuring angles and is used to measure unknown distances. This can be done by establishing a baseline length. From each point in Fig. 1, angles of distant points can be measured. Triangulation determine the distances by forming triangles from the three points, based on the lengths and angles measured. The Angle of Arrival (AoA) technique can be classified under this method.

Angle of Arrival (AoA) estimates the target's location from the intersection of the several pairs of angle direction lines, each formed by the circular radius from a base station. This method requires only two measuring units for 2D and three measuring units for 3D. AoA does not require synchronization between the measuring units. Furthermore, it is based on direction, which received signal arrives from one device to another [13]. However, this method does not work efficiently in indoor environment since the accuracy and precision decreases when there are signal reflections from surrounding objects.

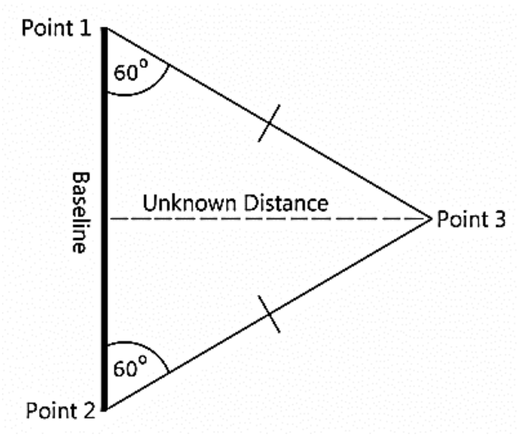


Fig. 1. Triangulation Localization Method.

Fingerprinting. Indoor environment produces many significant noises such as multipath fading, signal occlusions due to walls or objects, and signal diffractions depending on the object's material. Fingerprinting is the most popular localization method since it has higher accuracy compared to other methods. Most indoor localization approaches have adopted a matching fingerprint as the basic location determination scheme. The main idea is to collect scene features (fingerprint) from the surrounding signatures at each location in the areas of interest and then build a fingerprint database [7].

Proximity is the simplest method for localization. This localization method belongs to the range-free localization group. The assumption based on this method is, if the point is within the range of a known station, we can calculate the point's location to the known station. The location of user can be estimated to the position of access point for connection based on wireless communication. This method was utilized by the Cell-ID method standardized in GSM cellular systems, followed by RFID and Bluetooth-based systems [14].

Dead Reckoning (DR) is one of the useful method to overcome limitations where GPS or GNSS signals does not propagate very well within some environment such as underground passages, basements, or tunnels. When a vehicle pass through some environment where very strong multipath propagation occurs, GPS or GNSS signals cannot be received. By utilizing dead reckoning capabilities, the receiver will consistently output position with some hybrid positioning with GPS or GNSS signals. Information from various sensors are used to determine the current position which enables high accuracy localization. This method is widely used in automotive navigation systems.

B. Technologies

In this section, we present some representative examples and the description of indoor localization technologies which are Infrared Radiation (IR), Radio-Frequency Identification (RFID), Bluetooth, Wireless Local Area Network (WLAN), and Ultra-Wideband (UWB).

Infrared Radiation (IR) is an electromagnetic radiation (EMR) that is invisible to the human eye, although longer infrared waves can be sensed as heat. Infrared light, with wavelengths longer that the visible light can be used in wired or wireless operations depending on the situation [15]. Lower frequency light such as infrared light is usually used in fiber optic cables to transmit data due to the ability to travel farther down fiber optic lines with less power (intensity) required than higher frequency light. The main advantages of IR is it requires minimum power to operate and can be set up at a low cost [16]. Infrared transmission is also a secure way to transfer data between devices as the signal cannot pass beyond a room or chamber. However, the disadvantages of infrared are it can be used for a small range distance and the signals can be interpreted by objects, people, and impacted by weather conditions.

Radio-Frequency Identification (RFID) is a technology with the use of electromagnetic fields or radio waves to wirelessly identify and capture information of a tag attached to the object. There is a wide range of RFID applications in various industries such as automobile, manufacturing, retail, education, and healthcare. The advantages of RFID tags are it can be read without the line of sight, multiple tags can be read simultaneously, and the ability to read data without visual access. In spite of that, privacy is a main concern with the use of RFID on products as it can be easily tapped or intercepted and the external electromagnetic interference can limit the RFID remote reading [13].

Bluetooth is a short-range wireless communication technology that allows two different Bluetooth-enabled devices to connect by using low-energy radio waves in order to send the data [17]. Bluetooth can be considered as a wireless replacement of cables as common as Wi-Fi but with less power consumption and less implementation cost than Wi-Fi [18]. Bluetooth is used for voice and data transfer with a better range to be compared with Infrared since it is able to avoid interference from other wireless devices. However, there are some limitations of Bluetooth which it can lose connection in certain conditions and allow only short range communication between devices.

Wireless Local Area Network (WLAN) positioning refers to the process of allowing devices to connect and communicate through a WLAN infrastructure. A WLAN can provide a connection to the wider Internet through a gateway in order to achieve high flexibility for ad hoc communication. WLANs are based on IEEE 802.11 standards which refers to a family of specifications developed by the IEEE for WLAN technology [19]. WLAN positioning system use RSSI values from the Access Points (AP). Mobile devices will measure the signal strength receives from the APs. Nevertheless, APs are quite expensive to be compared with wires and hubs. In terms of flexibility, the nodes can communicate without further restriction. It is also a good alternative in terms of accuracy, precision, and cost since we can leverage on the existing infrastructure.

Ultra-Wideband (UWB) is an ultra-low power wireless technology for transmitting large amounts of digital data over a short distances. The UWB data rate is significantly higher than the Bluetooth and Wi-Fi technologies. The main advantage of UWB is the transmission using UWB are very secure due to the low power density that limits the interference potential with conventional radio systems. Furthermore, it is a promising indoor positioning technology which provides high accuracy since its bandwidth is very high that it can allow very high data throughput for communications devices [20]. It is also able to penetrate different type of materials easily. When there are obstacles, we can call this non-line-of-sight (NLOS). However, this technology requires higher initial implementation cost with slower adoption rate [21].

III. Methodology

A. Proposed Procedures

In this research, MATLAB software is used to evaluate the performance of the proposed localization method virtually based on the mathematical models in order to track and locate target tags within the selected area from RFID readers. The simulation can be categorized into two main phases which are 3D localization and multistory localization. Fig. 2 shows the phases in this research. However, this paper will discuss the simulation results of phase 1 only which is 3D localization.

Fig. 3 below shows RFID localization system block diagram. Signal strength of reference tags will be read from the RFID readers. In this research, the authors use the signal strength and reference position of readers from reference paper [22]. The RSSI values from RFID readers will be used for distance estimation between nodes by applying log-distance path loss model. To locate a target tag in 3D environment, trilateration localization algorithm will be applied. The target tag coordinates in x, y, and z plane will be located once we applied the localization algorithm. The localization accuracy can be determined by the analysis based on the number of positions that have been entered into the database.

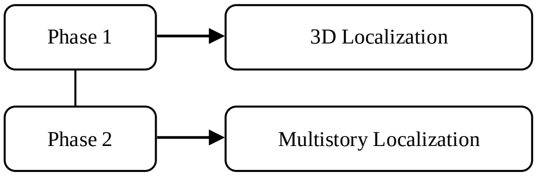


Fig. 2. Simulation Phases.

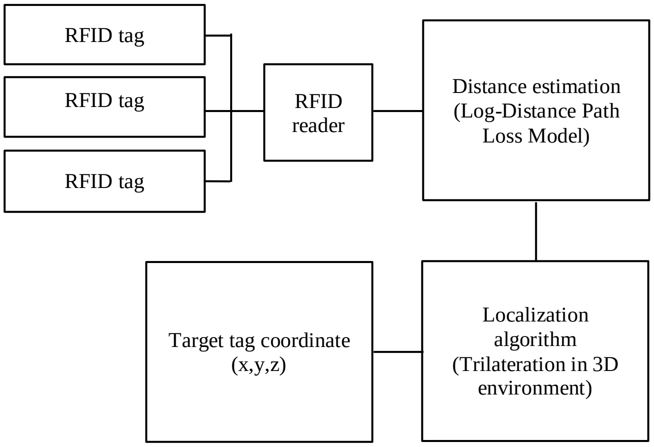

RFID tag
RFID tag
RFID tag
RFID reader
Distance estimation (Log-Distance Path Loss Model)
Localization algorithm (Trilateration in 3D environment)
Target tag coordinate (x,y,z)

Fig. 3. RFID Localization System Block Diagram.

The aim of this localization system is to estimate the position coordinate of a target node with the help of at least three reference nodes by using RSSI-based trilateration algorithm with an acceptable localization error. RSSI measures the signal power at the receiver. Based on the transmitted power, the propagation loss is calculated and the loss can be translated into distance estimation. Trilateration method helps to figure out the center point of three nodes surrounded by circles for which the distance and position is already known.

B. System Flowchart

The distances between readers and target tags can be calculated based on the RSSI values [23]. In this research, the authors use the signal strength and reference position of readers from reference paper [22] as shown in Table I. Factors of the RFID readers used in Table I are shown in the Table II below. Based on the RSSI values, we can estimate and calculate the distance between the nodes by using log-distance path loss model. Fig. 4 shows the system flowchart for this 3D trilateration localization.

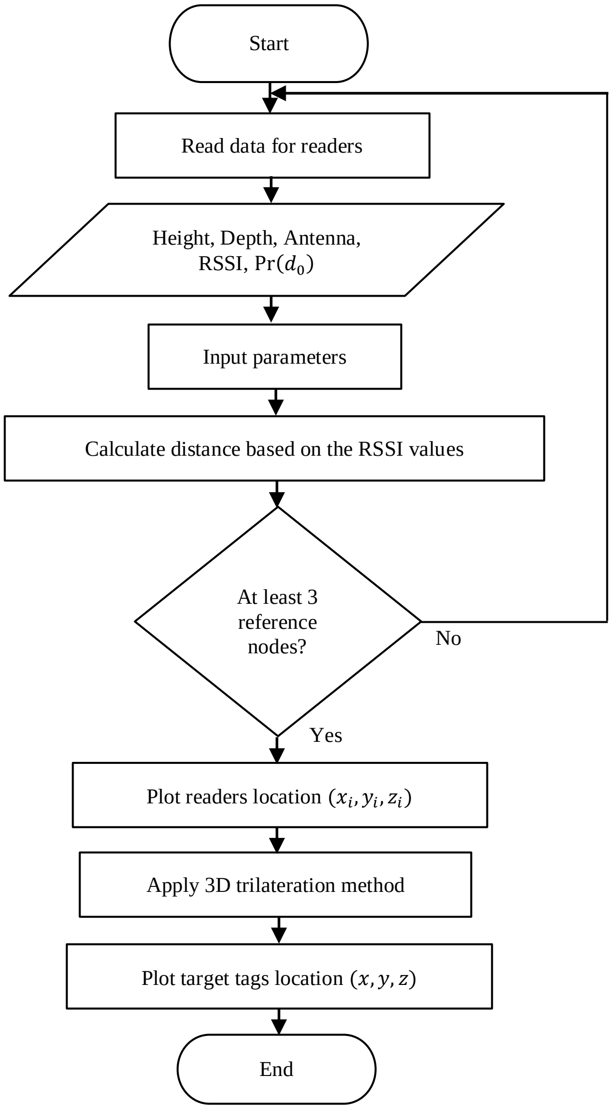

Start
Read data for readers
Height, Depth, Antenna, RSSI, $\Pr(d_0)$
Input parameters
Calculate distance based on the RSSI values
At least 3 reference nodes?
No
Yes
Plot readers location (x_i, y_i, z_i)
Apply 3D trilateration method
Plot target tags location (x, y, z)
End

Fig. 4. System Flowchart.

TABLE. I. RFID READERS WITH CORRESPONDING DATA

Reader	Height	Depth	Antenna	RSSI	$\Pr(d_0)$
1	1	1	1	-51	39.01
	1	3	3	-55	0
	3	1	1	-57	42.74
	1	3	1	-56	42.7
	3	1	3	-53	41.47
	3	3	3	-53	0
2	1	3	1	-60	41.47
	1	3	3	-53	0
	3	1	1	-63	42.74
	3	1	3	-56	42.7
	1	1	1	-53	44.16
	3	3	3	-52	0
3	3	1	3	-55	40.37
	1	3	3	-56	0
	1	1	1	-52	42.73
	3	1	1	-62	42.7
	1	3	1	-55	44.19
	3	3	3	-55	0
4	1	3	3	-57	40.37
	3	3	3	-58	0
	1	3	1	-61	38.46
	1	1	1	-54	43.9
	3	1	1	-60	44.19
	3	1	3	-53	0

TABLE. II. FACTORS OF RFID READERS

Factor	Symbol	Level	Description
Height	1	0.4m	Close to the ground
	2	1.2m	Far away from either the ground or the ceiling
	3	2m	Close to the ceiling
Depth	1	0.2m	Close to the wall
	2	0.6m	Neither too close nor too far away to the wall
	3	1m	Far away from the wall
Antenna	1	[\ /]	Antenna A and B are both in the vertical plane with 45 degrees to the ground and 90 degrees in between
	2	[|_]	Antenna A and B are both in the vertical plane with A in the vertical direction and B in the horizontal direction
	3	[. |]	Antenna A is in the horizontal plane and pointing straight forward, while antenna B is in the vertical plane and in the vertical direction.

In order to calculate the distance between the nodes, the authors use the log-distance path loss model or known as penetration thru buildings model based on the RSSI values collected before [24]. Average received signal power decreases logarithmically with distance, whether in outdoor or indoor environment. Log-distance path loss model is a general propagation model. It can be used in both indoor and outdoor environment. The distance can be calculated using the equation:

$$d = d_0 * \exp\left(\frac{\Pr(d_0)-\Pr(d)}{10n}\right) \tag{5}$$

where d_0 is the near earth reference distance which we used its value as 1 meter for indoor environment, $\Pr(d_0)$ represent received power at reference distance d_0 and n is the path loss exponent which its value is between 2 to 6 depending on the environment.

C. 3D Trilateration Method

The distance calculated in part *B* will be used to locate target tags location by trilateration. Trilateration localization algorithm will be applied in order to get the location coordinate output on the x, y, and z plane or in other words, in 3D localization. The quadratic equation for this trilateration localization algorithm is as follows:

$$(x-x_1)^2 + (y-y_1)^2 + (z-z_1)^2 = {d_1}^2 \tag{6}$$

$$(x-x_2)^2 + (y-y_2)^2 + (z-z_2)^2 = {d_2}^2 \tag{7}$$

$$(x-x_3)^2 + (y-y_3)^2 + (z-z_3)^2 = {d_3}^2 \tag{8}$$

$$(x-x_4)^2 + (y-y_4)^2 + (z-z_4)^2 = {d_4}^2 \tag{9}$$

where (x, y, z) is the estimated target tag coordinate, (x_i, y_i, z_i) is the position of the readers and d_i is the calculated distance from target tag to the four readers. Since the equations contain three unknowns and four equations, three equations are used to calculate unknowns. Fourth equation is taken as a reference to select the correct pair of variables.

D. Simulation Setup

In this research, simulation by using MATLAB was used for RFID readers to locate any target tag within the selected area based on the RSSI values. The simulation was carried out in a selected area with the dimension of 3m x 3m x 3m. There will be four RFID readers located in the area to locate or track any target tag in order to fulfil the objective of this simulation. Distance estimation from the four nodes to the target node is calculated by using the log-distance path loss model as shown in equation (5). In order to plot the location of the nodes in 3D localization, we will need to have at least three reference nodes. Trilateration localization algorithm as shown in equation (6), (7), (8), and (9) will be applied in order to get the location coordinate output of the estimated target tag in 3D localization.

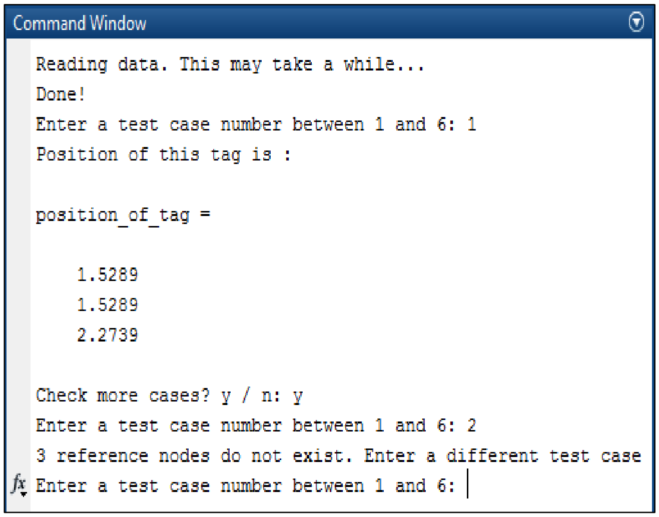
IV. Discussion

Each red dot in the figures below symbolized the RFID readers. The position of the four RFID readers are located at (1,1,1), (1,3,1), (3,1,3) and (1,3,3). The location of the target tag shown by a green dot within the red dots (RFID readers). There will be six test cases for each RFID reader. The position of the target tag for test case 1 is (1.5289,1.5289,2.2739) as shown in Fig. 5. The green dot in Fig. 6 below showing position of the target tag.

For test case 2 and test case 6, three reference nodes does not exist due to the path loss exponent. In the study of wireless communications, path loss can be represented by the path loss exponent, whose value is normally in the range of 2 to 4. In some environments, such as buildings, stadiums and other indoor environments, the path loss exponent can reach values in the range of 4 to 6. On the other hand, a tunnel may act as a waveguide, resulting in a path loss exponent less than 2. Path loss is usually expressed in dB. So, we can conclude that the value of path loss exponent is between 2 to 6 depending on the environment.

For test case 3, there are two target tags located at (1.5289,1.5289,2.2739) and (1.99229,3.1260,2.0074) as shown in Fig. 7.

Fig. 8 shows the three target tags located at (1.5289,1.5289,2.2739), (1.99229,3.1260,2.0074) and (3.3015,3.3122,2.0000) for test case 4.



```
Reading data. This may take a while...
Done!
Enter a test case number between 1 and 6: 1
Position of this tag is :

position_of_tag =

    1.5289
    1.5289
    2.2739

Check more cases? y / n: y
Enter a test case number between 1 and 6: 2
3 reference nodes do not exist. Enter a different test case
fx Enter a test case number between 1 and 6:
```

Fig. 5. Command Window for Test Case 1 and 2.

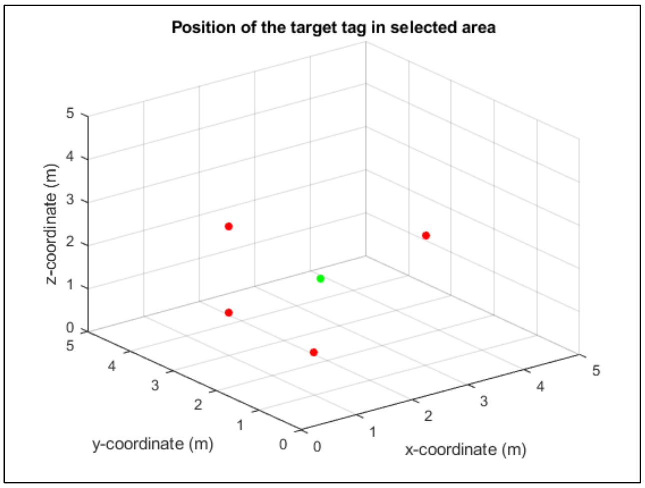


Fig. 6. 3D Trilateration with One Target Tag.

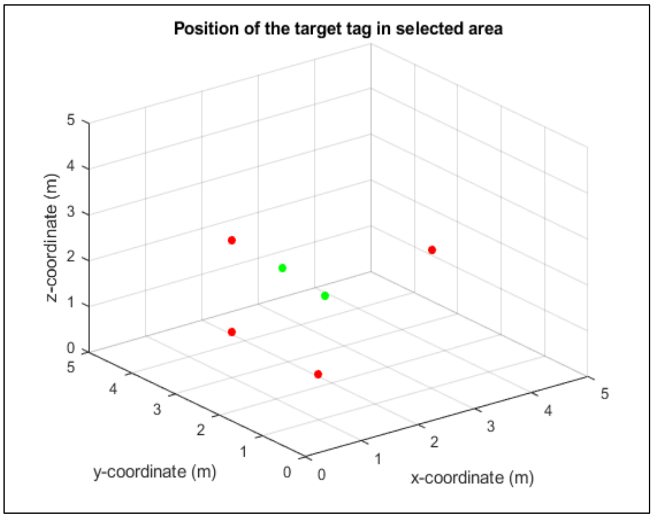


Fig. 7. 3D Trilateration with Two Target Tags.

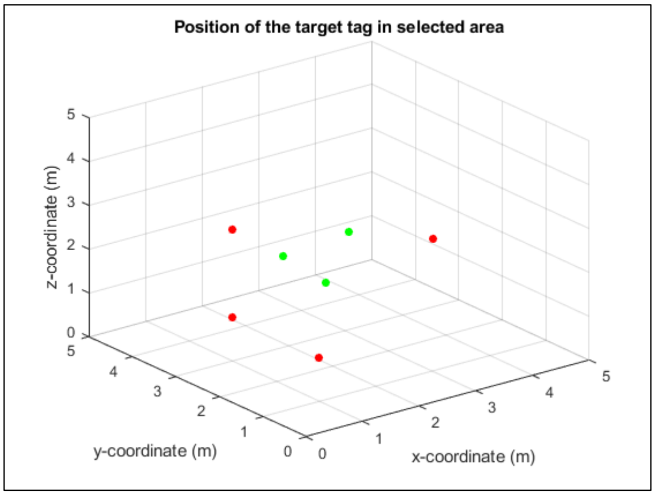


Fig. 8. 3D Trilateration with Three Target Tags.

Fig. 9 shows the location of four target tags for test case 5 at (1.5289,1.5289,2.2739), (1.99229,3.1260,2.0074), (3.3015,3.3122,2.0000) and (1.9383,1.9383,4.4331).

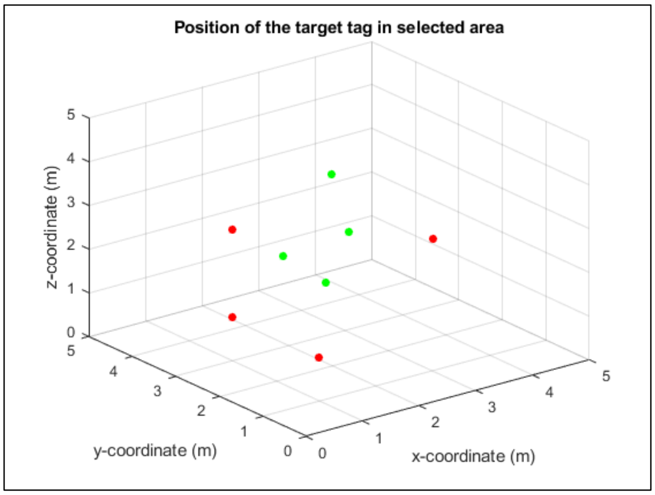


Fig. 9. 3D Trilateration with Four Target Tags.

This 3D trilateration localization simulation will be extended into the next phase which is multistory localization. The comparison between these two will be used in order to measure the accuracy of 3D indoor localization in multistory environment.

V. CONCLUSION

IPS use sensors and communication technologies to locate objects in indoor environment. IPS are attracting scientific and enterprise interest because there is a big market opportunity for applying these technologies. In this paper, trilateration method is used for 3D localization to locate target tag in the selected area accurately by using RSSI values. RSSI is a measure of the strength of a signal received from a tag. From the six test cases, we can conclude that RSSI signals is important to determine how close a target tag is to a reader antenna and to consider the real-time operation of sensor nodes. The RSSI values varies depending on the distance of the tag from the reader's antenna. There are several other factors to be considered when utilizing RSSI which are the orientation of the tag with respect to the reader's antenna, the material of the tag and blocking objects between the tag and the reader.

VI. FUTURE WORK

Current surveillance systems for multistory building lack in supporting real-time monitoring of every point of a level in multistory buildings at all times. Solutions using wireless sensor networks, on the other hand, can gather sensory data values from all points of a building continuously, day and night to provide fresh and accurate data. However, sensor networks face serious obstacle which is lack of accurate indoor localization for multistory buildings. Recommendation for further studies in this area is to propose the best method for indoor localization as well as to prove that the method can determine accurate indoor localization for multistory building. While continuously preserving the goal of building monitoring, it is also to construct a system that shall regards the lack of accurate localization that may hinder the network performance.

ACKNOWLEDGMENT

The authors would like to thank the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS for providing the information, guidance, facilities and financial support through Graduate Assistantship (GA) Scheme throughout this research activities.

REFERENCES

[1] M. Er Rida, F. Liu, Y. Jadi, A. A. A. Algawhari, and A. Askourih, "Indoor location position based on bluetooth signal strength," Proc. - 2015 2nd Int. Conf. Inf. Sci. Control Eng. ICISCE 2015, pp. 769–773, 2015.

[2] T. Kim and E. J. Kim, "A novel 3D indoor localization scheme using virtual access point," Int. J. Distrib. Sens. Networks, vol. 2014, pp. 6–11, 2014.

[3] L. Batistic and M. Tomic, "Overview of indoor positioning system technologies," 2018 41st Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2018 - Proc., pp. 473–478, 2018.

[4] N. Li, B. Becerik-Gerber, B. Krishnamachari, and L. Soibelman, "A BIM centered indoor localization algorithm to support building fire

emergency response operations," Autom. Constr., vol. 42, pp. 78–89, 2014.

[5] N. Li, Z. Yang, A. Ghahramani, B. Becerik-Gerber, and L. Soibelman, "Situational awareness for supporting building fire emergency response: Information needs, information sources, and implementation requirements," Fire Saf. J., vol. 63, pp. 17–28, 2014.

[6] M. J. Kenter, "Fire loss in the United States during 1989," Fire J. Boston, Mass., vol. 84, no. 5, 1990.

[7] Z. Farid, R. Nordin, and M. Ismail, "Recent advances in wireless indoor localization techniques and system," J. Comput. Networks Commun., vol. 2013, 2013.

[8] M. Md Din, N. Jamil, J. Maniam, and M. A. Mohamed, "Review of indoor localization techniques," Int. J. Eng. Technol., vol. 7, no. 2.14, p. 201, 2018.

[9] D. Zhang, F. Xia, Z. Yang, L. Yao, and W. Zhao, "Localization technologies for indoor human tracking," 2010 5th Int. Conf. Futur. Inf. Technol. Futur. 2010 - Proc., no. 60903153, pp. 1–6, 2010.

[10] P. Shirke, A. Potgantwar, and V. M. Wadhai, "Analysis of RFID Based Positioning Technique Using Received Signal Strength and Directional Antenna Global Positioning System (GPS), Indoor Positioning System (IPS), Radio Frequency Identification (RFID), Received Signal Strength (RSS)," vol. 7, no. May, pp. 80–89, 2016.

[11] G. Kul, T. Özyer, and B. Tavli, "IEEE 802.11 WLAN based real time indoor positioning: Literature survey and experimental investigations," Procedia Comput. Sci., vol. 34, no. August, pp. 157–164, 2014.

[12] N. Rosli et al., "Jurnal Teknologi," vol. 1, pp. 1–6, 2015.

[13] J. Kárník and J. Streit, "Summary of available indoor location techniques," IFAC-PapersOnLine, vol. 49, no. 25, pp. 311–317, 2016.

[14] C. Laoudias, A. Moreira, S. Kim, S. Lee, L. Wirola, and C. Fischione, "A survey of enabling technologies for network localization, tracking, and navigation," IEEE Commun. Surv. Tutorials, vol. 20, no. 4, pp. 3607–3644, 2018.

[15] 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," J. Sensors, vol. 2017, 2017.

[16] S. Ma, Q. Liu, and H. Tang, "An Overview of Location Semantics Technologies and Applications," Int. J. Semant. Comput., vol. 9, no. 3, pp. 373–393, 2015.

[17] I. Oksar, "A Bluetooth signal strength based indoor localization method," Int. Conf. Syst. Signals, Image Process., no. May, pp. 251–254, 2014.

[18] Y. Zhuang, J. Yang, Y. Li, L. Qi, and N. El-Sheimy, "Smartphone-based indoor localization with bluetooth low energy beacons," Sensors (Switzerland), vol. 16, no. 5, pp. 1–20, 2016.

[19] S. Dhanalakshmi and M. Sathiya, "An Overview of IEEE 802.11 Wireless LAN Technologies," Int. J. Comput. Sci. Mob. Comput., vol. 4, no. 1, pp. 85–93, 2015.

[20] A. K. Paul and T. Sato, "Localization in wireless sensor networks: A survey on algorithms, measurement techniques, applications and challenges," J. Sens. Actuator Networks, vol. 6, no. 4, 2017.

[21] J. Wu, "DigitalCommons @ University of Nebraska - Lincoln Three-Dimensional Indoor RFID Localization System," 2012.

[22] J. Wu, "Three-Dimensional Indoor RFID Localization System," Univ. Nebraska -Lincoln, p. 199, 2012.

[23] O. G. Adewumi, K. Djouani, and A. M. Kurien, "RSSI based indoor and outdoor distance estimation for localization in WSN," Proc. IEEE Int. Conf. Ind. Technol., pp. 1534–1539, 2013.

[24] M. Region, "RSSI based Trilateral Indoor Localization System using Reference Tags of Radio Frequency Identification System," vol. 07, no. 02, pp. 276–279, 2018.

Cloud based Power Failure Sensing and Management Model for the Electricity Grid in Developing Countries: A Case of Zambia

Janet Nanyangwe Sinkala[1]
Department of Electrical and Electronics Engineering
University of Zambia
Lusaka, Zambia

Jackson Phiri[2]
Department of Computer Science
University of Zambia
Lusaka, Zambia

Abstract—**In most developing countries, huge parts of the electric power grid are not monitored making it difficult for the service provider to determine when there is a power failure in the electric grid, especially if the power failure occurs in the Low Voltage level. Clients usually have to call and inform the utility's customer service centre to report a power failure. However, this system of addressing power outages is not very effective and usually results in long durations of system interruptions. This paper proposes a cloud based power failure sensing system to enable automatic power failure sensing and reporting as well as monitoring of the low voltage power network in Zambia, a developing country in Southern Africa. A baseline study was conducted to determine the challenges faced by both the electric power utility company called Zambia Electricity Supply Corporation (ZESCO) and the electricity consumers in the current power failure reporting management model. The results from the baseline study indicate that challenges are being faced by electricity consumers when it comes to reporting power failures. These include failure to get through to the customer call centre due to constantly engaged lines, unanswered calls, failed calls and network failure. The challenge faced by the electricity service provider is the inability to attend to all the customers through the call centre as customer calls are rejected due to limited Call Centre system resources. To address these challenges the proposed cloud based power failure sensor model made use of a Voltage sensor circuit, Arduino Microcontroller board, SIM808 GSM/GPRS/GPS module, cloud architecture, Web Application and Google Map API. Results from the proposed model show improved reporting time, location information and quick response to power failures.**

Keywords*—*Cloud architecture; power failure sensing; low voltage network; electric grid

I. Introduction

Electrical power systems are extremely huge and complex networks. Such electric power systems provide a secondary source of energy essential to meeting man's needs, improve living standards and boost socio-economic development [1] [2]. These systems are integrated to provide economic benefits, increased reliability, operational advantages making them the most important national and global infrastructure such that when they collapse it leads to major direct and indirect impacts on the economy and national security [1]. A power system is usually made of many components such as transmission lines, substations, switches and transformers with widely dispersed power generating systems and loads being the main features. A power system consists of three subsystems being generation, transmission and distribution. According to [5] the distribution network is a major element of the total electrical supply scheme, providing the final link between the bulk transmission systems and the customers with 80% of the outages that occur at the customer electricity service being due to failures in the distribution network.

Urbanization and computerization of society has increased demand for reliable electricity supply. Power outages also referred to as power failures cause an interruption to this urbanization and computerization resulting in welfare losses, slow economic growth and low productivity for firms [3] [4]. High-speed wind, flying objects, falling trees, physical contact by animals, lightning, snow storms, contamination of insulators, human errors, overloads, bad insulation and protection failure are some of the causes of power failures [4].

The aim of most power utility companies is to ensure reduced economic losses, lost productivity, and customer inconvenience brought about by power failures. To mitigate the customer interruption costs due to power failures, distribution systems need a distribution automation system for power failure detection [4] [6]. Automation of the power failure management model is critical for detecting a power failure and its location to enable rapid restoration of supply. Location of the geographical position of the power failure is important to keep the stability of the power system after fault detection. In [4] it is pointed out that the reduced number of customers interrupted and the associated customer minutes of interruptions are the primary major benefits of automated power failure sensing systems.

The electricity industry faces constant power failures, which require an effective and modern way of power failure and fault management. This study is an attempt to develop a power failure sensing and management model for the electricity grid in Zambia. This model will be cloud based.

II. Background

The Zambia electricity grid is an interconnected system of electric transmission lines linking generators to loads and comprises transmission lines and substations at 330 kV, 220 kV, 132 kV, 88kV, 66 kV, 11kV, 33kV and 0.4kV voltage levels [7]. The backbone of the grid is built on a robust 330 kV system from the southern part of the country through Lusaka and Central provinces to the Copper belt.

Due to the favorable economic development, the demand for electricity in Zambia has been increasing at average annual rates of 3-4 percent in recent years [8]. The Zambian government has projected an increase in the rural electrification rate from the current 2 percent to 50 percent by the year 2050, while urban electrification rate has been projected to increase from the current 48 to 90 percent by 2030 [8]. On 27th July 2019, the Zambia electricity service provider ZESCO connected the one millionth customer to the national grid. The corporation had witnessed growth of about 400 % between 2000 and 2019, with the customer base increasing from 200,000 in the year 2000, to about 900,000 in the year 2018 and consequently 1,000,000 customers in 2019 [9] .With Zambia being part of the sub Saharan African countries which have a total duration of outages averaging approximately 800 hours a year [10], it is important that reliable and efficient automatic power failure-sensing systems be put in place to reduce on outage durations as well as ensure customer satisfaction. This will ensure adequate, reliable and efficient electricity supply.

Power failure Remote sensing is necessary for acquisition of data from anywhere without the need for physical visits. It relies on sensory objects to sense and collect data from remote in real time and relay that information to a central place.

The degree to which the electric-powered technology has become embedded in all human activities makes the security and reliability of the electrical energy infrastructure of vital importance today more than ever.

III. Literature Review

Related work carried out includes research on the use of sensors to sense various parameters such as substation and transformer measurements, street lighting and condition monitoring of equipment.

Researchers in [11] propose a Global System for Mobile Communications (GSM) cloud based street light control and fault detection system. They use a Wi-Fi module for sending faulty light alert messages to the cloud so that information can be captured anytime and anywhere. A GSM module is used to send an alert SMS to a mobile phone and Arduino microcontroller is used to sense and control the streetlights. A light dependent resistor detects whether it is light or dark and switches the lights on and off as well as determines which light is faulty. The study in [12] propose and develop an internet of things (IOT) based prototype model using the APC220 transceiver, GSM, General Packet Radio Services (GPRS), radio-frequency identification (RFID), passive infrared sensor (PIR), Arduino microcontroller and cloud storage to curb theft of grain at the storage points of the food reserve agency(FRA). PIR sensors are used to sense motion and send a logical signal to the microcontroller which together with the GSM/GPRS wireless module and RFID is used to send alerts to the cloud and track bags of grain. They concluded that once this technology is adopted, theft will be reduced and grain management in the FRA satellite depots will improve. However there was no provision for locating the satellite depots in their prototype. In [13] a base line study was conducted on the challenges faced by Zambia Air Force (ZAF) in inventory management of spares and it was discovered that the major challenge was due to the manual inventory management which resulted into incorrect inventory reporting and pilferage of items. To address this challenge they propose a web based inventory management system using cloud architecture and barcode technology. The prototype application developed consists of the backend and the frontend components and users are created and managed by the system administrator in order to keep track of their activities. A barcode reader is incorporated to scan the barcode on the items. The barcode scan captures the barcode as well as other details on the scanned item which are then saved to the database. The developed prototype proved to be faster, efficient and more reliable than the manual and paper based system. The researchers however did not provide means of alerting ZAF through either a cloud application or an SMS to a mobile phone when spares are being pilfered.

In [14] a prototype based on remote sensor network including cloud and internet of things to aid the Food Reserve Agency in analytics, timely action and real-time reporting from all its food depots in Zambia is proposed. A baseline study was conducted which identified the challenges FRA faced, such as manual report generation, no connectivity to remote warehouses, inability to track stock on demand, theft and spoilage of stock due to lack of environmental monitoring. The proposed prototype was made up of Raspberry Pi microcontroller, temperature and humidity sensors, motion sensors, Global Positioning System (GPS) sensor, ZigBee transceivers and Wi-Fi access. Through these devices they were able to monitor temperature, humidity, location and motion via the cloud application. They concluded that modern warehousing relying on components such as sensors provide better grain storage, management, transparency of operations and hence lead to cost effective grain marketing which leads to better national food security. However the researchers concentrated on sending alerts using Wi-Fi which covers distances of about 100metre and did not look at how GPRS can be used for the same purpose. The researchers in [15] use GSM technology for detection and monitoring of transmission power line faults. The system is able to send an SMS to the utility and the utility has the ability to set current limits on the system. A Programmable Interface Controller (PIC) microcontroller is used to sense the current, voltage and frequency of the system. The PIC microcontroller is also able to detect short circuit limits by comparing the current sensed and the preset limit. When the preset limit is crossed the microcontroller sends a signal for tripping the system and an SMS alert is sent via the GSM network. Bidirectional communication was achieved making setting of the short circuit current limit from a mobile phone possible. The researchers however did not look at how cloud application can further assist in monitoring the transmission line parameters nor did they explore the sending of alerts together with the location of the fault.

In [16] [17] researchers propose a GSM based system for transmission and distribution fault detection. A microcontroller is used for sensing the voltage, current and frequency. This is then reported to a mobile and presented on a computer through serial RS232 communication. The researchers however do not explore the use of cloud technology through the use of GPRS nor do they explore the benefits of the use of location maps for locating the faults detected. The study in [18] [19] propose the use of GSM technology to monitor substation parameters such as voltage and current (over voltage, under voltage, over current) and send this information over to the operator mobile for further action. An SMS through GSM technology is used to send the change in status of the parameters to the operators and operators can send an SMS to read the parameters of the substation. They use a PIC microcontroller for sensing the substation parameters. The researchers however do not explore the use of cloud technology, GPRS nor do they explore the benefits of the use of location maps for locating the faults detected. The researchers in [20] use the combined GSM, SMS and Atmega16 microcontroller system to monitor transformer parameters such as oil level, temperature and load current. In [21] the researcher goes further to connect the microcontroller to a computer using RS232 to enable monitoring of the transformer condition. However the researchers did not explore the use of cloud services in monitoring the parameters.

From the proposed and implemented projects reviewed from literature, it has been observed that there is substantial potential to be tapped from the use of sensing and cloud technology in power systems and other sectors. It can also be noted that there exist a number of research gaps in the sense that there were no experiments or projects which made use of cloud and sensor technology applied in three phase power failure management and monitoring enabling alert, location, status, duration and measurement information to be sent to both the mobile phone and cloud services for storage and for location determination on location maps.

IV. Methodology

This section gives a description of the research process and an explanation of the methods used to gather data. The research involved conducting a baseline study whose results where then used to develop a power failure reporting management model and a prototype based on the model.

A. Base Line Study Methodology

1) Research design: This study employed survey, exploratory and Water fall system development research designs. Exploratory research design is a research design described as the problem-finding phase of research wherein the researcher focuses on the scope of study but with anticipation of arising problems at a later stage of the study [22]. Further, Survey research design is described as a type of research which is used to give a wider picture or an overview, assessing opinions, trends, beliefs and feelings of selected groups of individuals [23]. Exploratory and survey research was used to determine the challenges faced by both ZESCO the Service provider and the electricity clients in reporting and addressing power failure. Water fall system development research design was used for the development of a prototype for sensing and reporting power failure.

2) Data collection and sources: This study involved collecting information from both primary and secondary data sources. Secondary data sources included review of published reports on the subject such as books, essential Journals, conference papers, published academic papers and system documentation. Primary data sources on the other hand included questionnaire interviews with the electricity service provider and electricity clients.

3) Study population: The population target for this study was employees from ZESCO who work in the Call Centre customer service department, Fault Co-coordinators as well as electricity clients in Lusaka district.

4) Sampling technique: The study selected the people who use electricity in Lusaka district townships, ZESCO employees in the customer service department as well as fault coordinators. Questions in the data collection tools were both open and close-ended. This enabled the study extract both quantitative and qualitative responses. A Purposive Sampling method was employed in the selection of experts, as the interest was to target those with knowledge of the current power failure reporting management model. However, in selecting the electricity consumers, a multistage cluster sampling method was employed in which Lusaka district townships were selected using simple random sampling. Following that selection, convenience sampling was then applied in which households were selected for questionnaire distribution. Krejcie and Morgan sample selection technique guided the sample size for this study, which follows a curve of relational values between population and corresponding sample values at assumed standard error of 0.05. The sample size was determined to be 383.

5) Data analysis: Primary data collected from the respondents was entered and analyzed both qualitatively and quantitatively using Statistical Package for Social Scientists. Quantitative analysis was done using frequencies and percentages, presenting them in tables and bar charts. The qualitative data as well as secondary data analysis was carried out by use of content analysis.

B. System Design

The system requirements specification and model design phase of the research study employed the use of quantitative and qualitative data from interviews conducted with ZESCO personnel as well as data collected from electricity clients in Lusaka District through questionnaires. The quantitative and qualitative data from ZESCO personnel and electricity clients provided the information needed to come up with the current business process and thereafter develop the prototype based on cloud and sensing technology.

1) Current business process for power failure reporting: The current business process for power failure reporting is as shown in Fig. 1. It is derived from the baseline study that was conducted.

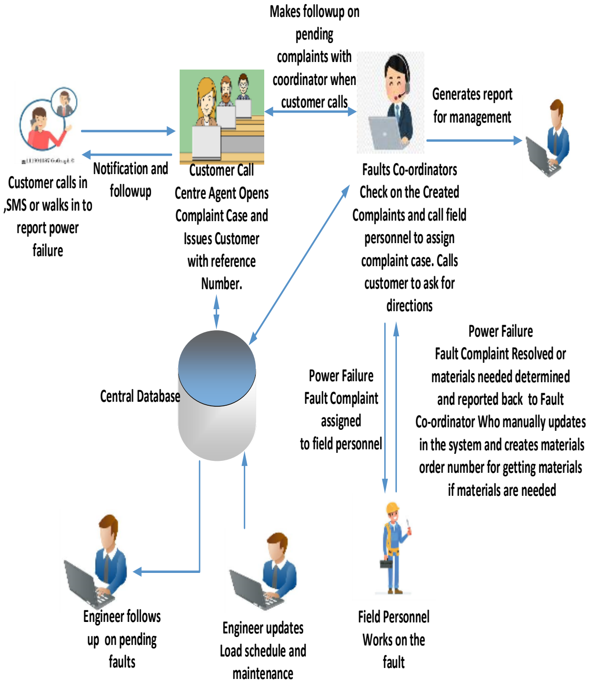

Makes followup on pending complaints with coordinator when customer calls
Generates report for management
Customer calls in ,SMS or walks in to report power failure
Notification and followup
Customer Call Centre Agent Opens Complaint Case and Issues Customer with reference Number.
Faults Co-ordinators Check on the Created Complaints and call field personnel to assign complaint case. Calls customer to ask for directions
Central Database
Power Failure Fault Complaint assigned to field personnel
Power Failure Fault Complaint Resolved or materials needed determined and reported back to Fault Co-ordinator Who manually updates in the system and creates materials order number for getting materials if materials are needed
Engineer follows up on pending faults
Engineer updates Load schedule and maintenance
Field Personnel Works on the fault

Fig. 1. Current Business Model for Power Failure Reporting.

The data collected from the service provider indicate that in the current power failure reporting management business model customers need to either call, SMS or walk into a customer call Centre to report a power failure fault. The customer is then given a complaint number to use for complaint case follow-up. This customer complaint case is then captured by the faults coordinator who calls and assigns the reported power failure fault case to the field personnel. In case the power failure location is not known, the fault coordinator calls the customer to get the location of the power failure. Engineers and managers also have access to the power failure information but only when they are connected to the ZESCO Local Area network (LAN).

2) Proposed business model for power failure reporting: Based on the results obtained from the baseline study a power failure reporting management business model based on sensor and cloud technology was proposed to address the challenges ZESCO and the clients are facing when making power failure reports. Fig. 2 shows the proposed model.

The proposed model was designed to make use of a microcontroller based electronic device installed on a distribution electricity line to automatically detect and report power failure and location to ZESCO personnel through cloud and mobile services. This eliminates the need for the customer to report the power failure thus addressing the challenges of call Centre access revealed in the baseline study. This model is intended to ensure that the overall response time of ZESCO personnel to power failures is improved as well as enable access to information regarding the status of the Low Voltage electricity network by anybody, anywhere and anytime through cloud services even when not connected to the ZESCO LAN.

3) Use case diagram: The Use Case diagram describes the proposed functionality of the new system. Use cases represent how a system interacts with its environment by illustrating the activities that are performed by the users of the system and the system's responses [24]. Use case diagrams model the functionality of a system using actors and use cases. The web application for the cloud platform was designed using a use case diagram. For this study the use case diagram consists of seven actors namely the Call Centre Agent, Faults Coordinator, Manager, Engineer, Customer, Field personnel and power failure detection system as shown in Fig. 3. The power failure detection system updates the status, location and duration of power failure. The call centre agent is able to view power failure faults and notify customers of the failure. The faults coordinator and engineer are capable of updating faults, generating materials order number and assigning faults whilst the field personnel is able to view faults, update faults, indicate the fault problem and materials needed. The cloud based nature of the model enables the customer to be able to login from anywhere and view areas affected by the power failure. The manager is able to view power failure statistics and generate reports. The web application was built on an open source cloud services provider which offers a platform for development of microprocessor based systems. This cloud based platform is built on Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), and Java Script as client software with ASP.net and Microsoft SQL database as server software. Google API was used for accessing Google Maps to determine location of the power failure.

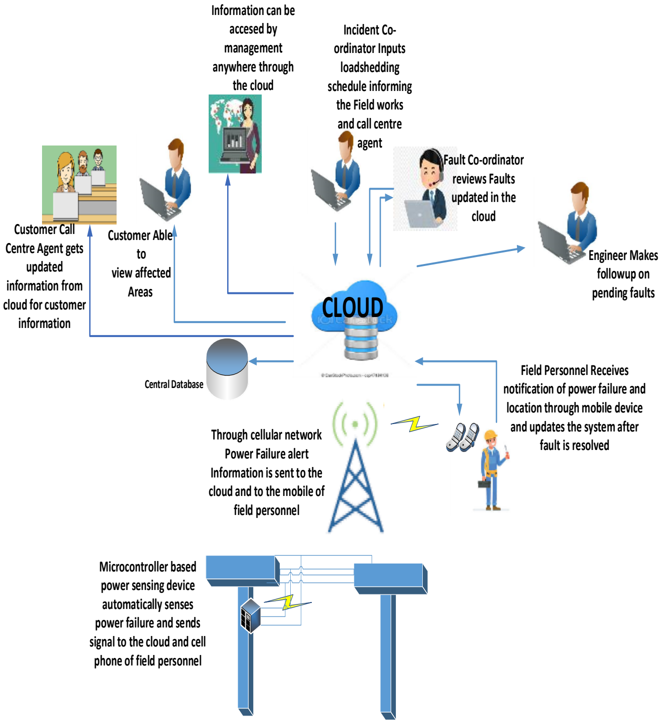

Information can be accesed by management anywhere through the cloud
Incident Co-ordinator Inputs loadshedding schedule informing the Field works and call centre agent
Fault Co-ordinator reviews Faults updated in the cloud
Customer Call Centre Agent gets updated information from cloud for customer information
Customer Able to view affected Areas
Engineer Makes followup on pending faults
CLOUD
Central Database
Field Personnel Receives notification of power failure and location through mobile device and updates the system after fault is resolved
Through cellular network Power Failure alert Information is sent to the cloud and to the mobile of field personnel
Microcontroller based power sensing device automatically senses power failure and sends signal to the cloud and cell phone of field personnel

Fig. 2. Proposed Model for Power Failure Reporting.

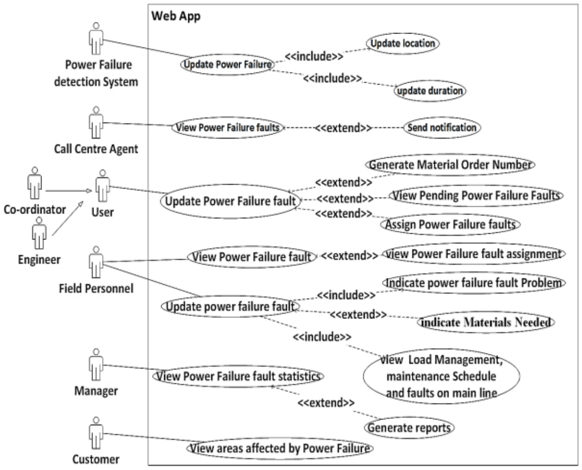

Web App
Power Failure detection System
Update Power Failure
<<include>>
Update location
<<include>>
update duration
Call Centre Agent
View Power Failure faults
<<extend>>
Send notification
Co-ordinator
User
Engineer
Update Power Failure fault
<<extend>>
Generate Material Order Number
<<extend>>
View Pending Power Failure Faults
<<extend>>
Assign Power Failure faults
Field Personnel
View Power Failure fault
<<extend>>
view Power Failure fault assignment
Update power failure fault
<<include>>
Indicate power failure fault Problem
<<extend>>
indicate Materials Needed
<<include>>
view Load Management, maintenance Schedule and faults on main line
Manager
View Power Failure fault statistics
<<extend>>
Generate reports
Customer
View areas affected by Power Failure

Fig. 3. Use Case Diagram for the Web Application.

C. Prototype Development

From the proposed model a prototype was developed which consisted of an electronic device made from a Voltage sensor circuit, Arduino Microcontroller board, Atmega 328 microprocessor, GSM/GPRS/GPS module, battery charge control circuit and Web Cloud services using cloud architecture, Web Application and Google Map API. Fig. 4 shows the block diagram of the Arduino Microcontroller based power sensing device. Fig. 5 shows the flow chart of the Power failure sensing functionality.

1) Prototype implementation materials and methods: The software used in prototype implementation was

- Arduino Integrated development environment (IDE) for programing the Arduino microcontroller board
- Java script
- HTML
- CSS
- ASP.net
- Microsoft SQL database
- Arduino Development board API

The hardware used was

- Arduino Development board
- Atmega 328 microprocessor
- MAX 232 Chip
- SIM808 GSM/GPRS/GPS Module
- Electronic components
- 12 Volts battery
- Liquid-crystal display (LCD)
- Laptop for accessing cloud services
- Mobile phone for receiving power failure alert notification

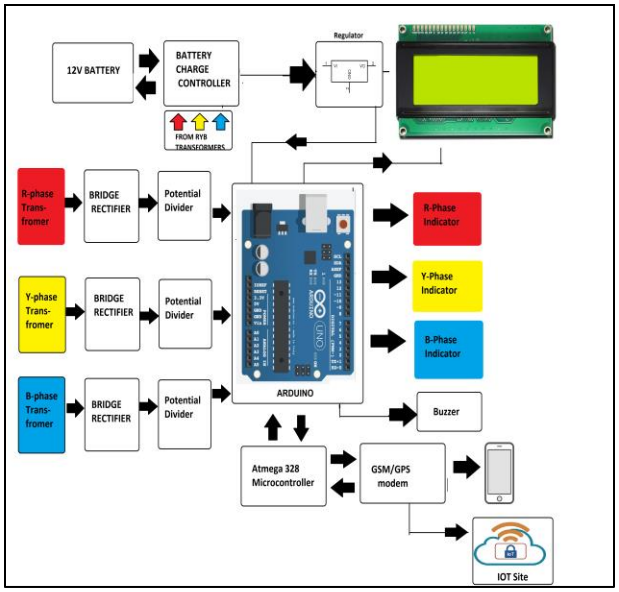


Fig. 4. Block Diagram of Arduino Microcontroller based Power Sensing Device.

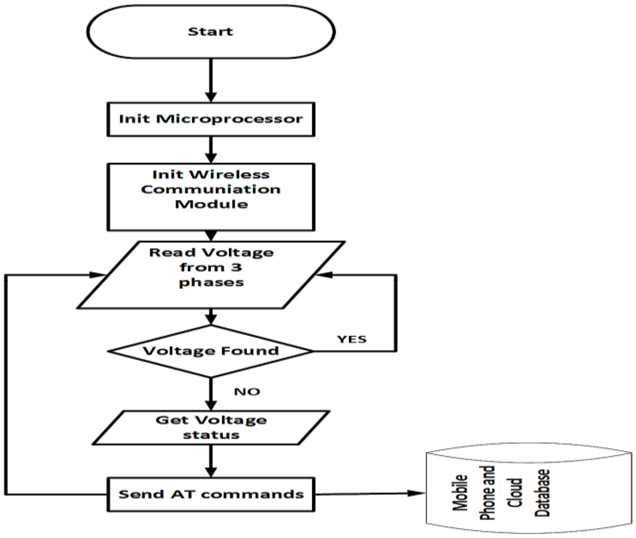

Start
Init Microprocessor
Init Wireless Communiation Module
Read Voltage from 3 phases
Voltage Found
YES
NO
Get Voltage status
Send AT commands
Mobile Phone and Cloud Database

Fig. 5. Flow Chart of Arduino Power Failure Sensor Functionality.

2) Power failure sensor functionality: The power failure sensor functionality is made possible by a voltage sensor circuit in which voltage is stepped down from 230V AC to 12 V AC then rectified and approximately 2V DC used as input to the Arduino. When there is loss of supply, the 2V DC input to the Arduino becomes zero and the embedded program in the Arduino microcontroller and Atmega microprocessor sends an SMS alert consisting of information on the status of the supply from the three phases as well as the power failure location to a predefined cell phone number using the connected SIM808 GSM/GPRS/GPS module over the GPRS cellular network. At the same time through the AT command HTTP POST request the application on the remote server is invoked to parse the power failure alert data and location of the power failure data and insert it into the cloud database for storage and display on the cloud platform.

3) Voltage sensor circuit: The voltage sensor circuit consists of three 230/12V Voltage transformers stepping down the 230V AC input voltage from the Red, Yellow and Blue phases to 12V AC output voltage. A diode bridge rectifier rectifies the 12V AC to DC and a potential divider ensures input of approximately 2V DC to the Arduino board analogue input pins. The analogue input pins take the voltage readings from the sensor and convert them into a number between 0 and 1023 which is later converted back to voltage values for display on LCD and cloud platform. Loss of power in any of the three phases results in the loss of the voltage input to the Arduino, the embedded program in the Arduino detects the loss of supply and sends alert messages to the cloud and predefined mobile phone number. The Arduino through a calculation in the embedded program is able to send the voltage values of the three phases to the cloud platform. The date and time of the alert messages are also displayed on the cloud platform. Fig. 7 shows the main circuit board. Apart from the voltage sensor circuit there is a regulator LM7805 for maintaining the 5V DC supply voltage to the main board circuit. In addition there is a battery controller circuit which controls the charging of the battery and an LCD display for display of voltage values on the electronic device.

4) Battery charge control circuit: The battery charge control circuit controls the charging of the battery. The Arduino microcontroller is programmed to read the battery voltage and depending on the value of the voltage it will send a signal to start or stop the charging of the battery through a Transistor-MOSFET switch combination circuit. To start the charging of the battery the Arduino microcontroller turns on the Negative-Positive-Negative (NPN) transistor which in turn turns on the MOSFET to connect the battery to the 13.5V for charging. Once the battery is charged the Arduino microcontroller turns off the NPN Transistor which in turn turns off the MOSFET to disconnect the battery from the 13.5V charging supply. The battery is needed to power the device when there is loss of supply from all three phases. The device also consists of the LM 2576 regulator to provide the 13.5V charging voltage as well as provide power supply to the SIM808 module. Fig. 6 shows the flow chart for the battery charging functionality and Fig. 8 shows the battery charge control circuit.

5) Arduino microcontroller board: The Arduino is an open source microcontroller interface based board consisting of a circuit board and software called Arduino IDE (Integrated Development Environment), which is used to write and upload the computer code to the physical board [25].

Arduino boards consist of digital and analogue pins which can be programmed to either input or output signals through the IDE [26].

The Arduino microcontroller board is programmed in simplified C++. It is used with the SIM808 module to transfer information to the mobile phone and cloud platform. SIM808 module is a complete Quad-band GSM/GPRS module which combines GPS technology for satellite navigation. It has high GPS receive sensitivity with 22 tracking and 66 acquisition receiver channels [27].

In this study the Arduino was programmed to use the analogue input pins to sense the power failure through the input of the 2 V DC. An Atmega 328 microprocessor is introduced to communicate with the SIM808 module through the MAX232 receiver/transmitter chip and RS232 serial interface. The MAX 232 converts the signals from the microprocessor to RS232 signals to interface with the SIM808 module. Through the use of AT- HTTP commands to the SIM808 module, power failure alert messages were sent to the mobile phone and cloud platform.

6) Cloud services: Cloud services used are offered by the IOTGECKO cloud services platform. IOTGecko cloud platform offers API support over Arduino, Raspberry Pi Microcontrollers and other controller boards. It provides a GUI builder and customized application creator system enabling developers to design desired IOT systems. On the client side the software used was HTML, CSS and JavaScript. HTML makes up the content of the website and enables the browser (like Internet Explorer or Google Chrome) to show the website content. CSS was used to describe the presentation (the look and formatting) of the website. JavaScript is a programming language used to create interactive effects within the web browser. These softwares are referred to as client side because they are executed by the browser on the personal computer to enable viewing of the website. The client side software enables the power failure alerts to be viewed in a certain presentation on a website. On the server side, ASP.net programming language was used to program custom functionality on the website such as enabling the updates of power failure alerts from the microcontroller based device. The server consists of a database engine based on Microsoft SQL for storage of data such as the status and duration of the power failure from each of the phases. The web server software used was Internet Information Services (IIS) with Microsoft windows as the operating system.

From the cloud platform it is possible to see which phases have no supply, what time the power failure occurred, what time power was restored and voltage values for each phase as well as the location of the power failure.

7) Prototype testing: To measure success and performance, the prototype was setup and tested in a lab environment by connecting it to a three phase power supply and a battery for backup supply as shown in Fig. 10. Connection to the cloud platform was established through a web browser interface on a laptop and a mobile phone was used to receive messages from the prototype.

When power supply was switched on, the voltage values of the three phases was successfully displayed on the LCD, cloud platform and an alert message was sent to the mobile phone. The time it takes the three phase voltage values to be displayed both on the local LCD display and on the cloud platform was measured and determined to be within 35 seconds. An SMS indicating the status of the three phase supply was also sent to a predefined mobile phone number within the same time. After this, the red phase supply was disconnected from the device. The red phase voltage value and status was successfully transmitted and displayed onto the local LCD, cloud platform and mobile phone and this transmission was measured to be within 35 seconds as well. This was repeated for the yellow and blue phases each time obtaining similar results. Once all the phases were disconnected a zero voltage value for all the phases was successfully reported to the local LCD, cloud platform and phase status to the mobile phone. With the three phases disconnected the device was able to run on the battery. The three phase supply was then connected back and the status updates were again successfully sent to the local LCD display, cloud platform and predefined mobile number.

Apart from the display of voltage values, location of the power failure was successfully displayed on the cloud platform and mobile phone. The location was also determined to be correct once it was compared with the actual location of the lab environment. The date and duration of the power failure was also indicated on the cloud platform. The start time, end time and date of the power failure displayed on the cloud platform was compared with a GPS synchronised time and date and found to be correct. To determine whether the voltage values displayed on the LCD and cloud platform were correct, they were compared with values of the voltages from the three phases measured using a digital multi-meter.

At the time of testing the prototype a call was made to the ZESCO call centre and it was noted that it took about four minute for the call to be answered. This tallies with the statistics from call centre which indicate that average waiting and processing time for a customer calling call centre is four minutes. When this is compared to the average of 35 seconds processing time in the automatic power failure reporting system, the automatic power failure system performs better.

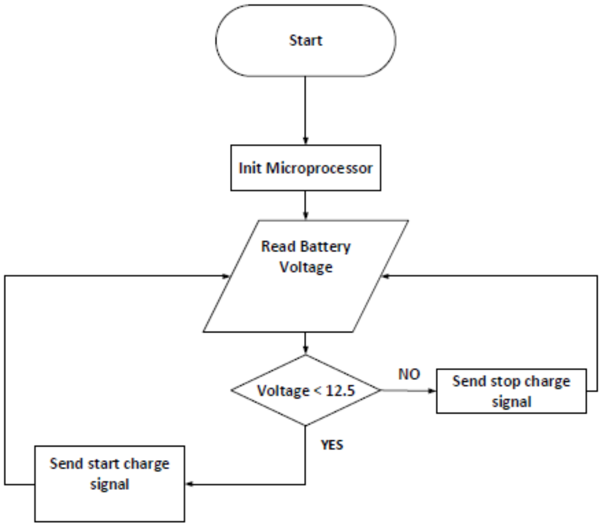


Fig. 6. Flow Chart of Battery Control Functionality.

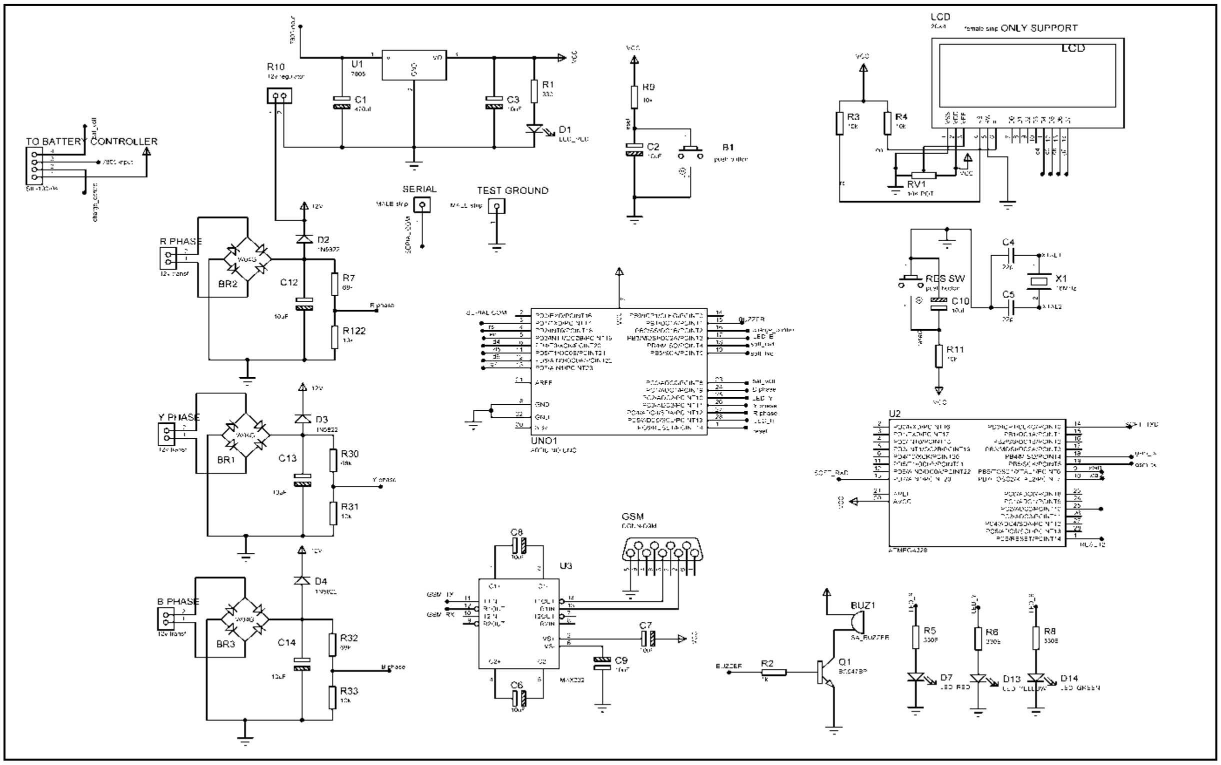


Fig. 7. Prototype Main Board Circuit.

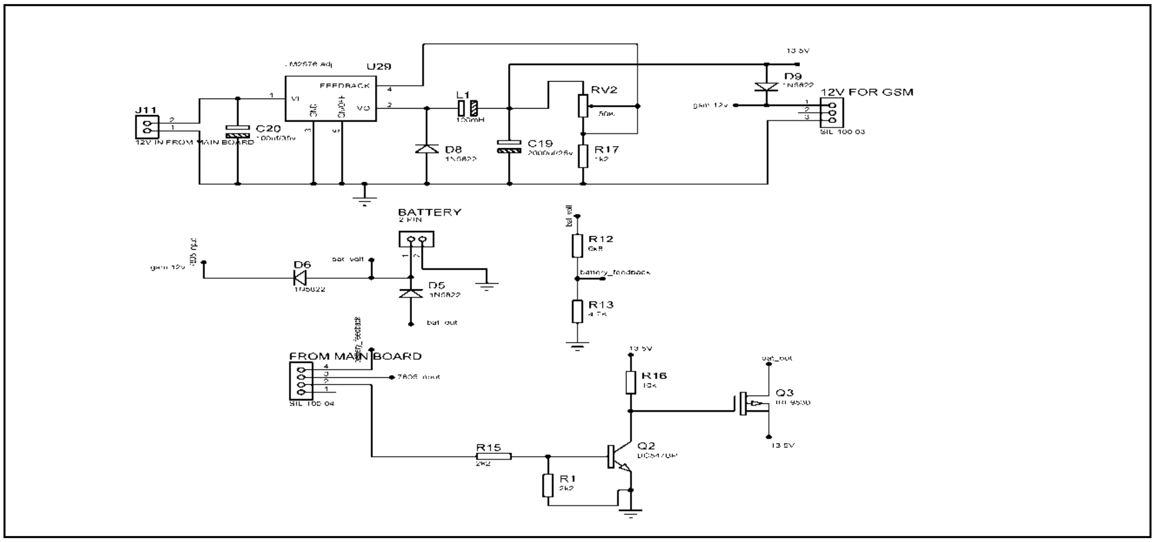


Fig. 8. Battery Charge Control Circuit.

V. Results

The results obtained from the baseline study and system prototype development and testing are presented in this section. The main purpose of conducting the baseline study was to ascertain the challenges that the electricity service provider ZESCO and the customer in the current power failure reporting management model face. The proposed prototype was developed to show as proof of concept of how the fully implemented system would work to alleviate the challenges currently faced by ZESCO and the electricity clients.

A. Baseline Study

The data collected from the baseline study was analyzed using descriptive statistics and the results were presented in form of charts. Data collected from the 383 electricity clients indicated that 100% of the respondents have access to electricity supply with 86% entirely relying on electricity from the power utility company ZESCO and 14% have alternative power sources in the case of power failure.

The results further indicated that 94% of the respondents had experienced power failure.

The majority of the respondents indicated that the longest power failure experienced lasted a number of days resulting in the loss of perishable goods, business customers and damage to equipment.

A cross tabulation of those who reported a power failure complaint to the service provider and those who called ZESCO Call Centre indicate that of the 383 respondents 73% have personally made a power failure complaint to ZESCO with 67.2% of these having made the complaint through the ZESCO call centre.

53% of the 383 respondents have reported power failure by calling the call centre and a cross tabulation of respondents who called the call centre and their experience showed that 86% of those who called call centre have experienced difficulties with accessing the call centre due to either line being constantly engaged or never answered.

When it came to service provider personnel asking for directions to the site of the power failure when a power failure is reported, 73% of the respondents agreed that the service provider personnel ask for directions when a power failure is reported. Further the study revealed that 90.4% of respondents complained of poor response time of ZESCO to power failures reported. Fig. 9 shows the baseline study statistics.

B. Dependency Tests

1) Complaint made to ZESCO by calling call centre:* In determining the extent to which the service provider (ZESCO) Call centre is being utilized by the electricity clients, a cross tabulation was carried out between those that had personally made a power failure complaint to ZESCO, and those that had made such a complaint by calling the Call Centre. Findings revealed that of the total 383 respondents who participated in the study, 202 respondents, representing 53% of the respondents indicated that they had made a power failure complaint to ZESCO by calling call centre, while 47% used other means to report power failure. Out of those that had personally reported power failure to ZESCO, 67.2% had done so by calling the call Centre, while the rest indicated that they had not used the Call Centre.

The null hypothesis for the above test was that the number of reports made to the service provider (ZESCO) through the Call Centre is not significantly high, compared to other methods of reporting. The results revealed a large Chi-square value of 78.226, with a P-value of 0.000 or $P<0.05$. Based on this result, the Null hypothesis was rejected, implying that the number of reports made through the call Centre is significantly high, at 5% significant level.

2) Call centre rating: Further, in determining the ease with which clients can get in touch with ZESCO through the call centre, a cross tabulation was carried out between those that had personally made a power failure complaint to ZESCO by calling call centre, and how they rated the ease with which they managed to get in touch with ZESCO call centre.

Findings of the results revealed that out of those that had reported to ZESCO through the Call Centre, only 16.3% of respondents indicated that they found it either easy, or very easy to get in touch with ZESCO Call Centre, while 48.6% indicated it was either difficult or very difficult. About 34.7% rated ZESCO Call Centre as average in terms of the extent to which it is easy to get in touch with.

In carrying out the dependence test for the above variables, the null hypothesis stated that it was not difficult to get in touch with ZESCO through the Call Centre. The results revealed a large Chi-square value of 31.694, with a P-value of 0.000 or P<0.05. Based on this result, the Null hypothesis was rejected, implying that getting in touch with ZESCO through the Call Centre was not easy at all, and this result was significant at 5% significant level.

The customer call centre further indicated that customer's calls were being rejected due to system resource limitations with the highest number of calls received being related to power failure reports. Statistics collected from the call centre indicate that in a typical month, 70% of customer calls are rejected due to limited system resources at the call centre.

C. System Prototype Results

As already outlined in the previous section the prototype developed consists of an electronic device based on the Arduino microprocessor and web application built on the cloud platform. The electronic device consists of the voltage sensor circuit, Arduino Microcontroller board, Atmega 328 microprocessor, MAX232 chip, SIM808 module and battery voltage control circuit. The hardware setup is as shown in Fig. 10.

The electronic device through its voltage sensor circuit is able to detect when there is a power failure and send the power failure alert message to the mobile phone of the field personnel and cloud application, making it possible to be viewed by the ZESCO customer service personnel. The power failure alert message showing the power supply status of each of the three phases and a link to Google maps to show the location of the power failure is sent to a predefined mobile phone number as shown in Fig. 11. A power failure status alert message for a phase is indicated as Phase Fail and a normal power supply status alert message for a phase is indicated as Phase OK.

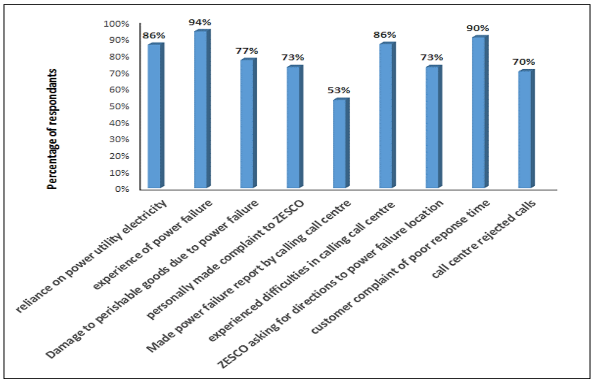


Fig. 9. Baseline Study Statistics.

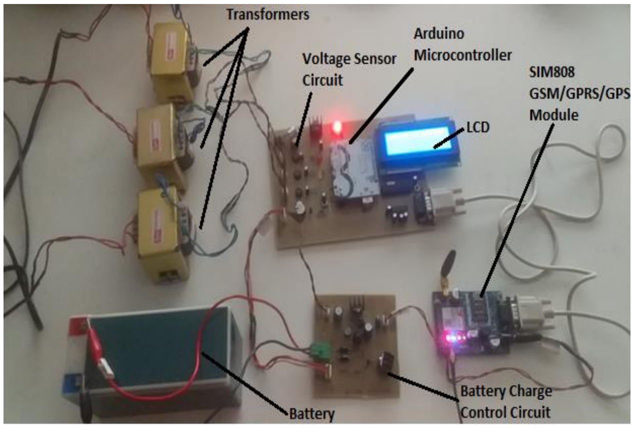
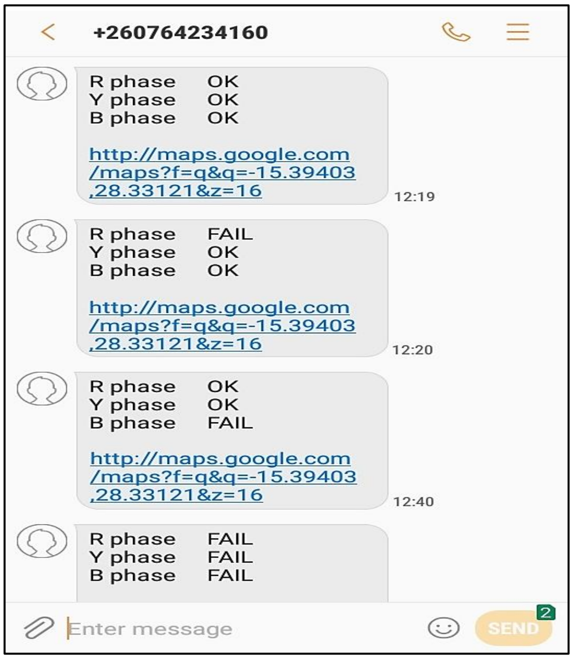


Fig. 10. Hardware Set up of the Microcontroller based System.


+260764234160

R phase OK
Y phase OK
B phase OK

http://maps.google.com/maps?f=q&q=-15.39403,28.33121&z=16

12:19

R phase FAIL
Y phase OK
B phase OK

http://maps.google.com/maps?f=q&q=-15.39403,28.33121&z=16

12:20

R phase OK
Y phase OK
B phase FAIL

http://maps.google.com/maps?f=q&q=-15.39403,28.33121&z=16

12:40

R phase FAIL
Y phase FAIL
B phase FAIL

Fig. 11. Power Failure Alert Information Sent to Mobile Phone.

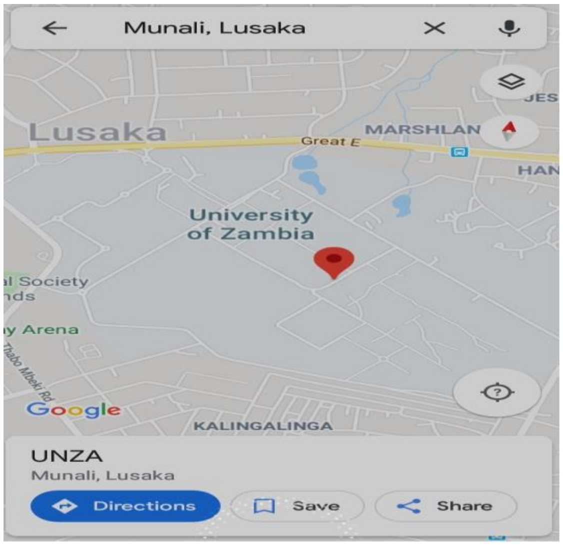
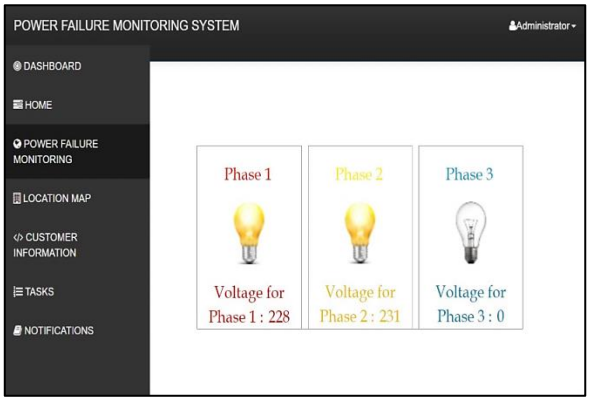


Fig. 12. Power Failure Location Information Sent to Mobile Phone.

When the link to the Google maps sent to the mobile phone is clicked the location of the power failure is shown. See in Fig. 12.

The power alert message indicating the phase which has lost power and the voltage values of the phases with power is also sent to the cloud platform as shown in Fig. 13.

The power failure alert to the cloud platform shows the location of the power failure through the Google API embedded in the cloud application as shown Fig. 14.



Phase 1
Voltage for Phase 1 : 228

Phase 2
Voltage for Phase 2 : 231

Phase 3
Voltage for Phase 3 : 0

Fig. 13. Three Phase Voltage Values on the Cloud Platform.

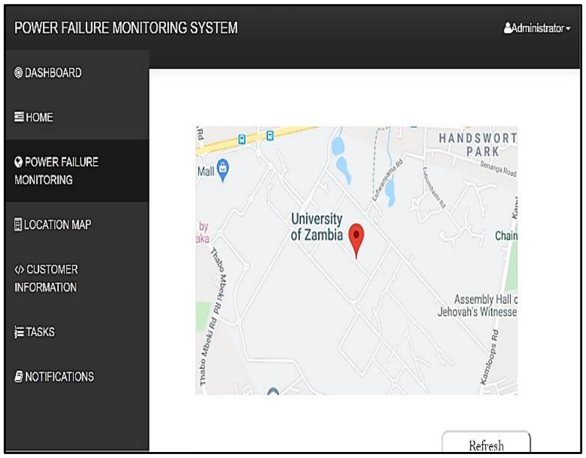
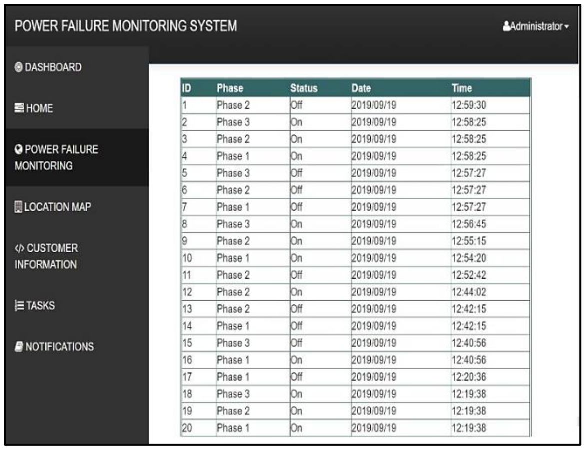


Fig. 14. Location of Power Failure on Cloud Platform.



ID	Phase	Status	Date	Time
1	Phase 2	Off	2019/09/19	12:59:30
2	Phase 3	On	2019/09/19	12:58:25
3	Phase 2	On	2019/09/19	12:58:25
4	Phase 1	On	2019/09/19	12:58:25
5	Phase 3	Off	2019/09/19	12:57:27
6	Phase 2	Off	2019/09/19	12:57:27
7	Phase 1	Off	2019/09/19	12:57:27
8	Phase 3	On	2019/09/19	12:56:45
9	Phase 2	On	2019/09/19	12:55:15
10	Phase 1	On	2019/09/19	12:54:20
11	Phase 2	Off	2019/09/19	12:52:42
12	Phase 2	On	2019/09/19	12:44:02
13	Phase 2	Off	2019/09/19	12:42:15
14	Phase 1	Off	2019/09/19	12:42:15
15	Phase 3	Off	2019/09/19	12:40:56
16	Phase 1	On	2019/09/19	12:40:56
17	Phase 1	Off	2019/09/19	12:20:36
18	Phase 3	On	2019/09/19	12:19:38
19	Phase 2	On	2019/09/19	12:19:38
20	Phase 1	On	2019/09/19	12:19:38

Fig. 15. Power Failure Alert Date and Time on Cloud Platform.

The power failure alert message sent to the cloud platform is able to display the status of each phase, the date and the time enabling the duration of the power failure to be determined as shown in Fig. 15.

VI. Discussion

The study conducted research on the power failure reporting management model for the Zambian electricity Provider ZESCO. A base line study was conducted which revealed the challenges being faced by both the service provider and electricity clients in the current power failure reporting management model. Amongst these challenges were difficulties in accessing call centre to report a fault, challenges in locating the power failure point, limited resources at the call centre leading to rejection of client calls and failure to respond quickly to the power failures reported. Cross table tabulations and chi-square tests revealed a strong relationship between those who made a power failure complaint to the service provider ZESCO and those who reported by calling the call centre implying that most of the complaints are reported through the call centre. A strong relationship was also established between those who called the call centre and those who experienced difficulties in calling the call centre, indicating that clients do face difficulties when calling call centre. The majority of those who called call centre have experienced difficulties in accessing the call centre due to either the line being constantly engaged or the line never answered with most of these calls being rejected as revealed from the Call Centre statistics.

From the findings of the base line study a power failure reporting management model based on sensing and cloud technology was proposed to address the challenges of reporting power failure. In this model power failure is sensed and automatically reported to the service provider without the customer having to report the power failure. Automatic indication of the location of the power failure is also included in the proposed model to enable fast response as there would be no need to call the customer to determine the location. Tests conducted on the prototype confirm that it performs better than the current power failure reporting system which relies on the customer for fault reporting. The proposed model would contribute to reduction of system outage durations and reduce the burden on call centre resources as power failure would be reported automatically.

The proposed model provides other benefits such as the ability to view the information about power failure anywhere and anytime through the cloud platform. Through the cloud platform managers can access the information about power failure and generate statistics anywhere they are in the world.

This system can be adopted in the Low Voltage network where it will be able to capture power failure and enable monitoring of the low voltage network as voltage readings of each phase are displayed on the cloud platform. From the cloud platform duration of the power failure is captured and this can be used to determine the performance of the service provider in responding to power failures.

VII. CONCLUSION

In the proposed model and prototype developed, power failure alerts are sent to the mobile phone and cloud platform in a matter of seconds which is definitely faster than having the customer report the fault. When adopted this system will.

- Improve response times to power failure reports and reduce system outage duration as the power failure will be reported automatically to the utility personnel through the cloud and mobile phone.
- Enable the service provider monitor the low voltage network and keep track of power failure durations through the cloud,
- Reduce the burden on call centre resources and remove the burden of the customer reporting power failures.
- Enable field personnel access to power failure information through the cloud platform.

VIII. RECOMMENDATIONS AND FUTURE WORKS

A. Recommendations

The study has revealed that automating the power failure reporting system is desirable and therefore this system should be fully implemented in order to realize its full benefits.

B. Future Works

The proposed future works, which should be done on this system, are to integrate SMS customer notification from the cloud platform, incorporate the detection of low voltage as it is one of the other main reported fault and investigate how the transmission time of the power failure alerts can be reduced.

REFERENCES

[1] Sachan, "Microcontroller Based Substation Monitoring and Control System with GSM Modem," Internationsl Journal of Electrical and Electronics Engineering, vol. 1, no. 6, pp. 13-21, 2012.

[2] E. Christie, A. Ademola and A. Olayinka, "The Daunting Challenges of the Nigerian Electricity Supply industry," International Journal of Energy Technologies and Policy, vol. 5, no. 9, pp. 25-32, 2015.

[3] H. Tao, L. Simona, A. P. Anca and E. Abouzar, "Analysis of Chain of Events in Major Historic Power Outages," Advances in Electrical and Computer Engineering, pp. 63-70, August 2014.

[4] M. Mohammed, A. Sherif and J. Kluss, "Review of Fault Types, Impacts, and Management Solutions in Smart Grid Systems," International Journal of Smart Grid and Renewable Energy, vol. 10, no. 4, pp. 98-117, 2019.

[5] H. M, R. S, W. S, J. M, A. Ghani and B. Z, "Development of a novel fault management in distribution system using distribution automation system in conjunction with GSM communication," International Journal of Smart Grid and Clean Energy, vol. 2, no. 3, pp. 329-335, 2013.

[6] K. S. Osmo, S. Amir, L. Matti and F.-F. Mahmud, "Optimal Distribution Network Automation Considering Earth Fault Events," IEEE Transactions on Smart Grid, vol. 6, no. 2, pp. 1010-1018, March 2015.

[7] ZESCO, "Our business - Transmission," ZESCO, Lusaka, 2015.

[8] Ministry of energy and Water Development, "Power System Development Master Plan for Zambia," Ministry of energy and Water Development, Lusaka, 2010.

[9] H. M. Zulu, "ZESCO Customer Base Hits one million," ZESCO, Lusaka, 2019.

[10] B. A. Thomas and D. Carl-Johan, "Power outages and economic growth in Africa," International journal of enrgy economics, vol. 38, pp. 19-23, 2013.

[11] T. Gowdhaman and D. Surendran, "Automatic Street Light Control and Fault Deection System with Cloud Storage," International Journal of Scientific & Engineering Research, vol. 8, no. 5, pp. 1-5, 2017.

[12] S. Chihana, J. Phiri and D. Kunda, "An IoT based Warehouse Intrusion Detection (E-Perimeter) and Grain Tracking Model for Food Reserve Agency," International Journal of Advanced Computer Science and Applications, vol. 9, no. 9, pp. 213-223, 2018.

[13] T. Muyumba and J. Phiri, "A Web based Inventory Control System using Cloud Architecture and Barcode Technology for Zambia Air Force," International Journal of Advanced Computer Science and Applications, vol. 8, no. 11, pp. 132-142, 2017.

[14] C. Mulima and P. Jackson, "A Remote Sensor Network using Android Things and Cloud Computing for the Food Reserve Agency in Zambia," International Journal of Advanced Computer Science and Applications, vol. 8, no. 11, pp. 411-418, 2017.

[15] K. Okokpujie, E. Amuta, R. Okonigene and J. Samuel, "Momitoring and fault detection system for power transmission using GSM technology," in International Conference of Wireless Networks, 2017.

[16] S. Sujatha M and K. M. Vijay, "On-line monitoring and analysis of faults in transmission and distribution lines using GSM technique," Journal of Theoretical and Applied Information Technology, vol. 33, no. 2, pp. 258-265, 2011.

[17] R. P. Vikramsingh, J. Shivani, D. Anand, S. Arti and C. Kapil, "Automatic Fault Detection in Transmission Lines using GSM Technology," International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, vol. 6, no. 4, pp. 90-95, 2018.

[18] D. Vihar, D. Avee, S. Rishi and J. Arjun, "Implementation of Remote Monitoring of Substation EquipmantsUsing GSM," International Journal of Advanced Research in Electrical,Electronics and Instrumentation Engineering, vol. 5, no. 6, pp. 5565-5571, 2016.

[19] D. Krupal, P. Jenish, S. Yasin, M. Anas and P. Krishn, "Substation Monitoring and Control Using Microcontroller & GSM," International Research Journal of Engineering and Technology, vol. 4, no. 4, pp. 398-403, 2017.

[20] A. Hassan, "Monitoring and Controlling of Distribution Transformer Using GSM Module (AVR Microcontroller Based)," International Journal of Advanced Research in Science and Engineering, vol. 6, no. 8, pp. 1645-1653, 2017.

[21] P. Sandeep, B. Ashish, B. Nitishsingh, K. Neelam and U. Prag, "Web based substation monitoring and protection and control," International Journal of Application or Innovation in Engineering & Management, no. Special Issue for National Conference On Recent Advances in Technology and Management, pp. 1-7, 2013.

[22] A. Gerald, A. R. Catherine and M. S. Scott, "Topic sensitivity and research design: effects on internet survey respondents' motives," Asia Pacific Journal of Marketing and Logistics, vol. 26, no. 1, pp. 147 - 161, 2014.

[23] C. Joseph and K. S. Russell, Research Methods in Education, Thousand Oaks, CA: SAGE Publications, 2011.

[24] D. Alan, H. W. Barbara and R. Roberta, System Analysis and Design, USA: John Wiley & Sons, Inc, 2012.

[25] S. Monk, 30 Arduino Projects for the evil genus, McGraw-Hill Companies, Inc publisher, 2010.

[26] B. Massimo, Getting Started With Arduino, California: O'Reilly Media Inc , 2011.

[27] S. ShengWu and W. Xuegang, "SIM808 Hardware design V1.00," Shanghai SIMcom Wireless Solutions Limited, Shanghai, 2014.

JobChain: An Integrated Blockchain Model for Managing Job Recruitment for Ministries in Sultanate of Oman

Vinu Sherimon[1], Sherimon P.C[2], Alaa Ismaeel[3]
Department of Information Technology, Higher College of Technology, Muscat, Sultanate of Oman[1]
Faculty of Computer Studies, Arab Open University, Muscat, Sultanate of Oman[2, 3]
Faculty of Science, Minia University, Minya, Egypt[3]

***Abstract*—Industries around the world has revolutionized with the arrival of blockchain technology. Blockchain applications and use cases are in the process of development in different domains. This research presents a blockchain platform "JobChain" to manage the job recruitments. The case study is conducted for the job recruitments in various Ministries in Sultanate of Oman. Currently, in Oman, citizens are aware of the job vacancies through the advertisements posted in newspapers or social media. A job seeker then applies for the desired job and thereafter the qualified candidates are called for tests/ interviews. To ease this process, a solution based on blockchain which includes various Ministries and the citizens/ residents of Sultanate of Oman is proposed in this research. Ministries can post the job vacancies in the blockchain and qualified citizen(s) can submit their application. Relevant cryptographic functions are used to verify the authenticity of the participants in the blockchain network. The citizens feel the existence of a trusted secure government, which is mandatory for the development of a country. Unlike traditional models, blockchain eliminates the need of intermediary agents (e.g. Job Consultancies) thereby providing direct communication between the participants of the blockchain. The proposed blockchain framework helps the citizens in Oman to get updated about the job vacancies. Hyperledger Composer Playground is used to design and test the proposed blockchain business network. Preliminary results show that the participants and assets are created successfully and the transactions to approve a job vacancy and a job application is done through the proposed blockchain network.**

Keywords—Blockchain; permissioned; chaincode; hyperledger composer playground; job recruitment

I. Introduction

As part of the Industrial Revolution 4.0, blockchain technologies are accelerating at a great velocity. It was first developed for bitcoin cryptocurrency [1]. Blockchain technology runs on distributed networks. The goal of this technology was to provide a global environment which is decentralized, transparent and secure to conduct transactions [1]. They have the potential to transform and reshape data governance. In centralized database systems, the system management, access control, protection against cyber-attacks, etc. are done by a single entity. But in blockchain, every node in the network updates the data independently. The transactions are kept with every participant in an openly distributed ledger securely and it is immutable. Data once added to the blockchain cannot be repudiated or updated. Also, the users can cross-check the validity of the data provided in the blockchain. The need for a third-party or a single entity control is eliminated here [1]. This technology offers digital trust between the participants involved in the network. Cryptographic algorithms will ensure the authenticity of the participants and the transactions will be carried out only by the authorized participants [2].

A blockchain system is classified into two different types of networks: Permissionless blockchains and permissioned blockchains. Ethereum and Bitcoin are examples of Permissionless blockchain. In this network, there is no central authority and any node can join the network to submit a transaction. These transactions are then validated by miner nodes, who receives incentives on calculating a complex hash value that results in a new block creation. The incentives (usually a bitcoin) encourage the miner nodes to stay in the network and they compete among themselves in mining a new block. On the other hand, in permissioned blockchain, a central authority decides about the members of a network. Only authenticated and authorized participants can join the network.

In Sultanate of Oman, Ministry of Civil Services (MOCS) is responsible for recruiting the citizens to the various ministries in Oman. The traditional method was to collect the requirements from different Ministries through Human Resources Management System (HRMS) and place notifications in the newspapers. Job seekers then apply for the desired job and thereafter the qualified candidates are called for tests/ interviews. Since it was very time-consuming, later SMS (Short Messaging Service) was introduced so that the job seeker can send an SMS with the desired job code and the employee code registered in Ministry of Manpower (MOM). MOM validates the applicant's data and other relevant documents. Then the suitable candidates go through aptitude tests and interviews.

In the existing system, the following are the drawbacks:

- The Job vacancy notifications are published in newspapers, which costs money.
- The Job seeker should remember the job codes to send SMS to MOCS.

- If a Job seeker wishes to update his CV, for example, to add a new qualification, he/she must submit the new certificates to MOM, and wait for validation, which is time consuming and difficult.

To enhance the existing recruitment process, we propose a blockchain based platform “JobChain” to ease the recruitment process.

Blockchains are digital decentralized ledgers. Blockchain applications builds trust, accountability and transparency among the participants in the network. A permissioned private blockchain [2] which requires the verification of the participants is proposed in this project. Every participant who is involved in the blockchain requires permission to execute transactions on the blockchain. Membership services authenticates and authorizes the identity of the participants. The blockchain creates a ledger which stores the whole history of blockchain transactions which are immutable, and it is distributed throughout the blockchain network. Each node in the blockchain maintains its own copy of the ledger. Whenever an update or delete transaction is initiated it is done based on the consensus between the nodes. The blockchain creates more transparency between the job seeker and the ministries. Here every job vacancy is posted in the blockchain by the concerned Ministry. MOCS approves/ rejects the posted job vacancy.

The rest of the paper is organized as follows: Section 2 presents the related work. Section 3 describes Permissioned blockchain networks and the different components. The architecture of the proposed model is shown in Section 4. Section 5 presents the implementation details of the model. Section 6 presents the results and Section 7 describes the significance of the proposed model followed by conclusion and future in Section 8.

II. Analysis of Related Work

Blockchain technology is disrupting different industries in many countries now. Estonia is the first country to test the blockchain technology in 2008, even before the white paper on bitcoin was published [3]. At that time, since the term “blockchain” was not in use, the technology was named as “hash-linked time-stamped”. The blockchain enabled eEstonia project, integrates all government services to a single platform. The country had developed a blockchain technology KSI, which secures e-services such as e-Healthcare records, e-Judiciary, e-Government, e-Land registry, e-Banking, etc. by storing the data in a distributed ledger [3]. This avoids data misuse and corruptions. A web-based blockchain voting program (e-Voting) was used in 2017 in Estonia [3]. Using this system, a citizen can cast his/her vote and he/she can also verify the data (voting choice) received by the government [3].

The first government endorsed blockchain platform is launched in Dubai on October 30, 2018. It is as part of the vision to make the Emirati government paperless by 2021 [4]. It integrates the digitized services to the lives of the normal citizens. Various blockchain applications in different sectors such as education, healthcare, energy, etc. are also launched in Dubai, the pioneer in blockchain technology.

In an urge to ensure the food safety and authenticity, in 2017, Dubai has launched ‘Food Watch’, a digital platform that digitize the food safety and nutritional information of all foods served in the country [5]. Initially, all establishments which handle high-risk foods are required to update the platform with the information related to the foods they handle. This enables the consumers and the food quality checkers to verify the authenticity of food. The future phase of this project includes to integrate blockchain, big data and IOT technologies to track food products from farm to table [5].

UAE Banks Federation has initiated the process of implementing blockchain to digitize various processes in bank. For example, adopting blockchain technology in KYC (Know Your Customer) processes to enhance the experience of customers [6]. The Swiss city of Zug pilot tested “e-voting pilot”, a blockchain application to cast votes. In 2017, an Ethereum-based application “uPort” was launched in Zug to digitize ID’s of residents. This ID was used by the residents to cast votes using the blockchain application [7].

A private blockchain based project was introduced in the Republic of Georgia to register land titles thus becoming the first government to introduce blockchain in this domain [8]. The application allows citizens to validate their land registrations securely. The project is in the process of expansion by including services such as purchase and sale of land titles, demolition of lands, mortgages, etc. [8].

Kuwait has introduced its first blockchain application for email which ensures secured transactions [9]. As part of Vision 2030, Saudi Arabia also has partnered up with IBM to implement several blockchain applications in various sectors [10]. Bahrain’s economic development board (EDB) and Abu Dhabi global market (ADGM) have agreed to collaborate on blockchain based developments and digital payments [11]. In Sultanate of Oman, Blockchain Solutions & Services Co (BSS) is the pioneer in conducting many forums and workshops about Blockchain [12]. BSS is working with different entities to establish the needed infrastructure to implement this new powerful technology [12].

As per the literature review, no research has been reported regarding the application of blockchain in job recruitments.

III. Permissioned Blockchain

Permissioned blockchains can be utilized by such business organizations where the participants are known to each other [ABB+18]. They are used in cases where there exists partial trust or no trust between the participants. Here, a central authority is authorized to decide on the rules for the participants to join the network and submit transactions. Transactions are validated by every participant in the blockchain by reaching a Consensus. They are executed by a Smart Contract, which updates the data in a distributed ledger and adds a new block to blockchain.

Consensus–Because of the distributed nature of blockchain, every participant should agree on certain rules to validate the block generated and to add it in the blockchain. These rules are called Consensus and they work together with Smart Contracts to guarantee the order, correctness and validity of transactions. There are different types of consensus algorithms based on several classifications. For instance, if the classification is based on the method of block selection, then voting based algorithms and lottery-based algorithms are the two types. In voting type, majority of the participants validate the blocks or transactions and consensus is achieved fast, whereas in lottery-based systems, the winner of the process proposes a new block and adds it to the blockchain.

Apache Kafka is a voting-based consensus algorithm. If the classification is based on the type of peers, there are two types of algorithms-competing peers and non-competing peers. In competing peers' algorithm, the peers compete on the next block to be appended on the blockchain whereas in non-competing peers' algorithm, only few peers work on block creation, and the rest of the peers validates the block. Some other consensus algorithms depend on the architecture of execution of the transactions. In the Order-Execute model, the ordering of the transaction takes place before the execution of transactions. Proof-of-Work (PoW) algorithm is an example of this type [ABB+18]. In Execute-Order-Validate model, the transactions are executed initially, followed by ordering of the executed transactions, and the validation of the outputs. The transaction execution model of Hyperledger Fabric is Execute-Order-Validate style.

Membership Service Provider (MSP) – Every trusted member of a blockchain network must have a digital identity/ certificate to allow access to resources and to submit transactions. Each of these identities must be issued by a trusted authority. An MSP is a component that determines the valid identities of an organization. The default MSP implementation uses X.509 certificates issued by Certificate Authority (CA). In blockchain networks, sometimes there may be more than one CAs. Also, the Root CA issues certificates to various Intermediate CAs, which in turn issues certificates to members of different organizations. In this way, a chain of trust is established. In a blockchain network, MSP identifies the trusted Root CAs and Intermediate CAs (if any) which are used to define the members of a domain. MSP can also determine specific member roles and define the access privileges of members in the network as per their roles.

Chaincodes – The business logic of the blockchain network is defined in Chaincodes. It forms the heart of the blockchain network. In a Permissionless blockchain like Ethereum, chain codes are referred as Smart Contracts. Also, both the terms, Smart Contract and Chaincode are used interchangeably in Permissioned Blockchains. A Chaincode is installed on all nodes of the blockchain network. The client application invokes the Chaincode with the support of APIs and submit requests for transactions. Also, the transactions can be executed by itself upon meeting certain criteria. After the transactions are validated by the network participants, the results are updated in the shared distributed ledger by all the participants of the network.

Authorization Policy – It defines the policies for the set of nodes who will execute the Chaincode and validates its results. The policies take the form of "MSP.Role". If the policy is defined as Org1.admin, then the policy designates the admin of Org1 MSP.

Channels – This is required if a large business network includes a subset of private networks which do not want to share some confidential information with the rest of the network. An organization can participate in multiple blockchain networks via separate channels. All the participants in a channel share information and coordinate with each other according to the authorization policy.

Ledger – There are two different components for a ledger – world state and a blockchain. The world state holds the current value of an object and the blockchain keeps the history of all the transactions. When a new transaction is successfully committed, the world state gets updated with the new results. A key-value pair represents a world state. The blockchain consists of blocks, connected with each other and each block includes a set of transactions.

Data Storage – The world state can be implemented in two ways – using CouchDB or LevelDB. LevelDB supports simple key-value pairs and CouchDB supports complex queries.

IV. Jobchain Architecture

Fig. 1 shows the architecture of the proposed blockchain application.

Ministries can post the job vacancies in the blockchain, which will be approved by the Ministry of Civil Services. As per the requirements, qualified citizen(s) can submit their application. This will be validated by the Ministry of Manpower. If the job application is initially approved, the concerned Ministry will inform the candidates about the exam/ interview details. Also, the results of the exam/ interview will be uploaded in the blockchain and the status of job vacancies are updated accordingly.

Every blockchain network consists of participants, assets, and transactions. The participant Ministry represents different ministries in Sultanate of Oman. MinistryStaff represents the staff under each ministry who posts the categories of jobs and the current vacancies. Here MOCS (Ministry of Civil Services) is responsible to approve the job vacancies posted by other ministries. Job Seeker's role is to view the job vacancies and to apply for the job. MOM (Ministry of Manpower) is responsible to validate the job seeker's data.

The asset Job Categories represent the categories of jobs posted by the ministries. Job Vacancies signify the current vacancies in the ministries and Job Application characterizes the application of job seekers.

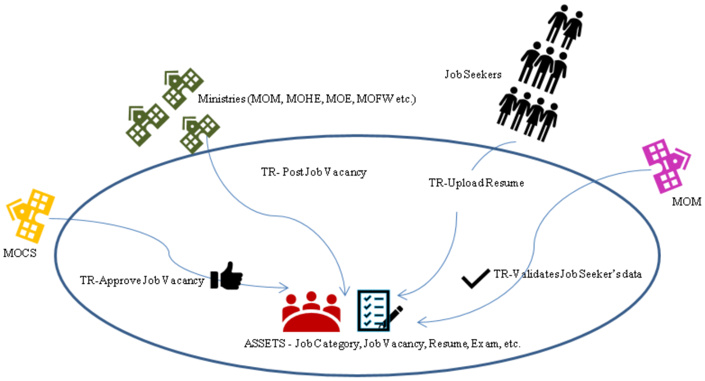

Job Seekers
Ministries (MOM, MOHE, MOE, MOFW etc.)
TR- Post Job Vacancy
TR-Upload Resume
MOM
MOCS
TR-Approve Job Vacancy
TR-Validates Job Seeker's data
ASSETS - Job Category, Job Vacancy, Resume, Exam, etc.

Fig. 1. Proposed Framework.

Post Job Vacancy transaction is used by Ministries to post the vacancies. Approve Job Vacancy is used by MOCS to approve the vacancies posted by other ministries. ValidatesJobSeeker's data is used by MOM to validate the data of a Job Seeker.

V. Model Implementation

An object-oriented modeling language is used to define the domain model for a business network definition. All the resources such as assets, participants, transactions, and events are defined in the model file.

Every Ministry participant is identified by an Id, followed by name and address. Here address is defined as a concept. It consists of multiple information combined in a single concept.

```
participant Ministry identified by
ministryid
{
      o String ministryid
      o String ministryName
      o Address address
}
```

An abstract object called JobChainParticipant is defined and the participants MinistryStaff and JobSeeker are extended from this abstract object as they both share common properties.

```
abstract participant JobChainParticipant
identified by participantId
{
      o String participantId
      o Contact contact
      o Address address
}
```

The MinistryStaff participant has additional attributes such as department, jobTitle, userStatus and a relationship to Ministry object. userStatus is defined as an enumeration to define the status of the user.

```
participant MinistryStaff extends
JobChainParticipant
{
      o String department
      o String jobTitle
      o UserStatus userStatus
default="Active"
      -->Ministry ministry
}
```

The assets of the blockchain are JobCategory, JobVacancy, etc. The asset JobCategory is given below:

```
asset JobCategory identified by
jobCategoryId {
      o String jobCategoryId
      o String jobCategoryName
      o String jobCategoryDes
      o jobCategoryStatus
jobCategoryStatus
      -->MinistryStaff creator
}
```

In addition to the attributes, jobCategoryId, jobCategoryName, jobCategoryDes, and jobCategoryStatus, a relationship exists to MinistryStaff object in the above asset.

Similarly, different transactions such as CreateMinistry, CreateMinistryStaff, PostJobCategory, PostJobVacancy, etc. are also defined in the model.

To execute the different transactions defined in the model file, a logic file should be implemented in the blockchain network. This file contains the codes required to implement each of these transactions. Corresponding to each transaction, a function is written in the logic file, which is used to update the attributes of the participants/ assets. These updates are saved to the participant and asset registries.

VI. Testing and Results

Hyperledger composer playground, an open source web application tool is used to develop the JobChain prototype. Hyperledger is a project of open source Blockchain and distributed ledger technology by the Linux Foundation. This tool is used to build the business model, test the model and deploy the network to runtime. It keeps the blockchain model in the browser storage.

A. Creation of Participants

Initially, the participant registry is empty. The admin submits the transaction CreateMinistry in the blockchain and register two members of the Ministry participant in the blockchain.

Ministry of Health – MOH (id: 100), who can post a Job Category and Job Vacancy.

```
100
{
 "$class":
"org.example.empty.Ministry",
 "ministryid": "100",
 "ministryName": "Ministry of
Health",
 "address": {
  "$class":
"org.example.empty.Address",
  "street": "Darsait",
  "city": "Muscat",
  "country": "Oman"
 }
}
```

Ministry of Civil Service – MOCS (id: 102) who can post a Job Category and Job Vacancy and can approve/ reject the Job Vacancies posted by other ministries.

```
102
{
 "$class":
"org.example.empty.Ministry",
 "ministryid": "102",
 "ministryName": "Ministry of Civil
Service",
 "address": {
  "$class":
"org.example.empty.Address",
  "street": "Ruwi",
  "city": "Muscat",
  "country": "Oman"
 }
}
```

Later, MOH submits the transaction `CreateMinistryStaff` to register an MOH staff in the blockchain. For example, the data of a MOH staff (id: 1) is registered as shown below:

```
1
{
 "$class":
"org.example.empty.MinistryStaff",
 "department": "IT",
 "jobTitle": "Manager",
 "userStatus": "Active",
 "ministry":
"resource:org.example.empty.Ministr
y#100",
 "participantId": "1",
 "contact": {
  "$class":
"org.example.empty.Contact",
  "firstName": "Ahmed",
  "lastName": "Al Busaidi",
  "email": "ahm@gmail.com",
  "phone": 96321456
 },
 "address": {
  "$class":
"org.example.empty.Address",
  "street": "Azaiba",
  "city": "Muscat",
  "country": "Oman"
 }
}
```

A job seeker can register in the blockchain through `CreateJobSeeker` transaction (id: 0738).

```
0738
{
 "$class":
"org.example.empty.JobSeeker",
 "dob": "1974-05-28T00:00:00.000Z",
 "qualification": [
  {
   "$class":
"org.example.empty.Qualification",
   "qname": "Master",
   "qnameInstitution": "MGU",
   "dateObtained":          "1997-01-
01T00:00:00.000Z",
   "grade": "A"
  },
  {
   "$class":
"org.example.empty.Qualification",
   "qname": "PhD",
   "qnameInstitution": "MGU",
   "dateObtained":          "2002-01-
01T00:00:00.000Z",
   "grade": "A"
  }
 ],
 "totalExp": 10,
 "jobStatus": "Unemployed",
 "participantId": "0738",
 "contact": {
  "$class":
"org.example.empty.Contact",
  "firstName": "Ameer",
  "lastName": "Al Kalbani",
  "email": "amr@hotmail.com",
  "phone": 24455789
 },
 "address": {
  "$class":
"org.example.empty.Address",
  "street": "Al Khuwair",
  "city": "Muscat",
  "country": "Oman"
 }
}
```

B. *Creation of Assets*

Initially, the asset registry is empty. Now, assume that MOH would like to register a JobCategory asset in the blockchain. The staff of MOH submits the transaction PostJobCategory and adds an instance of the asset JobCategory (id: C1) in the asset registry.

```
C1
{
 "$class":
"org.example.empty.JobCategory",
 "jobCategoryId": "C1",
 "jobCategoryName": "Academic",
 "jobCategoryDes": "Teaching and
Research",
 "jobCategoryStatus": "Active",
 "creator":
"resource:org.example.empty.Ministr
yStaff#1"
}
```

After the category C1 is registered in the blockchain, an instance of the asset JobVacancy under C1 is also posted by MOH by submitting PostJobVacancy transaction.

```
8421
{
 "$class":
"org.example.empty.JobVacancy",
 "jobCode": "8421",
 "jobTitle": "Assistant Lecturer",
 "jobDes": "Assisting the
lecturer",
 "salary": 925,
 "jobLocation": "Muscat",
 "benefits": "Free medical",
 "essentialQualification":
"Masters",
 "desirableQualification": "PhD",
 "essentialExperience": "4",
 "desirableExperience": "5",
 "status": "Open",
 "jobVacancyApprovalStatus":
"PENDING",
 "postedDate": "2019-12-
29T17:43:31.696Z",
 "lastDate": "2020-12-
01T17:43:31.696Z",
 "jobCategory":
"resource:org.example.empty.JobCate
gory#C1",
 "creator":
"resource:org.example.empty.Ministr
yStaff#1"
}
```

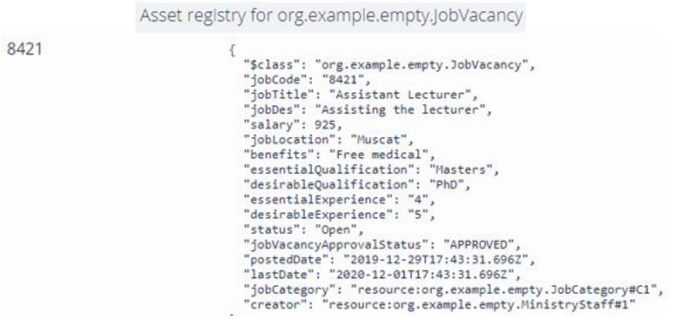
Next, assume that the JobSeeker Ammer Al Kalbani (id: 0738) wish to apply for the job 8421 given above. This is done by submitting the transaction PostJobApplication into the blockchain.

```
7758
{
 "$class":
"org.example.empty.JobApplication",
 "JobApplicationId": "7758",
 "jobVacancy":
"resource:org.example.empty.JobVaca
ncy#8421",
 "jobSeeker":
"resource:org.example.empty.JobSeek
er#0738",
 "applicationDate": "2019-12-
29T17:46:22.690Z",
 "jobApplicationStatus": "APPLIED"
}
```

C. *Approval of Job Vacancy*

Now, MOCS is given the rights to update the JobVacancy asset. This is done through the ApproveJobVacancy transaction. The attribute jobVacancyApprovalStatus is set to the value APPROVED, once it is approved by MOCS.

For instance, MOCS submits the transaction ApproveJobVacancy to approve the JobVacancy (id: 8421). This results in the change in the attribute value of jobVacancyApprovalStatus to APPROVED in the asset JobVacancy (Fig. 2).



Asset registry for org.example.empty.JobVacancy

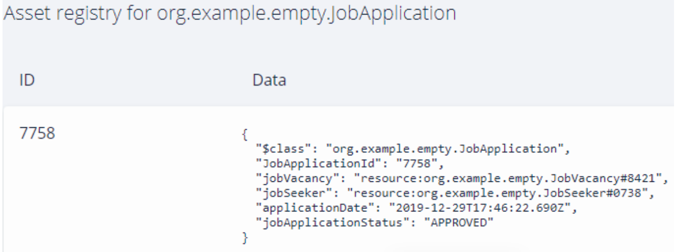
```
8421        {
              "$class": "org.example.empty.JobVacancy",
              "jobCode": "8421",
              "jobTitle": "Assistant Lecturer",
              "jobDes": "Assisting the lecturer",
              "salary": 925,
              "jobLocation": "Muscat",
              "benefits": "Free medical",
              "essentialQualification": "Masters",
              "desirableQualification": "PhD",
              "essentialExperience": "4",
              "desirableExperience": "5",
              "status": "Open",
              "jobVacancyApprovalStatus": "APPROVED",
              "postedDate": "2019-12-29T17:43:31.696Z",
              "lastDate": "2020-12-01T17:43:31.696Z",
              "jobCategory": "resource:org.example.empty.JobCategory#C1",
              "creator": "resource:org.example.empty.MinistryStaff#1"
```

Fig. 2. Job Vacancy Asset.

D. *Approval of Job Application*

When a job seeker registers a job application, the current status of the job application is APPLIED. The concerned Ministry is responsible to check the JobApplication and update its status. If the initial requirements are met, it is changed to APPROVED by the concerned ministry Otherwise the status is changed to REJECTED.

For instance, MOH checks the application submitted against its Job vacancy. Since the initial requirement are met, MOH submits the transaction ApproveJobApplication, which approves the JobApplication (id: 7758) submitted by the JobSeeker (id: 0738). The asset registry given in Fig. 3 shows the change in jobApplicationStatus.



Asset registry for org.example.empty.JobApplication

```
ID          Data

7758        {
              "$class": "org.example.empty.JobApplication",
              "JobApplicationId": "7758",
              "jobVacancy": "resource:org.example.empty.JobVacancy#8421",
              "jobSeeker": "resource:org.example.empty.JobSeeker#0738",
              "applicationDate": "2019-12-29T17:46:22.690Z",
              "jobApplicationStatus": "APPROVED"
            }
```

Fig. 3. Job Application Asset.

VII. SIGNIFICANCE OF JOBCHAIN

A blockchain network is a form of distributed ledger design, which is spread across multiple nodes where each node maintains an identical copy of the ledger. The main feature of this technology is that there is no central authority to administer the data. It is updated independently and ensure data synchronization. The data is secured using cryptographic techniques. The records generated are real-time, so each node keeps an up-to-date copy of the data or transactions [13]. The transaction records created are permanent and immutable [13].

The proposed research will impact the country in many ways. A common platform exists where the Ministries collaborate with each other [14]. Citizens often find difficulty to understand the category of jobs, the nature of jobs, current vacancies, etc. Also, currently the job vacancies are advertised in the newspapers and the job seeker should remember the job code and send SMS to apply for the job. Distributed Ledger Technology eliminates the need for intermediaries in any transactions. Through the proposed blockchain framework, the citizens can directly view the job vacancies in each Ministry. If the job profile matches with his/ her education and experience, the citizen can apply for the job.

VIII. Conclusion and Future WORK

Blockchain applications offer several unique features. A blockchain is not owned by a single entity; it is the asset of all the participants in the blockchain network. Data stored in a blockchain is secured against deletion/alteration. The stakeholders of a blockchain have a common platform to collaborate with each other. As every participant owns a copy of the blockchain data, it is resilient against any attacks. Also, every participant can verify the correctness of the transactions. We have presented a blockchain architecture to manage the job recruitments in various Ministries in Sultanate of Oman. The proposed model will enhance the trust, validity and accountability between Job Seekers and Ministries in the Sultanate. The full implementation of the proposed framework in Hyperledger Composer Playground is planned as the future scope of this paper.

Acknowledgment

The research leading to these results has received funding from the Research Council (TRC) of the Sultanate of Oman under the Block Funding Program BFP/ RGP/ ICT/ 18/ 114.

References

[1] Yli-Huumo, J., Ko, D., Choi, S., Park, S., & Smolander, K. (2016). Where is current research on blockchain technology? —a systematic review. PloS one, 11(10), e0163477.

[2] Greenspan, G. (2015). MultiChain private blockchain—White paper. URl: Retrieved from http://www. multichain. com/download/Multi Chain-White-Paper. pdf.

[3] Drechsler, W. (2018). Pathfinder: e-Estonia as the β-version. JeDEM-eJournal of eDemocracy and Open Government, 10(2), 1-22.

[4] Emirates Blockchain Strategy 2021. Retrieved from https://government.ae/en/about-the-uae/strategies-initiatives-and-awards/federal-governments-strategies-and-plans/emirates-blockchain-strategy-2021.

[5] Darin Detwiler. (2018, Feb 27). Retrieved from https://www.ibm.com/blogs/blockchain/2018/02/one-nations-move-to-increase-food-safety-with-blockchain/.

[6] UBF’s CEOs Advisory Council explored adoption of Blockchain in banks. Retrieved from https://www.unlock-bc.com/news/2018-12-17/ubfs-ceos-advisory-council-explored-adoption-of-blockchain-in-banks.

[7] Wofie Zhao. (2018, June 11). Swiss City Plans Blockchain Voting Pilot Using Ethereum-Based IDs. Retrieved from https://www.coindesk.com/swiss-city-plans-to-vote-on-blockchain-using-ethereum-digital-id.

[8] Laura Shin. (2017, Feb 7). Retrieved from https://www.forbes.com/sites/laurashin/2017/02/07/the-first-government-to-secure-land-titles-on-the-bitcoin-blockchain-expands-project/#3a7e53d04dcd.

[9] Prepare for Blockchain Disruption. Retrieved from https://www.protiviti.com/KW-en/node/94216.

[10] Saudi Arabia Partners With IBM To Use Blockchain for Better Government Services. Retrieved from https://ethereumworldnews.com/saudi-arabia-ibm-blockchain-services/.

[11] Arabian Business Industries. Retrieved from https://www.arabianbusiness.com/technology/403858-bahrain-minister-lauds-blockchain-as-true-mark-of-progress.

[12] MuscatDaily.Com. Retrieved from https://www.muscatdaily.com/Archive/Business/Blockchain-infrastructure-to-be-rolled-out-soon-in-Oman-58s7.

[13] Michael, JW, Alan Cohn, and Jared R. Butcher. "Blockchain technology." The Journal (2018).

[14] Peck, M. "Reinforcing the Links of the Blockchain." IEEE Future Directions Initiative White Paper (2017).

Automated Machine Learning Tool: The First Stop for Data Science and Statistical Model Building

DeepaRani Gopagoni[1], P V Lakshmi[2]
Department of Computer Science and Engineering
GIT GITAM (Deemed to be University)
Vishakhapatnam, Andhra Pradesh, India

Abstract—**Machine learning techniques are designed to derive knowledge out of existing data. Increased computational power, use of natural language processing, image processing methods made easy creation of rich data. Good domain knowledge is required to build useful models. Uncertainty remains around choosing the right sample data, variables reduction and selection of statistical algorithm. A suitable statistical method coupled with explaining variables is critical for model building and analysis. There are multiple choices around each parameter. An automated system which could help the scientists to select an appropriate data set coupled with learning algorithm will be very useful. A freely available web-based platform, named automated machine learning tool (AMLT), is developed in this study. AMLT will automate the entire model building process. AMLT is equipped with all most commonly used variable selection methods, statistical methods both for supervised and unsupervised learning. AMLT can also do the clustering. AMLT uses statistical principles like R2 to rank the models and automatic test set validation. Tool is validated for connectivity and capability by reproducing two published works.**

Keywords—Automated machine learning; regression models; support vector machines; QSAR; QSPR; artificial neural networks; k-means clustering; R program; shiny web app; drug design; market analysis; supervised learning; Naive Bayes classification

I. INTRODUCTION

In drug discovery and environmental toxicology, QSAR models regarded as a scientifically credible tool for predicting and classifying the biological activities of untested chemicals[1]. There is a lot of correlation and overlap between QSAR methods and general machine learning methods. In both QSAR and machine learning (ML) methods, descriptors have derived from factors that can affect the response variable. ML techniques designed to derive knowledge out of existing data on the fundamental of “Stored data becomes useful only when it has analyzed and turned into information that can make use of, for example, to make predictions” [2-4]. Many studies are available highlighting a successful application of ML techniques to diverse problems, which ranges from pharmaceutical industries to environmental sciences, mobile viruses, e-commerce, and business problems [5-7]. Machine learning methods have multiple applications like Genomic Medicine, Econometric Approach, Manufacturing, Solar radiation forecasting, big data learning, discovering phase transitions and Banking industry. Training set selection coupled with variable reduction and selection of statistical methods will make the number of models and combinations are high in number[8, 9]. Thus takes up a lot of time for model building and analysis followed by rebuilding the model with the different training set. Therefore having the tool which makes automatic data processing, training set selection, dimensional reduction, model building using multiple machine learning algorithms and model reporting will significantly save time by removing the multiple hit and trial approach. Thus the practitioner/Data scientist can focus on analysis for the progress of the project. AMLT provides users an integrated and friendly tool to build all machine learning models at one place. AMLT is freely available at "https://automatedmachine learning-gitamcse.shinyapps.io/MLPv3/"

One of the attempts made to automate the model building is in microsimulation. Modgen uses a single Montecarlo-based method to automate the micro simulation model building for different new data types, another attempt is AutoML, and automated prediction of the enzymatic functions of uncharacterized proteins, is an important topic in the field of bioinformatics. Both tools have focused only on single statistical method[10, 11].

Some challenges that data-set can contain are, e.g. missing values, high-dimensional data, mixing of numerical and text variables, non standard data, irrelevant and redundant information which may impact the performance of learning algorithms[12]. Today, most machine learning techniques handle only data with continuous and nominal values[12,13]. Missing values issue represents a very common challenge, there is a large amount of literature and practical solutions (e.g. in R) available[14, 15, 16, 17]. Pre-processing of data has a critical impact on the results. This can present challenges for the training of certain algorithms.

Proposed tool will enable practitioners to focus on the collection of good data sets and analyze the models for productivity. The tool will automate the model building process by picking the training set, variables and statistical methods that suits best to the input dataset.

A. *Defining the Problem Statement*

Performance of each algorithm depends on the data available, data pre-processing and parameter settings. The best fitting algorithm has to be found by testing various ones in a realistic data.

As of today, the generally accepted approach to select a suitable ML algorithm for a certain problem is as follows:

First, one looks at the available data and how it is described (labeled, unlabeled, available expert knowledge, etc.) to choose between a supervised, unsupervised approach.

Secondly, the general applicability of available algorithms with regard to the research problem requirements (e.g. able to handle high dimensionality) has to be analyzed. A specific focus has to be laid on the structure, the data types and overall amount of the available data, which can be used for training and evaluation.

Thirdly, previous applications of the algorithms on similar problems are to be investigated in order to identify a suitable algorithm. The term 'similar' in this case, research problems with comparable requirements e.g. in other disciplines or domains.

Another challenge is the interpretation of the results. It has to be taken into account that not only the format or illustration of the output is relevant for the interpretation but also the specifications of the chosen algorithm. Within the interpretation of the results, certain more distinct limitations (again depending on the chosen algorithm) can have a large impact. Among those are, e.g. immune to over-fitting, bias, and variance (therefore bias-variance tradeoff)[18, 19].

However, one of the promising approach to select a suitable algorithm is to look for similar and analyze what ML algorithm was used to solve it and interpretation of results. Once the algorithm is applied, based on first results different methods can be applied to improve the model. However, this is very time consuming and iterative process to compare the models for selecting the best data, right algorithm, and ease of model interpretation. Therefore, this project is proposed on automation tool which considers the good practices of machine learning methods to build best predictive model suites to the problem. Fig. 1 explains the machine learning model building.

II. Methodology

The program is written in blocks to incorporate data processing, supervised learning and unsupervised learning. Individual steps are described below.

A. Building an R program

R began in the early 1990s as the personal project of Ross Ihaka and Robert Gentleman, R is the most popular open source statistical software. R and its add-on packages provide a wide range of options for data processing, statistical methods, and high performance computing. R program, which can transform the data in a flow is mentioned in Fig. 1 will be highly useful. The program could enable practitioners to focus on the collection of data sets, descriptors calculation and analyzing the models for productivity. Automatic tool will not contain any new or modified algorithms intentionally; it uses the collection of published algorithms. That makes tool can be benchmarked against known data sets and users can easily jump on to tool to start working without having the questions related validation of algorithms.

B. Implementation and Features in the Tool

Coding the program is step by step process. Each step has few key feature implementations. Total code can be split into five main sections viz. Exploratory Data analysis (EDA), Data set selection, Feature selection, Selection of statistical method and results visualization. More details are given below.

1) Exploratory Data analysis: Returns the data header. Displays the column headings in data tables. Describes basic statistics of data i.e. summary (df), Plot Response variable/Output data to visualize distribution i.e. scatter or smooth plot.

2) Data set selection: Automatic random splitting of data into training and test sets. This splitting can be doable at different compositions of training and test tests.

3) Feature selection: Dimensionality reduction is a common technique used to reduce the number of variables in Machine Learning. The tool is equipped with multiple feature selection methods.

4) Statistical method: Most of the available statistical methods are coded together in this important section. All methods coded in a way that they process the same dataset for model building and test set validation. Parsing the single data set at onetime give the user to select the right algorithm for chosen data and set of variables. Both training and Test set validation is also automated to enable the user to choose the right model. This section divided into Classification, Regression and Clustering techniques. In classification, it supports (K-NN, Random Forest, Naive Bayes, SVM, ANN). In Regression, it supports (Linear Regression, Logistic Regression, Random Forest Regression, PCR, and PLSR). In clustering, tool supports K-MEANS Clustering

5) Results visualization: It's always very important to visualize the results in a manner results can be compared and interpreted. This is more important especially when multiple models are generated for training set and test set validation happens automatically.

Results visualization includes, For Classification TP= True Positive, FP=False Positive, TN=True Negative, FN=False Negative, Q=Q-Value, SE=Sensitivity. For Regression R2=R-Squared Value, RMSE=Root Mean Squared Error.

6) Export results: All algorithm results are exported at once to .csv file with predicted output added with testing data. Fig. 2 depicts the flow of the data how it is implemented.

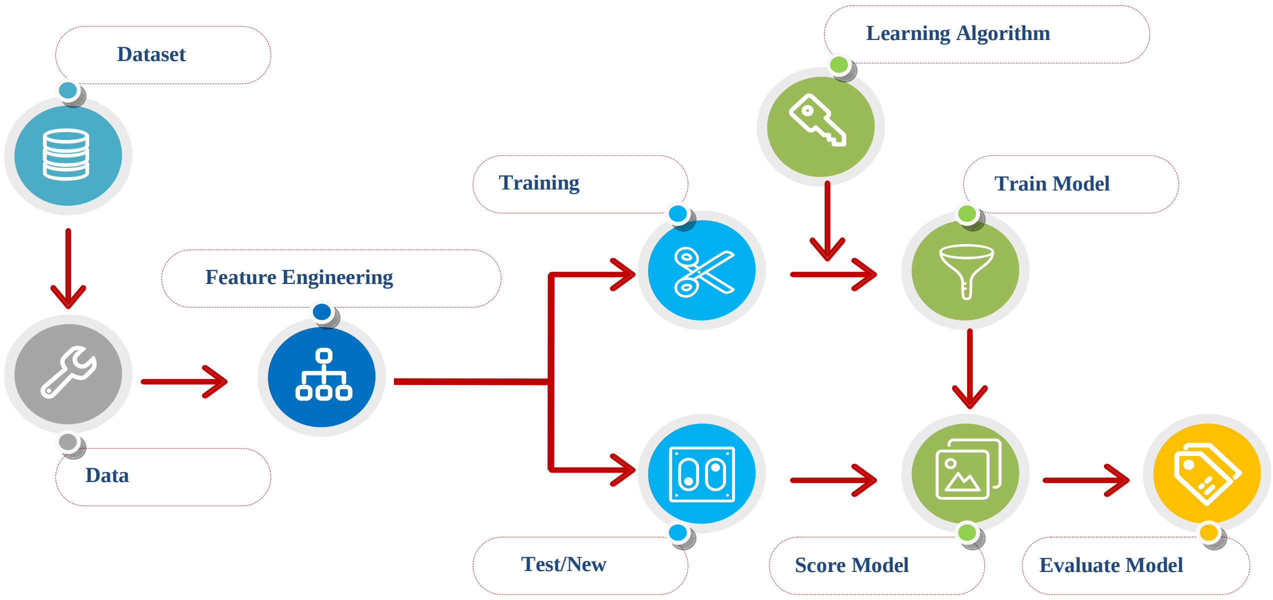


Fig. 1. Example flow for Machine Learning Model Building.

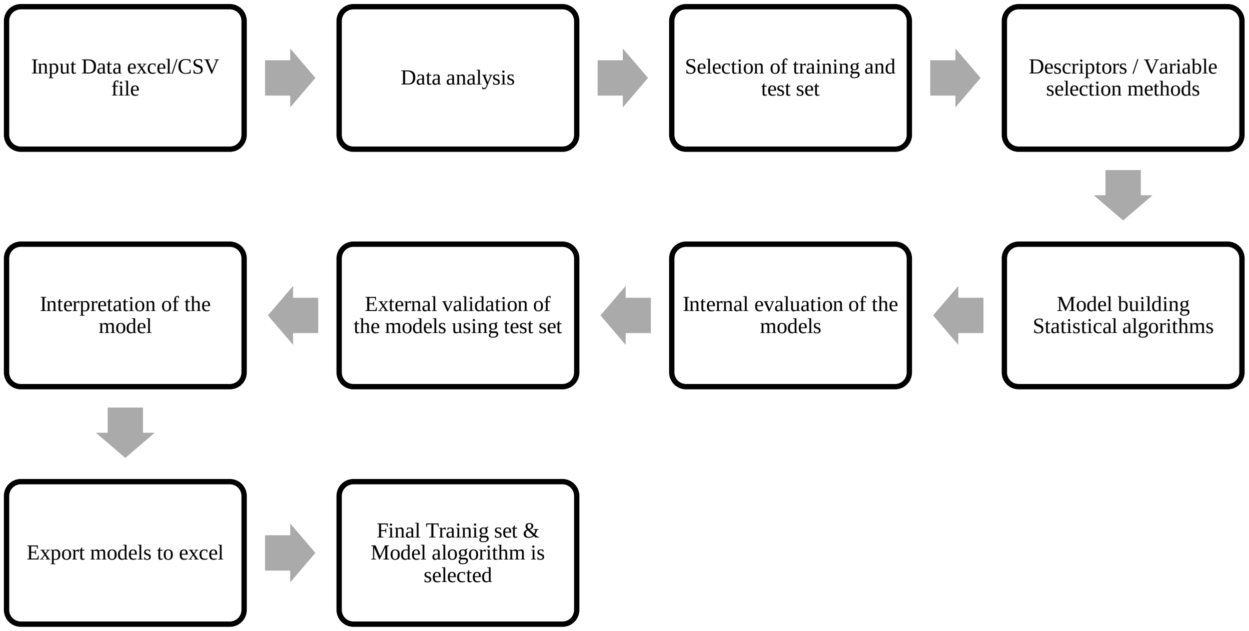

Input Data excel/CSV file
Data analysis
Selection of training and test set
Descriptors / Variable selection methods
Model building Statistical algorithms
Internal evaluation of the models
External validation of the models using test set
Interpretation of the model
Export models to excel
Final Trainig set & Model alogorithm is selected

Fig. 2. Flow Chart Implemented in Automated Machine Learning Program Architecture.

III. Validation of the Program

The functionality of the program and connection between different flows was verified with two data sets, one for blood-brain barrier (BBB) penetration and the other for Caco2 permeability [20, 21]. Both of these are of two different data types. Viz. are classification based data and quantitative data, respectively. Both of these datasets have published statistical models using manual model building approach. It will be good validation to test the program for reproducibility of published models. The program can also be tested for how automation of model building could benefit the model building for these datasets. It will also help to check if a new algorithm could result in a better model than published.

A. Description of the First Dataset

Dataset1 consists of 1839 molecule entries with fingerprint-based calculated descriptors. Data set is with known blood-brain barrier (BBB) penetration data is collected. The entire dataset was collected from Shen's work[20], which included 1438 BBB+ and 401 BBB- compounds. In the published work by Shen's group has use support vector machine (SVM) algorithm is applied to predict the blood-brain barrier (BBB) penetration. The built model could able to explain the training set and test set with a Q value of 0.9429. Overall, predictive accuracies of the best BBB model for the training and test sets were 98.8 and 98.4%.

B. Description of the Second Dataset

Dataset2 consists of Caco2 permeability, which is an important parameter, needs to be assessed for estimating the new chemical entities druggability.

Hai Pham The, et al. [21] were managed to general build a QSAR model for caco2 permeability. 21 QSAR models were with discriminate compounds with high Caco-2 permeability (Papp≥8*10−6 cm/s) from those with moderate-poor permeability (Papp<8*10−6 cm/s) were developed on a novel large dataset of 674 compounds. A general model combining all types of molecular descriptors was developed and it classified correctly 81.56 % and 83.94 % for training and test sets, respectively [21].

C. Model Building for Dataset 1

A dataset with 1839 molecules used to make an automated model building using AML tool (Automated Machine Learning). Total 287 descriptors are calculated using the Chemistry Development Kit (CDK) software.

Data statistics module used to understand the data distribution and the head part of the data. This module is also helpful to understand the data summary like mean, the median of the overall data divided into four quarters. In Fig. 3, output distribution map is depicted which is helpful to understand the dependent variable distribution. In this case, it is categorical distribution i.e. BBB+ or BBB- . For the model, building this kind of data is considered as one and zero internally.

Fig. 4 show cases the primary view of the tool. This shows Data statistics module and results in the side plane. For model building instance 1, data divided into 70:30 percentage for training and test sets respectively. All the classification models implemented in AML tool were applied. In this case, the feature selection module not applied, as this is a classification model building.

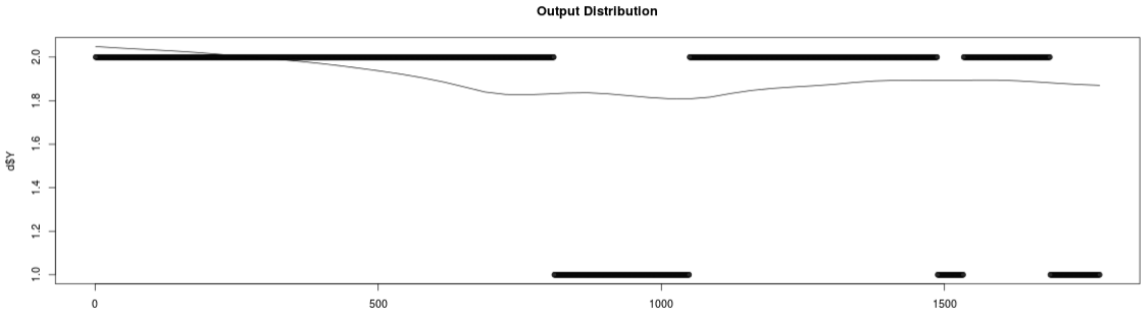
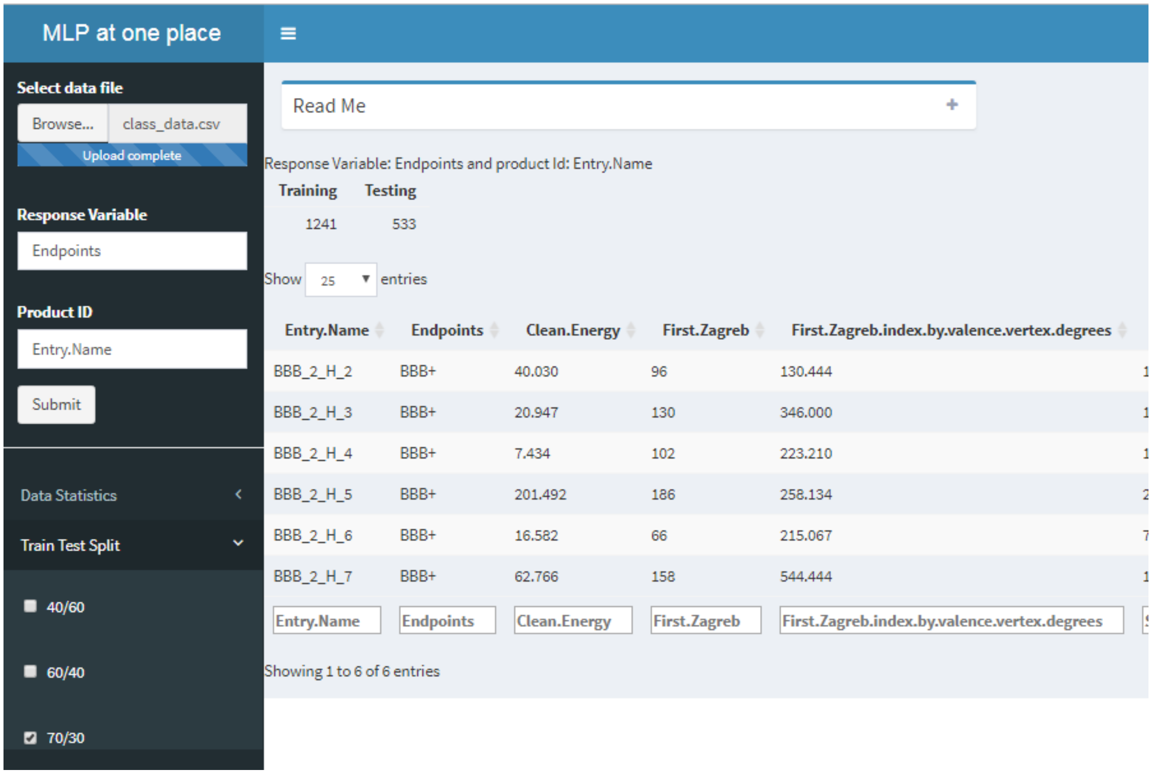


Fig. 3. Output Distribution Map Illustrates Two-Point Distribution of Data.



Response Variable: Endpoints and product Id: Entry.Name

Training	Testing
1241	533

Entry.Name	Endpoints	Clean.Energy	First.Zagreb	First.Zagreb.index.by.valence.vertex.degrees	
BBB_2_H_2	BBB+	40.030	96	130.444	1
BBB_2_H_3	BBB+	20.947	130	346.000	1
BBB_2_H_4	BBB+	7.434	102	223.210	1
BBB_2_H_5	BBB+	201.492	186	258.134	2
BBB_2_H_6	BBB+	16.582	66	215.067	7
BBB_2_H_7	BBB+	62.766	158	544.444	1
Entry.Name	Endpoints	Clean.Energy	First.Zagreb	First.Zagreb.index.by.valence.vertex.degrees	

Showing 1 to 6 of 6 entries

Fig. 4. Snapshot of BBB Model Data Selection Page.

Fig. 5 illustrates the multiple models built in less than a min. Model performance with confusion-matrix validation also generated at the same time, including test set performance assessment for different models. Classification model assessment was done using most commonly used methods like confusion matrix, Precision value, and Cohen's kappa coefficient (κ) value.

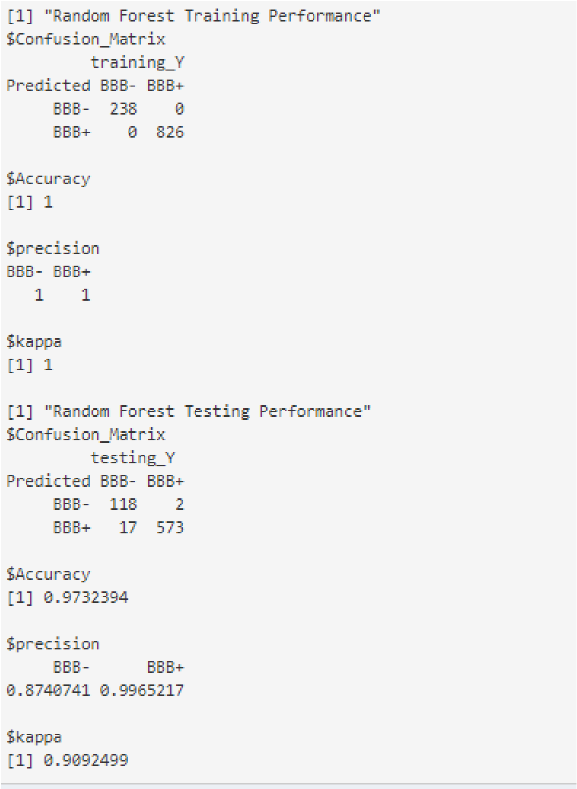
AML tool could able to generate 7 models in a flash of time. The user can pick the best performing model after observing both training and test performance. In this case, Random Forest classification algorithm is best performing with an accuracy of 1 and 0.96 for training and test sets respectively.

In instance 2, data divided into 60:40 percentage training and test sets respectively. Total 1064 molecules are in the training set. All 7 models were generated using AML tool. Again Random forest algorithm could able to explain the data well over SVM methods, KNN and Naive Bayes methods.

Interestingly random forest method accuracy performance was improved to test set 0.97 compared to 0.96 with 30% data in the test set. During instance 3, where training set reduced to 40% has also reduced algorithm test set performance to 0.95 accuracies. Hence, it is identified as Random forest method with 60:40 split of data is giving the good model. In AML tool all these three instances are created, users can able to identify the best algorithm suites to the problem and data set of interest by covering all machine learning space in a few minutes of times.

Total 42 models were generated using AML tool in three instances with a variation in %training set. Random forest algorithm couple with %60 in training set could able to generate the model with %accuracy >80 for both training and tests. Automatic validation and results generation will projected for test set, all models % accuracy is plotted in Fig. 5.

Comparison with published model: When compared the model performance with published models. AML could able to reproduces the equivalent model to the published one, where the test set Q value of 0.9429. The new model developed has test set accuracy of 0.97. Hence, this result validates the performance of the tool and connectivity between workflows and an alternative algorithm to explain the BBB permeation data is established. The earlier publishers did not try random forest models. Fig. 6 shows the Random forest algorithm results in AML tool. Tables I-III compares the different models using different training and test set compositions generated using AML tool.



```
[1] "Random Forest Training Performance"
$Confusion_Matrix
         training_Y
Predicted BBB- BBB+
     BBB-  238    0
     BBB+    0  826

$Accuracy
[1] 1

$precision
BBB- BBB+
   1    1

$kappa
[1] 1

[1] "Random Forest Testing Performance"
$Confusion_Matrix
         testing_Y
Predicted BBB- BBB+
     BBB-  118    2
     BBB+   17  573

$Accuracy
[1] 0.9732394

$precision
     BBB-      BBB+
0.8740741 0.9965217

$kappa
[1] 0.9092499
```

Fig. 5. Model Building Results Output Page with Confusion-Matrix Validation.

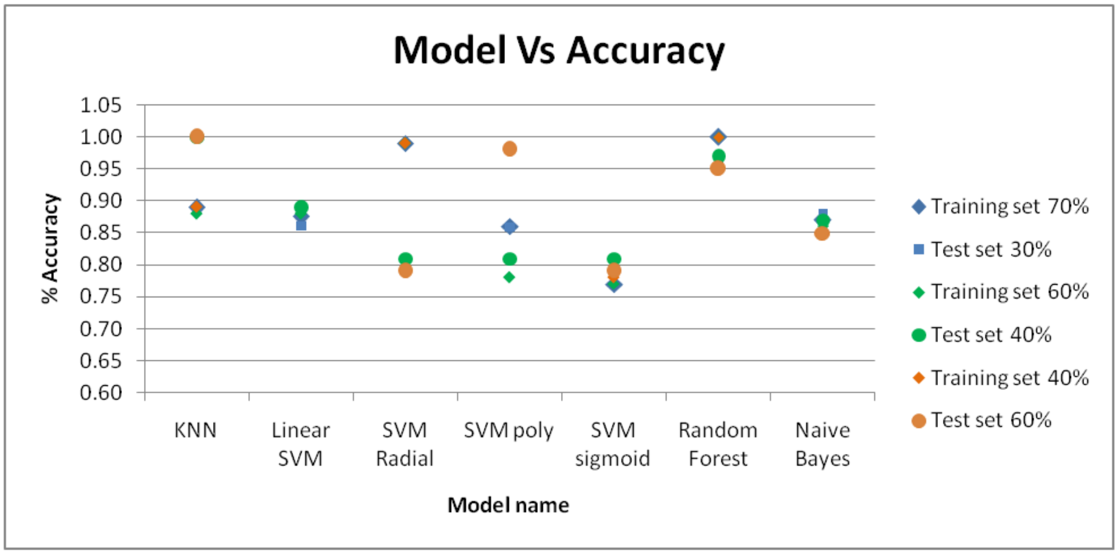


Fig. 6. Multiple Algorithms Predicted Accuracy was Compared Against Training Set Selection. Random Forest Method with 60% Training Set could able to Give Better Accuracy.

TABLE. I. TRAINING AND TEST SET 70:30

	Training set			Test set		
Model	Training set 70% **Accuracy**	**Precision (BBB-, BBB+)**	**Cohen's kappa coefficient (κ)**	Test set 30% **Accuracy**	**Precision (BBB-, BBB+)**	**Cohen's kappa coefficient (κ)**
KNN	0.89	0.729 , 0.946	0.70	1.00	1, 1	1.00
Linear SVM	0.88	0.854, 0.882	0.67	0.86	0.765,0.882	0.58
SVM Radial	0.99	0.996,1.0	0.99	0.81	0,1	0.00
SVM poly	0.86	0.587, 0.946	0.58	0.86	0.489,0.942	0.47
SVM sigmoid	0.77	0,1	0.00	0.81	0,1	0.00
Random Forest	1.00	1,1	1.00	0.96	0.785,1.000	0.85
Naive Bayes	0.87	0.697,0.921	0.62	0.88	0.612,0.949	0.59

TABLE. II. TRAIN TEST 60:40

	Training set			Test set		
Model	**Accuracy**	**Precision (BBB-, BBB+)**	**Cohen's kappa coefficient (κ)**	**Accuracy**	**Precision (BBB-, BBB+)**	**Cohen's kappa coefficient (κ)**
KNN Training Performance	0.88	0.705, 0.940	0.66	1	1,1	1
Linear SVM	0.88	0.689,0.940	0.65	0.89	0.637, 0.949	0.62
SVM Radial	0.99	0.995,1.000	0.99	0.81	0.007, 1.000	0.01
SVM poly	0.78	0.033, 1.000	0.05	0.81	0.022, 1.000	0.03
SVM sigmoid	0.77	0,1	0	0.81	0,1	0
Random Forest	1	1,1	1	0.97	0.874, 0.996	0.91
Naive Bayes	0.86	0.705, 0.906	0.61	0.87	0.651, 0.930	0.59

TABLE. III. TRAIN TEST 40:60

	Training set			Test set		
Model	Training set 40% **Accuracy**	**Precision (BBB-, BBB+)**	**Cohen's kappa coefficient (κ)**	Test set 60% **Accuracy**	**Precision (BBB-, BBB+)**	**Cohen's kappa coefficient (κ)**
KNN	0.89	0.686, 0.948	0.66	1.00	1,1	1.00
Linear SVM	0.57	0.773, 0.522	0.18	0.55	0.743, 0.505	0.15
SVM Radial	0.99	0.993, 1.000	0.99	0.79	0,1	0.00
SVM poly	0.98	1.000, 0.978	0.95	0.98	1.000, 0.985	0.96
SVM sigmoid	0.78	0,1	0.00	0.79	0,1	0.00
Random Forest	1.00	1,1	1.00	0.95	0.801 0.995	0.85
Naive Bayes	0.85	0.753, 0.881	0.59	0.85	0.666, 0.902	0.56

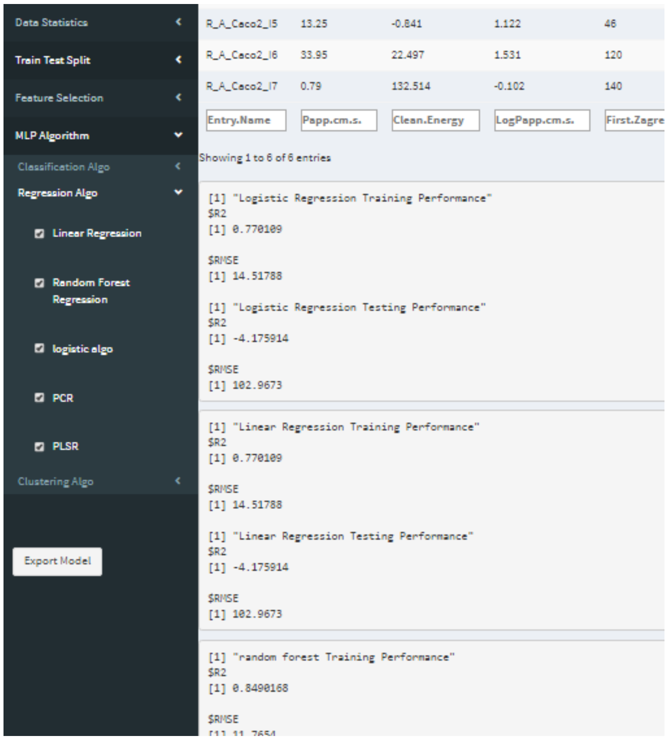
D. Model Building for Dataset 2

Total 674 molecules used to build a regression model for estimating the Caco2 permeability of chemical entities. Descriptors are calculated using CDK software. Using data distribution in AML tool, it was decided to use initially split the data to 60:40 training set and test set.

Total five different algorithms are applied at the same time for model building viz. Linear Regression, Logistic Regression, Random forest, PCR and PLS regression methods which is clearly shown in Fig. 7.

- Caco2 model results.

Random forest algorithm could able to explain training set with R2value of 0.83, with a good RMSE value for test set prediction. Random forest model could able to explain the data much better than the published caco2 model for the same reference set viz. R2 of 0.564, RMSE 0.339. However new descriptors are added to the data set which could improve the model dataset. Tables IV-VI, Fig. 8 and Fig. 9 shows the comparison of other models generated and test set performance results.



R_A_Caco2_I5	13.25	-0.841	1.122	46
R_A_Caco2_I6	33.95	22.497	1.531	120
R_A_Caco2_I7	0.79	132.514	-0.102	140
Entry.Name	Papp.cm.s.	Clean.Energy	LogPapp.cm.s.	First.Zagre

Showing 1 to 6 of 6 entries

```
[1] "Logistic Regression Training Performance"
$R2
[1] 0.770109

$RMSE
[1] 14.51788

[1] "Logistic Regression Testing Performance"
$R2
[1] -4.175914

$RMSE
[1] 102.9673
```

```
[1] "Linear Regression Training Performance"
$R2
[1] 0.770109

$RMSE
[1] 14.51788

[1] "Linear Regression Testing Performance"
$R2
[1] -4.175914

$RMSE
[1] 102.9673
```

```
[1] "random forest Training Performance"
$R2
[1] 0.8490168

$RMSE
[1] 11.7654
```

Fig. 7. Snapshot of the Tool Shows different Regression Algorithms Generated in AML. Picture also shows Clustering and Classification Algorithm Tabs.

TABLE. IV. TRAIN TEST CACO2_70:30

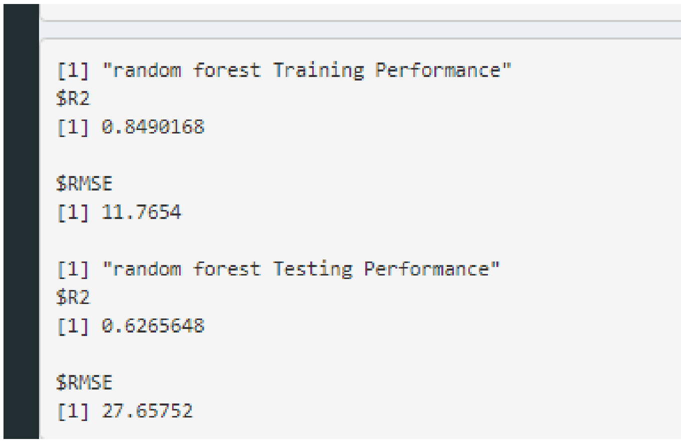
	Training set 70%		Test set 30%	
Model	**R2**	**RMSE**	**R2**	**RMSE**
Logistic Regression	0.77	14.52	-4.18	102.97
Linear Regression	0.77	14.52	-4.18	102.97
random forest	0.85	11.77	0.63	27.66
PCR	0.32	24.88	0.32	24.88
PLSR	0.36	24.26	0.17	24.26

TABLE. V. TRAIN TEST CACO2_60:40

	Training set 60%		Test set 40%	
Model	**R2**	**RMSE**	**R2**	**RMSE**
Logistic Regression	0.77	14.48	-2.90	84.95
Linear Regression	0.77	14.48	-2.90	84.95
random forest	0.84	12.30	0.66	25.34
PCR	0.33	24.91	0.33	24.91
PLSR	0.37	24.14	0.12	24.14

TABLE. VI. TRAIN TEST CACO2_40:60

	Training set 40%		Test set 60%	
Model	**R2**	**RMSE**	**R2**	**RMSE**
Logistic Regression	0.80	13.75	-4.38	89.70
Linear Regression	0.80	13.75	-4.38	89.70
random forest	0.84	12.13	0.69	21.25
PCR	0.32	25.21	0.32	25.21
PLSR	0.37	24.28	0.19	24.28



```
[1] "random forest Training Performance"
$R2
[1] 0.8490168

$RMSE
[1] 11.7654

[1] "random forest Testing Performance"
$R2
[1] 0.6265648

$RMSE
[1] 27.65752
```

Fig. 8. Model Building Results Output Page with Confusion Matrix Validation.

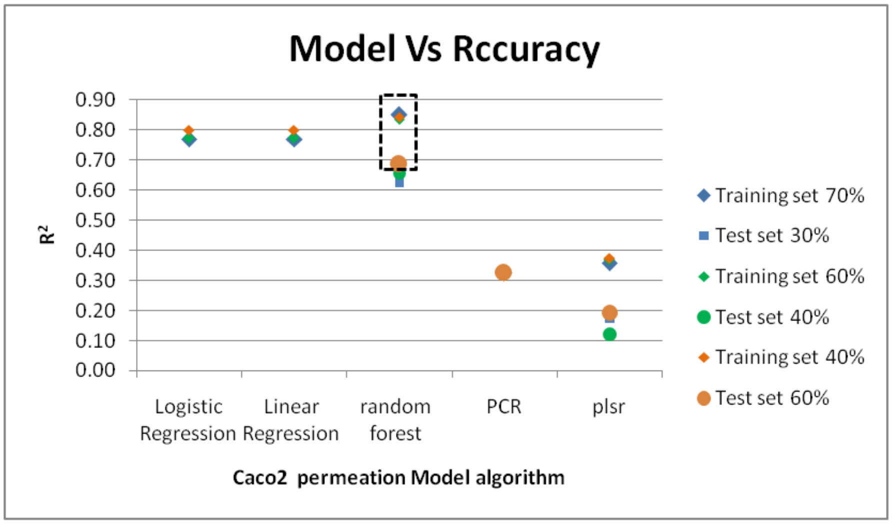


Fig. 9. Multiple Algorithms Predicted Accuracy was Compared Against Training Set Selection. Random Forest Method with 60% Training Set Could Able to Give Better Accuracy.

IV. Conclusion

Considering multiple machine learning applications, it is recommended to automate the model building process. Automated machine learning (AML) tool is developed and deployed in web portal. The tool is validated for technical capabilities, program stability and tested for seamless connectivity for automating the model building process. The pipeline and interface provides the means to (i) Perform initial data analysis, ii) Identify the data split, (iii) Selection of suitable variable selection method, (iv) Selection of suitable statistical algorithm for model building, (v) Model selection & interpretation. Program is validated against published models and datasets for do-ability and model reproducibility of published models. The tool not only able to reproduce the published results also suggested alternative algorithms, which can explain data variability up to 90% accuracy for training and test sets. Validation carried out on both regression and classification models. AML tool is also tested for potential bugs and abnormal shutdowns. AML tool has potential to generate all kinds of machine learning models at one place; this can be a first place to start with and get an initial combination about data and suitable algorithm to explain the data variations. The workflow is highly flexible, permitting modifications such as a choice of data set, level of theory, validation, or model selection. This can be used for large data sets, by doing the sampling of data from big databases. The tool is hosted in web portal "https://automatedmachinelearningitamcse.shinyapps.io/MLPv3/". The tool can be accessed by anyone who has access to the website.

References

[1] Ravi Shekar Ananthula, Kishore Madala. (2008) Strategies for generating less toxic P-selection inhibitors: Pharmacophore modeling, virtual screening and counter pharmacophore screening to remove toxic hits. Journal of molecular graphics & modelling. 27(4)546.

[2] Abdelrahman, M. A., Salama, I., Gomaa, M. S., Elaasser, M. M., Abdel-Aziz, M. M. and Soliman, D. H. (2017) Design, synthesis and 2D QSAR study of novel pyridine and quinolone hydrazone derivatives as potential antimicrobial and antitubercular agents. Eur J Med Chem. 138 698-714.

[3] Alpaydin, E. (2010) Introduction to machine learning. MIT Press2nd ed.

[4] Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N. and Lloyd, S. (2017) Quantum machine learning. Nature. 549(7671)195-202.

[5] Galvan, R. A. R. M. J. M. V. I. M. A Study of Machine Learning Techniques for Daily Solar Energy Forecasting Using Numerical Weather Models. Intelligent Distributed Computing VIII. DOI: 10.1007/978-3-319-10422-5_29269-278.

[6] Gong, G. I. J. L. M. C. H. H. Z. Z. P. (2019) A review on machine learning methods for in silico toxicity prediction. Journal of Environmental Science and Health, Part C 36(4)169-191.

[7] RT., M. G. S. (2019) A user-generated data based approach to enhancing location prediction of financial services in sub-Saharan Africa. Appl Geogr.105 25-36.

[8] CH, L. S. L. J. C. Y. Y. H. K. J. J. (2019) Machine learning models based on the dimensionality reduction of standard automated perimetry data for glaucoma diagnosis. Artif Intell Med.94 110-116.

[9] K, Z. M. T. Y. K. H. H. (2018) Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer. Cancer Inform.17 1-7.

[10] modgen. https://www.statcan.gc.ca/eng/microsimulation/modgen/modgen.

[11] Eggensperger, M. F. A. K. K. (2015) Efficient and Robust Automated Machine Learning. Advances in Neural Information Processing Systems 28 (NIPS 2015)1-9.

[12] Jordan, M. I. (2015) Machine learning: Trends, perspectives, and prospects. Science255-260.

[13] Koohy, H. (2018) The rise and fall of machine learning methods in biomedical research. F1000Research. 6 1-16.

[14] Vink, R. M. S. P. L. G. (2018) Generating missing values for simulation purposes: a multivariate amputation procedure. Journal of Statistical Computation and Simulation. 88(15)2909-2930.

[15] De Silva AP;, M.-B. M. D. L. A. L. K. S. J. Multiple imputation methods for handling missing values in a longitudinal categorical variable with restrictions on transitions over time: a simulation study. BMC Med Res Methodol. 19(1).

[16] Kang, H. (2013) The prevention and handling of the missing data. Korean J Anesthesiol.64(5)402-406.

[17] Hogan, C. J. H. L. E. C. J. W. (2015) Are all biases missing data problems? Curr Epidemiol Rep. 2015 Sep 1; 2(3): 162–171.2(3)162-171.

[18] JJ, K. H. S. S. G. R. S. S. G. (2016) Dimension reduction and shrinkage methods for high dimensional disease risk scores in historical data. Emerg Themes Epidemiol. 2016 Apr 5;13:5. doi: 10.1186/s12982-016-0047-x. 13 5.

[19] Hawkins, D. M. (2004) The Problem of Overfitting. J. Chem. Inf. Comput. Sci.44(1)1-12.

[20] Shen, J. C., F.; Xu, Y.; Li, W.; Tang, Y.;. (2010) Estimation of ADME properties with substructure pattern recognition. J Chem Inf Model. 50(6)1034-41.

[21] Pham The, H., Gonzalez-Alvarez, I., Bermejo, M., Mangas Sanjuan, V., Centelles, I., Garrigues, T. M. and Cabrera-Perez, M. A. (2011) In Silico Prediction of Caco-2 Cell Permeability by a Classification QSAR Approach. Mol Inform. 30(4)376-85.

Towards a Powerful Solution for Data Accuracy Assessment in the Big Data Context

Mohamed TALHA[1], Nabil ELMARZOUQI[2], Anas ABOU EL KALAM[3]
ENSA Marrakech, Cadi Ayyad University, Marrakech, Morocco[1]
ENSET Rabat, Mohammed V University in Rabat, Rabat, Morocco[2]
ENSA Marrakech, Cadi Ayyad University, Marrakech, Morocco[3]

***Abstract*—Data Accuracy is one of the main dimensions of Data Quality; it measures the degree to which data are correct. Knowing the accuracy of an organization's data reflects the level of reliability it can assign to them in decision-making processes. Measuring data accuracy in Big Data environment is a process that involves comparing data to assess with some "reference data" considered by the system to be correct. However, such a process can be complex or even impossible in the absence of appropriate reference data. In this paper, we focus on this problem and propose an approach to obtain the reference data thanks to the emergence of Big Data technologies. Our approach is based on the upstream selection of a set of criteria that we define as "Accuracy Criteria". We use furthermore a set of techniques such as Big Data Sampling, Schema Matching, Record Linkage, and Similarity Measurement. The proposed model and experiment results allow us to be more confident in the importance of data quality assessment solution and the configuration of the accuracy criteria to automate the selection of reference data in a Data Lake.**

***Keywords*—Big data; data quality; data accuracy assessment; big data sampling; schema matching; record linkage; similarity measurement**

I. Introduction

Data Quality is an essential topic for any organization to get accurate data and to make good decisions accordingly. Concerns about this topic can be addressed in three ways: data quality evaluation, data quality improvement and data protection [1]. In this work, we will focus on the evaluation of data quality. Different dimensions characterize the quality of the data but there is no general agreement on the complete list of these dimensions and their exact meanings [2]. The most studied dimensions in literature are Accuracy, Completeness, Currency and Consistency [3], [4]. Several studies have identified Accuracy as the key dimension of Data Quality [1], [5], [6], [7] and the hardest to assess [8]. Knowing beforehand the Accuracy of their data brings many benefits to organizations. Indeed, the decision-making process is improved when we know the degree of confidence that can be attributed to the data. On the other hand, inaccurate data can lead to erroneous conclusions and can significantly compromise a range of decision-making processes which can lead to lost opportunities, lost revenue, strategic mistakes, etc. Thus, evaluating data accuracy is a process that requires planning before any data exploitation. It is with these elements in mind that we will devote this work to the evaluation of data accuracy.

Many works in literature attempt to find solutions to evaluate the accuracy of data. Not all, but most require a key step of comparing the data to evaluate with the correct data without giving sufficient details on how to obtain this reference data. Today, thanks to the emergence of Big Data, Cloud Computing and IoT, organizations can more easily collect, store and manage very large volumes of data. In this article, we will offer a solution to obtain reference data in a Big Data environment. After exposing an overview of the related work, we will present a state of the art about data accuracy in order to deduce a clear definition that will be our basis for this work. We will then present a set of techniques and concepts necessary for the implementation of our solution. Next, we will detail our model that we will apply on a case study in order to experiment our approach. The analysis of the results will allow us to deduce a set of very interesting findings to successfully implement our solution for any other use case. We will end this paper with a conclusion and a glimpse into future work.

II. Related Work

Many studies are interested in evaluating the quality of data, particularly looking into the accuracy of the data. The authors of [8] propose a data accuracy assessment tool based on a collection of datasets and three different phases: training, record linkage and accuracy assessment. The idea is to be able to identify the reference data which will then serve as a basis for the comparison process. In their approach, they use machine learning techniques to choose, from the datasets already present in the lake of an organization, the closest dataset that they deem correct. This assumption can be correct if it is guaranteed that the data present in the lake are correct and up to date, which is generally not the case; the new data collected may even be of better quality in some cases. In addition, the correct information can exist in more than one dataset, not just one as they assume. In this case, it will be necessary to manage many aspects such as the heterogeneity of the schemas, the choice of the dataset offering the best quality for a particular data perimeter, the lack of correspondence for certain data, etc. Moreover, the use of Google’s Word2Vec word embedding as a basis for data training can slow down the assessment process. The idea of using word embedding to determine the closest dataset remains logical but it will be necessary to test its performance in an environment that hosts very large volumes of data.

Another work done by Taleb et al. [9] in which they propose a system for assessing the accuracy, completeness and consistency of Big Data based primarily on data sampling. The adopted principle is to create a set of samples, without replacement, from the original dataset. Then, from each sample created, generate a set of samples using the BLB (Bag of Little Bootstraps) resampling technique. For each sample thus generated, a data profiling process is applied to extract descriptive information from the data such as the description of the data format, the different attributes, their types and their values, the possible constraints, the ranges of the authorized values, etc. All this information obtained through the profiling process is then used to select the appropriate metrics for each dimension before proceeding with their evaluation. For the accuracy dimension, a metric can be defined to satisfy a certain number of constraints related to the type of data such as a zip code, an email, a social security number, or an address. For example, an attribute can be defined as a range of values between 0 and 100, otherwise it is incorrect. The accuracy of the attribute is then calculated based on the number of correct values divided by the number of observations or rows. The authors of this article have only dealt with syntactic accuracy which is much simpler to verify than semantic accuracy as we will see a little further.

One last interesting work we wish to present concerns the evaluation of the quality of unstructured data. In [19], the authors are interested in evaluating the quality of data collected from social networks through the integration of a metadata management system into Big Data architecture. In their approach, the authors distinguish five groups of metadata:

- Navigational metadata used to identify the location of each dataset.
- Process metadata used to describe the source and the processing performed on each dataset.
- Descriptive metadata consists of business metadata that describes the meaning of a dataset from a business perspective, and technical metadata that provides technical information about the dataset such as data size, content description, data creator, data type and format of content, etc.
- Quality metadata including dimensions and metrics used to describe the quality of the data.
- Administrative metadata used to describe the data provider, applicable licenses and access rights on the datasets, the copyright holder and the data privacy level indicator, etc.

The use of metadata to evaluate the quality of unstructured data seems to be a good solution especially when it is difficult, if not impossible, to compare these data with data that represent the real world. However, managing metadata could be a very expensive and complex process especially for quality metadata. The high volume and velocity of Big Data are real challenges to overcome. The use of metadata in conjunction with other techniques of comparison with correct data seems to us to be more efficient.

There are many other works in literature that are concerned with assessing the accuracy of data. For example, in [1], Motro and Rakov present a solution for assessing the accuracy and completeness of databases using Set Theory. Redman, for its part, provides in [6] a framework for assessing the accuracy of data based on four factors, namely where to take measurements, the choice of data to include, the measurement device and the scales on which the results are reported.

III. Data Accuracy in Literature

A. Definition

A multitude of definitions for data accuracy exist in literature. Each definition involves aspects of the context in which it was given. Generally, there is a reasonable consensus that the accuracy of the data is linked to a specific concept, namely the magnitude of an error [10]. For Ballou and Pazer [11], data are accurate when their values stored in a database correspond to the real values. Authors in [12], [13], [14] and [15] link the accuracy of data to the percentage of objects that do not contain errors in the data such as misspellings, values outside the allowed range, and so on. Several works [4], [5], [7], [16] define accuracy as a measure of the proximity of a given value ν to another value ν' considered to be correct.

Furthermore, ISO [17] and many studies [2], [5], [16] distinguish between syntactic accuracy and semantic accuracy. Each of them presents a particular aspect of the accuracy and has its own metrics. Syntactic accuracy is defined as the proximity of data values to a set of defined values in a domain considered to be syntactically correct. It concerns the structure of the data [18] and expresses the degree of syntactic error-freeness [14]. Semantic accuracy, as for it, is defined as the proximity of data values to a set of defined values in a domain considered to be semantically correct. It represents the correctness and the degree of validity of the data [12], [14]. It describes the extent to which data represent real-world conditions [18].

Although they diverge on some particularities, all of these definitions implicitly or explicitly imply a comparison between the data of a system and the real world. Therefore, we adopt the following definition: Accuracy reflects the degree of correctness at which the data in an information system represents the real world. More formally, let v be the value of a datum in an information system and ν' the corresponding reference value considered as correct; the accuracy of ν represents the degree of similarity between ν and ν'.

Whether for structured or semi-structured data, comparing the values of the data with those of the real world allows us to deduce the degree of accuracy. However, this definition does not seem well-suited to unstructured data such as files with free text (tweets, studies, personal reports, etc.), multimedia files (image, audio or video), etc. Certainly unstructured data may contain information that can be compared with the real world (a person's photo, information about an object, information in a story, a mathematical formula, etc.) but the problem lies in the information that only concerns the people who gave it (personal impressions, opinions about a subject, intentions, desires, etc.). Unstructured data is more complex to evaluate and cannot be evaluated in the same way as structured or semi-structured data. If we take the example of a scientific paper, we cannot evaluate its quality by analyzing its content or by comparing the information it contains with reference data.

Instead, we must analyze some information attached to it, such as the opinions of the reviewers, the importance of the magazine in which it was published, the reputation of the authors and their institutional affiliations, the type of publisher (academic, commercial ...), the source of the article (peer-reviewed journals, unpublished articles …), etc. and anything else that is relative to it. Thus, the evaluation of unstructured data quality will be more relevant if it concerns the information relating to the data rather than the data themselves. In [19], confirming our hypothesis, the authors present a solution to evaluate the quality of data collected from social networks by integrating a metadata management system in the Big Data life cycle. For this reason, and in order to limit the scope of this work, we will remain focused on structured and semi-structured data.

B. *Reference Data*

According to Redman [6], it is impossible to tell by direct examination if a data value is correct; all measurements of data accuracy must refer to human knowledge, other reference data or the real world. Comparing the data values with real-world values makes the measurement of their accuracy complex and costly because, very often, these real values are unknown [3], [20] or are hypothetical, really unavailable [1]. The degree of complexity changes according to the type of accuracy to assess; syntactic accuracy is usually simpler than semantic accuracy. Indeed, it can be verified by comparing the data values with reference dictionaries such as name dictionaries, domain dictionaries (list of product categories, commercial categories ...), address list, range of values, etc. On the other hand, semantic accuracy is more complex to measure because the terms of comparison must be derived from the real world, which is almost always costly [6].

One systematic way to verify semantic accuracy when multiple data sources are available is to compare information about the same instance stored in different locations. According to [3], a typical process for checking semantic accuracy consists of two phases: a searching phase and a matching phase. The first one is to identify the matching instances, while the second one is to make a decision on correspondence, non-correspondence or possible correspondence. Different criteria can be applied to make the comparisons. Generally, the values are considered correct if they come from a reliable source. In some cases, a data expert may be required to estimate the accuracy.

Moreover, when collecting data, there is no guarantee that the information collected is accurate. Today, to check the accuracy of the information provided by the consumers of a service, new methods are used such as automatically sending a secret code by mail to check a postal address, by email to check an email address, by SMS or voice call to check a phone number, etc. These methods, while effective in data collection, cannot be applied to estimate the accuracy of data already collected because of their high cost and the time it may take. They are therefore not suitable for checking the timeliness of information. In practice, the comparison is made with data collected from a reference source considered sufficiently reliable [21]. When multiple data sources provide the same types of information, the most reliable ones can be considered as data references for comparisons. The reliability of data sources can be determined through the trust and the reputation of the information provider. Other strategies include considering other quality factors to determine the most reliable source of data, for example, data consistency [4].

Reference data can be obtained through different mechanisms such as:

- Identification: we can assign to each data set available in an organization's information system a reliability level. Data sets with a high level of reliability can be used as reference data during the accuracy evaluation process.
- Collection: if reference data do not exist, some reference dictionaries such as addresses (postal codes, city names and codes, streets, ...), product catalogs, lists of names and surnames, the possible values for certain fields (diplomas, professional activities, ...), etc. can be collected independently and serve as reference data for checking syntactic inaccuracies. Business information can also be collected from external reliable providers to update data or fill in missing values.
- Correction: obtaining reference data is possible thanks to the improvement of the data quality of an organization. This can be done by implementing different techniques such as data cleaning, updating obsolete data values, correcting incorrect values, etc.

Reference data represent then the reality and make it possible to evaluate the accuracy of others data by calculating their degree of similarity. We consider this as the basic element for any process of evaluating the accuracy of structured and semi-structured data. In the next section, we outline the main accuracy criteria that allow us to identify reference data.

C. *Accuracy Criteria*

In this section, through examples, we present some of the criteria that data and their providers must meet in order to consider them as reference data.

The very first criterion one can think of is reliability in the source of the data. If the data are collected from a competent source that we are sure will provide correct data, we can consider them as reference data. As an example, to verify someone's personal identity information, the best solution is to compare them with the data of the civil registry office of his/her country. However, this operation can be more complex for other cases. For example, to verify the validity of the diplomas declared by a person, the ideal is to validate this information with the institutions having issued these diplomas which can be a very difficult or impossible task. It would then be easier to verify it with an organization providing this service and having a good reputation for doing so. We deduce from these two examples that the Trust and the Reputation of data providers are two key criteria for identifying reference data.

Now let us take another illustrative example using financial market data which has instrument values that change continuously. When making a financial decision, traders usually rely on market data prepared by the internal departments of their organizations. Financial institutions, during orders validation process, require a crucial step of validating the data upon which traders make their decisions, referring to up-to-date data acquired from third-party organizations such as Bloomberg, Thomson Reuters, CQG, etc. who are specialized in this field and guarantee a real-time update of their data. The third criterion to add to our list is the Timeliness of the data. This time it is a criterion that relates to the data itself and not the data providers.

Lastly, in Big Data environments, many data providers can feed an organization's data lake with information about subjects that may be redundant or contradictory. To determine the most reliable source, we can of course refer to the criteria mentioned in the examples above, but if this is not possible, we can take into account other quality factors such as data Consistency [4].

We understand that different criteria related to the data providers or the data themselves can be used to identify the reference data. Among others, these criteria include:

- Trust: this criterion plays a central role in assessing the quality of information [22]. It reflects the reliability and the trustworthiness of the provider. We define a trusted source as a competent data source in a particular field that can provide accurate data.
- Reputation: in the field of IT security, different approaches exist to build trust models, among which are those based on reputation. These models consider interactions and past experiences between entities [23]. We define reputation as a measure that reflects public opinion about the data reliability of a source of information. This value evolves over time based on people's experiences with the source.
- Timeliness: data timeliness can be a fundamental criterion in some contexts as illustrated in the previous example. For Fox et al. [24], a datum is said to be current or up-to-date at time t if it is correct at time t and is out-of-date at time t if it is incorrect at t but was correct at some moment preceding t. So to be up-to-date is to be correct right now and to be out-of-date is a special case of inaccuracy; an inaccuracy caused by a change in time. Timeliness reflects the mechanisms and processes put in place by the data provider to refresh and update their data in real time.
- Consistency: for Rafique et al. [25], consistency represents the degree to which information has attributes that are free from contradiction and are coherent with other information in a specific context of use. A consistency error would be that a 5-year-old child has a "married" marital status (semantic error) or postal codes that are not within an allowed range (syntactic error). Consistency of data indicates whether the logical relationship between correlated data is correct and complete [26]. We can then use consistency as a criterion that justifies the accuracy of the data [4].

The accuracy criteria presented so far can be used as indicators to determine reference data. Obviously, the list is not exhaustive and the choice of accuracy criteria is strongly dependent on the context of application.

In many cases, a single criterion would not be sufficient and the combination of two or more accuracy criteria would be necessary. If we take the example of financial market data, it may happen that two providers diverge on a particular datum, which requires, in addition to the timeliness, to take into account others accuracy criteria such as trust and/or reputation.

In the case where several accuracy criteria are pooled together to identify the reference data, the ranking of these criteria in order of importance would be mandatory. For this, it will be necessary to assign to each criterion a weight which represents its importance during the resolution of the possible conflicts. But before that, the values should be normalized so that all the criteria are represented by the same unit of measure. For example, if for a given case, three accuracy criteria are used, the trust represented by a binary value (0 or 1), the reputation represented on a scale (from 1 to 5) and the consistency represented by a percentage (between 0 and 100), the pooling together of the three criteria requires the normalization of the values using a mathematical technique such as Min-Max Scaling defined by the following formula:

$$X_{normal} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

With:

- X: represents the current value of the criterion.
- X_{min}: represents the minimum value that the criterion can have.
- X_{max}: represents the maximum value that the criterion can have.

This formula allows us to transform values with heterogeneous units into values in a range $[0, 1]$ which then enables us to apply comparative and analytical studies to different accuracy criteria.

D. Metrics

Quality evaluation can be quantitative or qualitative [19]. Quantitative evaluation is a systematic and formal process. It relies on the existing knowledge of an organization and applies computational methods as the result of a condition, a mathematical equation, an aggregation formula, etc. to reach the values of objective metrics. The results of the quantitative evaluation are therefore objective and more concrete than in the case of a qualitative evaluation. The latter is based on subjective metrics, measuring the perceptions and experiences of the stakeholders. They are generally carried out by data administrators or users through satisfaction questionnaires, user surveys, etc. [2], [14].

We denote by "data unit" the set of data belonging to a level of granularity. This is the basic element on which the accuracy assessment operations are applied: the finer the level of granularity, the longer is the calculation time, but the more precise the evaluation of the accuracy. Different metrics exist in literature to measure the accuracy of the data, among which we find:

- Boolean measure: this type of metrics takes Boolean value to indicate if a data unit is accurate or not.

- Degree measure: this metric, used to express the degree of confidence in data, is calculated by dividing the number of correct data units by the total number of data units.
- Distance measure: this is a numeric value that captures the distance between a data unit in the system and a reference data. Generally, this metric is calculated by the distance between the objects. The smaller the distance, the more similar the objects are.

In practice, determining what constitutes a unit of data and what is an error requires a set of clearly defined criteria [14] that depend on the context of each project. As an example, it is possible for an incorrect character in a text string to be tolerable in one circumstance but not in another.

Moreover, to assess the quality of data in a Big Data context, we often use the data sampling technique, which, as will be explained later, is extremely useful in circumventing the problem of large volumes of data. Quantitative measurement of accuracy requires the establishment of a set of aggregation functions. Let S be a sample of n data units to be evaluated and A_i the value of the accuracy of the i^{th} element in S. Inspired by [2], [14], [27] and [28], the following two typical aggregation functions are the most used:

- Ratio: this function measures the ratio of the number of correct data units in the sample, divided by the cardinality of the sample.

If we consider that the accuracy of the data units is expressed using a boolean measure $\mathrm{A_i} \in \{0,1\}$ with $1 \leq \mathrm{i} \leq \mathrm{n}$, then the accuracy of S is calculated as follows:

$$\text{Accuracy } (S) = \frac{|\{\mathrm{A_i}\}, \mathrm{A_i} = 1|}{\mathrm{n}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{A_i}}{\mathrm{n}} \tag{2}$$

$|\{\mathrm{A_i}\}, \mathrm{A_i} = 1|$ denotes the cardinality of data units with correct accuracy.

If we consider that the accuracy of the data units is expressed using measurements in degree or distance $A_i \in [0, 1]$ with $1 \leq i \leq n$, then it will be necessary to consider a threshold θ from which the data is considered as correct. In this case, the accuracy of S is calculated as follows:

$$Accuracy\ (S) = \frac{|\{A_i\}, A_i \geq \theta|}{n} \tag{3}$$

$|\{A_i\}, A_i \geq \theta|$ denotes the cardinality of data units having accuracy greater than or equal to θ.

- Average: this function measures the average of the correct data units. Whatever the metric used, the accuracy of S is calculated as follows:

$$Accuracy\ (S) = \frac{\sum_{i=1}^{n} A_i}{n} \tag{4}$$

If we consider that the data units do not all have the same importance, we can assign each unit a weight w_i and calculate the accuracy of S with the following method:

$$Accuracy\ (S) = \frac{\sum_{i=1}^{n} w_i A_i}{\sum_{i=1}^{n} w_i} \tag{5}$$

IV. Comparison Techniques

In this section, we briefly present some techniques and concepts needed to understand our study, especially in performing data comparison in a Big Data environment such as Big Data Sampling, Schema Matching, Record Linkage and Similarity Measurement.

A. Big Data Sampling

The need for a quick response is sometimes more important than a precise answer, especially in the case of evaluating the accuracy of the data. Data Sampling is extremely useful for making Big Data usable for analysis [29]. To analyze large sets of data in order to assess their quality, one can be satisfied with the selection and analysis of a representative sample of all data units. For certain types of problems, sampling gives results as good as performing the same analysis using all the data [30], but for particular cases, especially in the analysis of large volumes of data, sampling seems to be the most appropriate solution [1], [20], [31].

As we presented in [32], to create a sample of a dataset, different techniques exist such as Simple Random Sampling, Stratified Sampling, Cluster Sampling, Multistage Sampling, Systematic Sampling, etc. Several techniques can be used together to create an effective sample, the main rules are that the sample must be representative of all data and all data units must have the same chance of being selected in the sample. Moreover, to know the size of the sample, it will be necessary to know in advance the size of the data to be sampled which is not easy to obtain in a Big Data project. To meet these constraints, there is an effective approach called "Reservoir Sampling" initially introduced by Vitter [33]. It's a family of randomized algorithms that randomly select a sample of k elements from a large set of n-sized data or from a data stream of size n, where n is unknown or difficult to know. All elements have the same probability to be selected in the sample. The principle is: Let S be the set or the stream of data to be sampled. We start by creating a sample of the first k elements that will be called the Reservoir R and then, by sequential access on the rest of the elements of S, we randomly replace elements in R. Algorithm 1 is a typical example.

Algorithm 1: Example of Reservoir Sampling Algorithm

```
1. Let S be the data set or data stream to be sampled;
2. Create an empty array R of maximum size k;
3. Fill the array R by the first k elements of S;
4. For each element from position k+1 to the last element in S,
repeat the following process:
4.1. Let i be the position of the current element in S;
4.2. Let j be a digit generated randomly between 0 and i;
4.3. If j < k, then replace R [j] by S [i];
```

The advantage of this algorithm is that it makes it possible to create a sample by crossing the data only once, as the sample is created, by sequential access, without having to know the size of the data to be sampled and guarantees that all elements have the same chance of being selected.

B. Schema Matching

The heterogeneity of data sources is a challenge that makes data manipulation processes such as data integration, data fusion, application interoperability, software reuse, etc. complex. Data accuracy assessment also experiences this challenge, especially when comparing the data to be evaluated with those representing the real world. The heterogeneity between these data requires matching their schemas.

Bernstein et al. [34] define a schema as a formal structure that represents an engineered artifact, such as a SQL schema, XML schema, entity-relationship diagram, ontology description, interface definition, or form definition. They also define a correspondence as a relationship between one or more elements of one schema and one or more elements of another. In relational databases, Schema Matching consists in linking the tables and columns of a database to those that represent the same concepts in another database. The authors in [34] and [35] present and detail a taxonomy of techniques and methods used to achieve the schema matching, among which we find:

- Linguistic matching based on an element’s name or description, using stemming, tokenization, string and substrings matching, and information retrieval techniques.
- Matching based on auxiliary information such as thesauri, acronyms, dictionaries, and lists of mismatches.
- Instance-based matching which considers that the elements of two schemas are similar if their instances are similar based on statistics, metadata, or trained classifiers.
- Structure-based matching that considers elements in two schemas to be similar if they appear in similarly-structured groups, have similar relationships, or have relationships to similar elements.

Sutanta et al. [36], with reference to others research work, carried out a comparative study based on 34 prototype models and diagrams corresponding to different aspects such as the input type, the methods used, the field of use as well as the existence of a graphical interface allowing users to adapt the results of the prototype. According to this study, one of the most successful schema mapping prototypes is COMA 3.0 [37], which is an evolution of COMA++ [38]. This prototype accepts different types of input data (XSD, XDR, OWL, CSV, SQL), uses different matching algorithms (Linguistic based, Structure based, Instance based), is not specific to a particular field of use and interactive via a GUI. As part of our work, we used this prototype to implement our case study.

C. Record Linkage

Record linkage consists of gathering information from two records that are assumed to be related to the same entity. This involves linking records within a single file or between two or more files to identify similar records. The challenge is to collect the records of the same individual entities by searching for exact matches [39]. Record linkage can be used when assessing data quality (to detect similarities between data), when improving data quality (including data cleaning and the deletion of duplicates processes), when merging data sets, etc. Two main types of record linkage exist: deterministic and probabilistic. Deterministic record linkage is a relatively straightforward method, which usually requires exact agreement on a match key, which may be a unique identifier (e.g. national identity number, social security number, etc.) or a collection of partial identifiers (e.g. a key consisting of full name, year of birth and the postal code of the city of birth). A record pair is considered as a link only if the match keys (unique identifier or partial identifier collection) are identical. Deterministic linkage is unfortunately not always obvious. Errors or lack of information in the records may exist. To overcome these limitations, probabilistic models have been proposed to determine the linkage in the presence of recording errors and/or without using the matching keys. Newcombe et al. [40] were the first to propose probabilistic methods, suggesting that a matching weight could be created to represent the probability that two records actually correspond given the agreement or disagreement on a set of partial identifiers.

D. Similarity Measurement

Different methods exist to measure the similarity between data. The choice of a method depends largely on the type of data that need to be compared (characters, strings, numbers, binary values, etc.) and the context of how it will be used (for example, it could be considered that two strings of characters are similar even if they have one or two different characters which cannot be the case for other situations).

To compare strings, many methods exist such as Levenshtein and Jaro-Winkler distances. The Levenshtein distance is defined as the minimum number of changes needed to convert one string to another. However, depending on the context of each project, adjustments may be necessary to adapt this method to specific needs (case sensitivity, accented and special characters, use of acronyms, etc.). Jaro-Winkler distance, for its part, measures the similarity between two strings by calculating the number of characters that they have in common. It’s a variant proposed by Winkler derived from the Jaro distance used in the field of record linkage for duplicates detection. Many other methods exist in these topics such as Cosine similarity, q-Gram, Damerau-Levenshtein, etc.

For numbers, the similarity can be calculated as the difference between values [41] taking into consideration a threshold from which one can consider that two numbers are similar. The choice of the threshold depends on the context of comparison and the order of magnitude of the numbers (a difference between two small numbers does not have the same impact as that between two very large numbers). Other types, such as dates and geographical coordinates, can follow the same principle since they are convertible into numbers by retrieving the timestamp dates and latitude/longitude geographical positions.

V. Big Data Accuracy Assessment

In this section, we will demonstrate our solution to evaluate the accuracy of data in a Big Data context. We will first present our model to understand the main steps of the evaluation process. Then, to prove our concept, we will present our case study as well as its implementation in a Big Data environment.

Finally, to go back to the objective of this research work, we will analyze the results of the study and present our findings in the form of a conclusion.

A. *Assessment Data Accuracy Process*

Our model, as shown in Fig. 1, consists of five steps:

1) Master data set: the first step is to continuously collect data from different data providers and store them in their raw state in the data lake of the Information System. Data providers can be external or internal services of the organization.

2) Golden data set: before implementing a data accuracy assessment solution, as explained in the state of the art, an organization will need to determine the accuracy criteria that will enable it to determine the quality of its source data. In this step, each data set of the Master Data Set should be assigned a value for each of the accuracy criteria. Since these values are likely to change over time, this pre-processing step will have to be recurrent.

3) Mapped data: this step corresponds to the schema matching between the data to be evaluated (Input Data) and those present in the Golden Data Set. If the desired level of granularity is finer (values or objects for example), this step will consist in linking the columns of the data to evaluate with their correspondents in the Golden Data Set. If $X\,\{x_1, \dots, x_n\}$ represents the set of columns to be evaluated and $Y\,\{y_1, \dots, y_m\}$ represents the set of columns of all datasets in the Golden Data Set, then we will have 3 scenarios:

- Simple scenario: for each column x_i corresponds one, and exactly one, column y_j.
- Conflict scenario: for each column x_i corresponds a set of columns $\{y_1, y_2, \dots, y_k\}$ of cardinality $k \in [2, p]$ where p is the number of datasets of the Golden Data Set. In case of conflict, it is the column that belongs to the dataset that best meets the required accuracy criteria that will be considered for the mapping operation.
- Incomplete scenario: there exists a set of columns $\{x_1, \dots, x_l\}$ of cardinality $l \in [1, \mathrm{n}]$ of which no element has a correspondence in Y. Note that the larger the l, the more accuracy calculation loses its reliability. If $l = n$ the calculation of the accuracy cannot be carried out because no reference data will be found.

4) Reference data: for each accuracy criterion, the organization will need to determine a threshold for a set of data to be considered sufficiently correct. This step consists firstly in eliminating all the data sets that do not meet the levels of accuracy criteria required by the organization and in resolving the various possible conflicts from the previous step. Then, and in order to get around the problem of the large volume of data hosted in the lake, a process of sampling the data may prove necessary. Finally comes the step of extracting records that need to be evaluated via a Record Linkage process. In this way, we will have dynamically constructed reference data whenever an assessment process is launched.

5) Data accuracy: the last step is to calculate the similarity between the related records. Depending on whether the granularity is about the objects or the values, it will be necessary to determine the good processes of computation of similarity. It will also be necessary to determine if all the columns are involved in the similarity calculation or only particular columns.

B. *Proof of Concept*

As part of this work, and to demonstrate the feasibility and reliability of our solution, we have put in place a proof of concept. For our case study, we are interested in evaluating the accuracy of the data concerning the railway stations in Paris and its suburbs. Our approach is to prepare a Data Lake from open data sources found on the Internet. Our data are collected from three open databases:

- The website data.iledefrance.fr: an open platform of public data concerning the Ile-de-France region. The platform is managed by the communication department of the Regional Council of Ile-de-France.
- The website data.ratp.fr: an open platform of public data concerning public transportation in the Paris region. This platform is managed by the RATP (Régie Autonome des Transports Parisiens) which is a public establishment of an industrial and commercial nature fully owned by the French government. It is a control unit that ensures the operation, maintenance and engineering of networks of part of public transport in Paris and its suburbs.
- The website data.sncf.com: an open platform for public data on railway transport in France. This platform is managed by the SNCF (Société Nationale des Chemins de Fer) which is a public establishment of industrial and commercial character fully owned by the French government. It is a board that manages the transport of passengers and goods and carries out the management, operation and maintenance of the railway network in France.

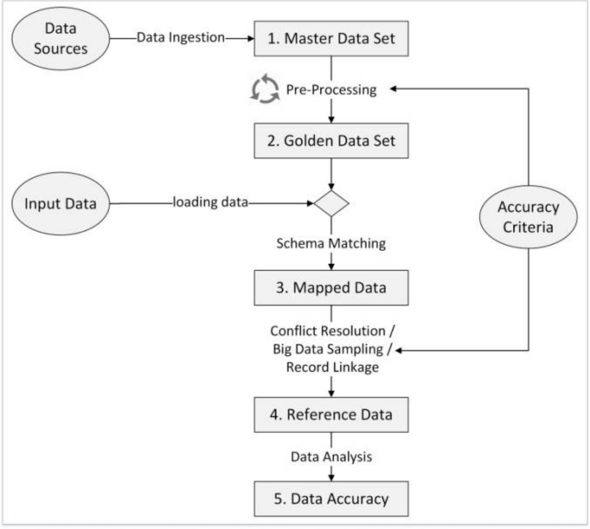


Fig. 1. Data Accuracy Assessment Process.

The very first step in implementing our solution is to define the accuracy criteria. Each organization is free to choose and define the criteria that suit it according to its activity, its projects and the nature of its data. For this case study, we will work on three criteria: Trust (T) represented by a percentage, Reputation (R) measured on a scale of 1 to 5 and Consistency (C) also represented by a percentage. The heterogeneity of the units of measurement is solved thanks to the normalization of the values by applying the Min-Max Scaling formula (1) as explained above.

C. *Solution Setting up*

We implemented our solution in six steps:

1) Big data platform: we have developed our solution on Hadoop 2.7.7 installed on Ubuntu 18.04.2 LTS 64-bit with an 8 GB RAM and 100 GB SATA disk. We have developed the different modules in python (version 3.6.7) and spark (version 2.3.1). We used HBase database (version 1.1.0) for metadata management.

2) Master data set: our Data Lake is composed of 3 Data Sets downloaded from the websites data.iledefrance.fr, data.ratp.fr and data.sncf.com. Data are stored in their raw state on HDFS (Hadoop Distributed File System).

3) Golden dataset: for each dataset, we assigned values to each of the three accuracy criteria (T, R, and C) as metadata. In a real life setting, these values must be calculated according to a well-defined approach. We will see at the end of this study that this stage is crucial. For this case study, the objective is to demonstrate the impact of accuracy criteria on the selection of reference data and, consequently, on the accuracy calculation reliability. We will then study several scenarios and in each scenario we will assign, for each of the three datasets, hypothetical values for each criterion to cover all possible cases.

4) Mapped data: this step consists in loading the data and selecting the columns to be evaluated and then matching each of them to those existing in the Golden Data Set. To achieve this step, we used the prototype COMA 3.0 [37].

5) Reference data: from the previous step, we were confronted with different situations. To perform schema matching, the evaluated column can be mapped to zero, one or many columns in the Golden Data Set. In case of mapping with multiple columns, a conflict exists and requires its resolution. For this, we have implemented a conflict resolution algorithm that consists in assigning a weight to each accuracy criterion and, in the case of a conflict, the Data Set with the highest weighted sum of the values of the accuracy criteria will be considered for the mapping. Once the mapping between the columns to evaluate and those of Data Lake is determined and all conflicts are resolved, we can select reference data through a record linkage process. We have implemented this mechanism using a Python's library called RecordLinkage [42], which provides indexing methods, similarity measurements, and classifiers.

6) Data accuracy: the record linkage result is a mapping table between records to be evaluated and reference records. All that remains now is the comparison of values. For this, the RecordLinkage library adopted for this study presents a class named *Compare* that compares the attributes of records while choosing the appropriate method for each type of data (character strings, numbers, geographic positions, etc.). By calculating the similarities of all the values of all records, we can deduce the accuracy for each column as well as for the entire table we want to evaluate.

D. *Experiments*

To justify the performance of our solution and the reliability of the results, we need to study the relationship between the accuracy criteria assigned to each dataset and the accuracy calculation result. The objective is to study the impact of the solution parameters to determine the best configuration that guarantees the most reliable result. For this, we have to execute a set of scenarios whose results are known beforehand and compare them with those calculated. To better understand our analysis approach, here is a use case:

"data.csv is a file that contains information about railway stations in Paris and its suburbs. We want to calculate the accuracy of this file by referring to data whose sources have a level of Trust ≥ 95%, with a Reputation ≥ 4.4/5 and whose Consistency ≥ 90 %."

To answer this use case, we must retrieve from our Data Lake all datasets that meet all the required accuracy criteria (i.e., after normalization, $T \geq 0.95$, $R \geq 0.85$ and $C \geq 0.90$). To get the reference data, the columns of the data.csv file must be matched with those of the selected data sets. A conflict resolution stage may be necessary. Then, the records will need to be matched through a Record Linkage process. Finally, to obtain the accuracy, it only remains to calculate the similarity between the matched records.

The reference data are extracted from datasets having at least the required values for each of the accuracy criteria. Since we have in our Data Lake three Data Sets; DS1 whose source is data.iledefrance.fr, DS2 whose source is data.ratp.fr and DS3 whose source is data.sncf.com, we can then distinguish between three groups of scenarios:

- Group A: for each scenario in this group (Scenario 1 – Scenario 3 of Table I), a single Data Set holds the maximum values for the three accuracy criteria (T, R, and C).
- Group B: for each scenario in this group (Scenario 4 – Scenario 21 of Table I), a Date Set holds the maximum values for two accuracy criteria and the maximum value of the third criterion is held by another Data Set. In this group we have 18 possible scenarios.
- Group C: for each scenario in this group (Scenario 22 – Scenario 27 of Table I), a Data Set can only have one maximum value for one of the three accuracy criteria. In this group we have 6 possible scenarios.

TABLE. I. CASE STUDY AND SCENARIOS

	Scenarios	***DS1***			***DS2***			***DS3***		
		T	***R***	***C***	***T***	***R***	***C***	***T***	***R***	***C***
Group A	Scenario 1	X	X	X	-	-	-	-	-	-
	Scenario 2	-	-	-	X	X	X	-	-	-
	Scenario 3	-	-	-	-	-	-	X	X	X
Group B	Scenario 4	X	X	-	-	-	X	-	-	-
	Scenario 5	X	X	-	-	-	-	-	-	X
	Scenario 6	-	-	X	X	X	-	-	-	-
	Scenario 7	-	-	-	X	X	-	-	-	X
	Scenario 8	-	-	X	-	-	-	X	X	-
	Scenario 9	-	-	-	-	-	X	X	X	-
	Scenario 10	X	-	X	-	X	-	-	-	-
	Scenario 11	X	-	X	-	-	-	-	X	-
	Scenario 12	-	X	-	X	-	X	-	-	-
	Scenario 13	-	-	-	X	-	X	-	X	-
	Scenario 14	-	X	-	-	-	-	X	-	X
	Scenario 15	-	-	-	-	X	-	X	-	X
	Scenario 16	-	X	X	X	-	-	-	-	-
	Scenario 17	-	X	X	-	-	-	X	-	-
	Scenario 18	X	-	-	-	X	X	-	-	-
	Scenario 19	-	-	-	-	X	X	X	-	-
	Scenario 20	X	-	-	-	-	-	-	X	X
	Scenario 21	-	-	-	X	-	-	-	X	X
Group C	Scenario 22	X	-	-	-	X	-	-	-	X
	Scenario 23	X	-	-	-	-	X	-	X	-
	Scenario 24	-	X	-	X	-	-	-	-	X
	Scenario 25	-	X	-	-	-	X	X	-	-
	Scenario 26	-	-	X	X	-	-	-	X	-
	Scenario 27	-	-	X	-	X	-	X	-	-

A special case not covered by any of the previous scenarios is the case where no Data Set satisfies all the accuracy criteria required by a use case. To be able to study the behavior for this particular case, we will assign to each criterion values close to, but less than, the maximum possible value.

For each scenario, we calculated the accuracy for all possible cases for the accuracy criteria, that is, for {T, R, C} ranging from {0.00, 0.00, 0.00} to {1.00, 1.00, 1.00}. Fig. 2, Fig. 3 and Fig. 4 show respectively the results of Scenario 1 of Group A, Scenario 4 of Group B and Scenario 22 of Group C (the first scenario of each group). The results of the other scenarios follow the same logic of those in the same group. For each scenario, we have 9261 iterations (for each variable T, R and C, from 0.00 to 1.00 with a step of 0.05, we have 21 iterations, and $21^3 = 9261$). By analyzing all the iterations of all the scenarios, we find that the accuracy can take on one of four values:

- 50%: when reference data are extracted from DS1.
- 83.14%: when reference data are extracted from DS2.
- 33.33%: when reference data are extracted from DS3.
- None: If none of the three data sets meets the criteria required for a given iteration, the accuracy value cannot be calculated for this iteration; our implementation returns then the value "None".

The analysis of the results of the different scenarios allows us to deduce that:

- For Group A, the value of the accuracy depends on the reference data set which holds the maximum of the values of the accuracy criteria. If for certain iteration no data set satisfies the criteria required, the value of the accuracy is None.
- For Group B, two values of the accuracy are possible and depend on the two data sets sharing the maximum values of the three accuracy criteria. If however no data set meets the criteria required by certain iteration, the value of the accuracy is None.
- For Group C, three values of accuracy are possible. For each iteration, the conflict resolution mechanism determines the data set that will be the source of the reference data. If no data set meets the criteria required by the current iteration, the value of the accuracy is None.

Fig. 5, Fig. 6 and Fig. 7 show the distribution of values as well as the execution time for each group. The analysis of these diagrams allows us to deduce that the smaller the number of Data Sets holding the maximum values of the accuracy criteria, the less we have None values, and the more the calculation of the accuracy is reliable. On the other hand, the execution time is longer. This is explained by the large number of data involved in the process of record linkage, reference data extraction and similarity measurement.

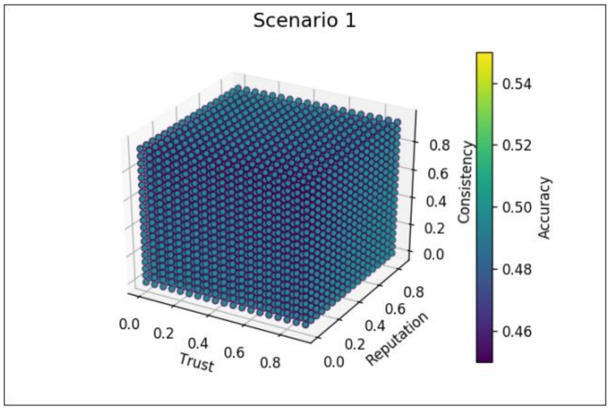


Fig. 2. Accuracy Calculation for Scenario 1 (Group A).

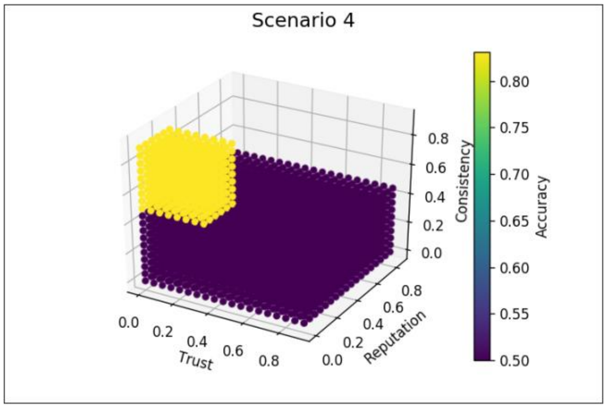


Fig. 3. Accuracy Calculation for Scenario 4 (Group B).

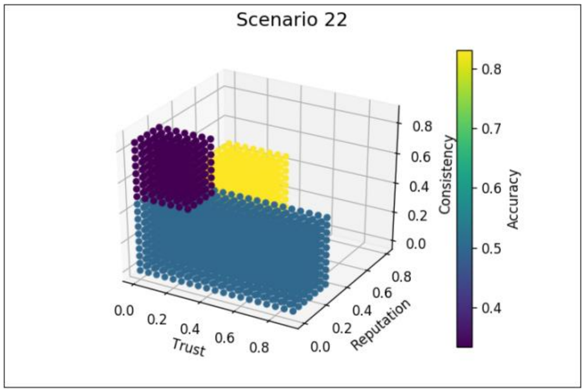


Fig. 4. Accuracy Calculation for Scenario 22 (Group C).

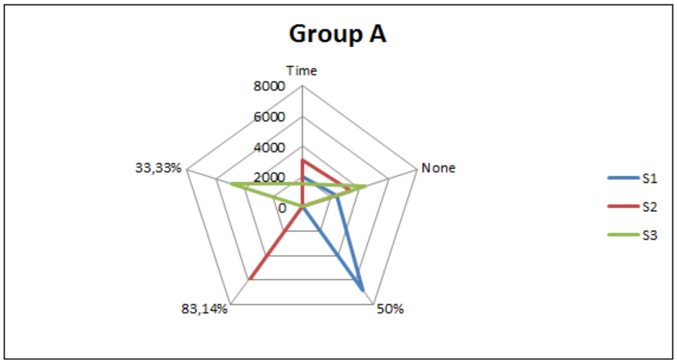


Fig. 5. Accuracy Assessment Metrics for Group A.

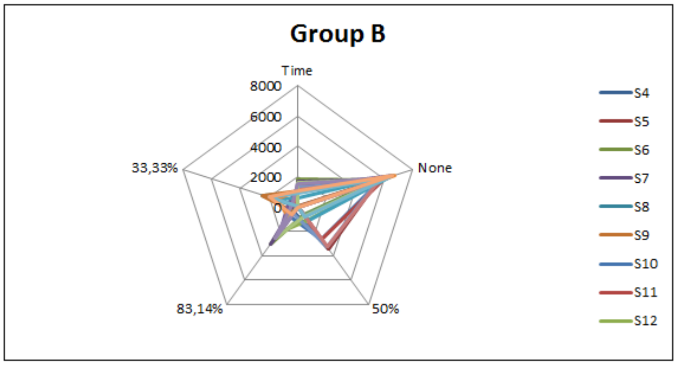


Fig. 6. Accuracy Assessment Metrics for Group B.

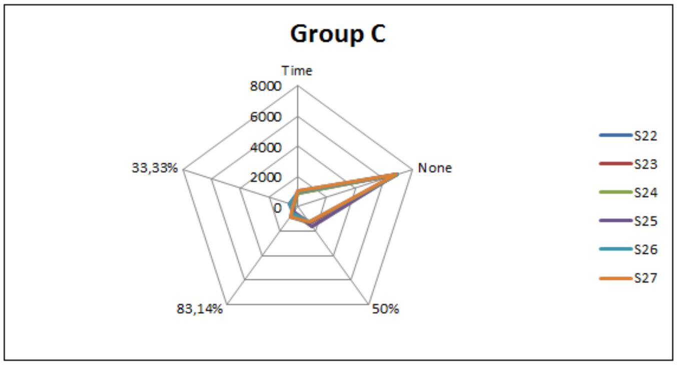


Fig. 7. Accuracy Assessment Metrics for Group C.

VI. Conclusions

In this work, we have highlighted a topical problem, namely the quality of data in Big Data through the evaluation of Data Accuracy. Whether syntactic or semantic, the evaluation of data accuracy requires their comparison with correct data called reference data. Obtaining such reference data is a very complex process and requires the establishment of a prior study to identify the quality criteria that measure the reliability of the data and their sources. We have proposed a solution allowing the configuration of the accuracy criteria in order to automate the selection of reference data in a Data Lake. Our study allows us to deduce that the implementation of a Big Data Accuracy Assessment System depends on several elements mainly related to the context of each project. The main steps to set up such a system are:

1) Having a data lake with data of good quality: The organization's Data Lake is the only source of reference data. The better the data lake, the more accurate the reference data.

2) Defining the right accuracy criteria that best characterize the notion of "data of good quality": For our case, we considered Trust, Reputation and Consistency of data sources as accuracy criteria. For other projects, other criteria may be more relevant such as Timeliness of the data. These criteria must be clearly measurable and assigned to each Data Set before initiating the assessment process.

3) Implementing the solution: For our case study, we have developed a demonstrator that exactly meets our needs in order to justify the reliability of our model. It is quite possible to develop a more generic application to define and manage the accuracy criteria used, to automate the mapping, to model the conflict resolution rules, etc.

VII. Future Work

As we have detailed in [43], confidentiality involves setting up a set of rules and restrictions to limit access to confidential data. It is generally handled with access control and cryptographic mechanisms. However, data quality assessment requires read access to the whole data. As for improving data quality, it requires write access to the data. We can therefore deduce that data security can make data quality management processes slower, more complex or even impossible. For our data quality assessment solution, we assumed that all datasets in the Data Lake are accessible, which cannot be the case in a professional setting in which data are often protected by different mechanisms and security policies even if they are hosted within the same system. Our next work will focus on this issue. We will work on implementing an effective solution to access all data without compromising their security. Our goal is to implement a data quality assessment solution in a Big Data context without compromising data security and without it being a barrier.

References

[1] A. Motro and I. Rakov, “Estimating the Quality of Databases,” Springer-Verlag Berlin Heidelberg, 1998, 298-307, 10.1007/BFb0056011.

[2] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for Data Quality Assessment and Improvement,” ACM Computing Surveys 2009, 41, 10.1145/1541880.1541883.

[3] M. Fugini, M. Mecella, P. Plebani, and M. Scannapieco, “Data Quality in Cooperative Web Information Systems,” Kluwer Academic Publishers, Netherlands (2002).

[4] T. Redman, “Data Quality for the Information Age,” Artech House 1996, isbn:0890068836.

[5] C. Batini and M. Scannapieco, “Data Quality: Concepts, Methodologies and Techniques,” Springer-Verlag Berlin Heidelberg, 2006, isbn:978-3-540-33172-8, 10.1007/3-540-33173-5.

[6] T. Redman, “Measuring data accuracy: A framework and review, Information Quality,” London and New York: Taylor & francis group 2005, 0-7656-1133-3, 21-36, isbn:9781317467991.

[7] Y. Wand and R. Wang, “Anchoring Data Quality Dimensions in Ontological Foundations,” Commun, ACM 1996, 39, 86-95, 10.1145/240455.240479.

[8] G. Mylavarapu, J.P. Thomas, and K.A. Viswanathan, “An Automated Big Data Accuracy Assessment Tool,” IEEE 4th International Conference on Big Data Analytics (ICBDA), Suzhou, China, 2019, 193-197, 10.1109/ICBDA.2019.8713218.

[9] I. Taleb, H. El Kassabi, M. Serhani, R. Dssouli, and C. Bouhaddioui, “Big Data Quality: A Quality Dimensions Evaluation,” Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress 2016, 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122.

[10] T. Haegemans, M. Snoeck, and W. Lemahieu, “Towards a Precise Definition of Data Accuracy and a Justification for its Measure,” Proceedings of the International Conference on Information Quality (ICIQ), 2016, Article 16.

[11] D. Ballou and H. Pazer, “Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems,” Management Science, February 1985, 31(2):150, doi:10.1287/mnsc.31.2.150.

[12] R. Wang and D. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” Journal of Management Information Systems 1996, 12, 5-33, 10.1080/07421222.1996.11518099.

[13] F. Naumann, U. Leser, and J.C. Freytag, “Quality-driven Integration of Heterogenous Information Systems,” Proceedings of the 25th VLDB Conference, Edinburgh, Scotland, 1999, 447-458.

[14] L. Pipino, Y. Lee, and R. Wang, “Data Quality Assessment,” Communications of the ACM 2003, 45, 10.1145/505248.506010.

[15] D. Loshin, “Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph,” San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2013, isbn:9780124186644.

[16] M. Scannapieco, P. Missier, and C. Batini, “Data Quality at a Glance,” Datenbank-Spektrum 2005, 14, 6-14.

[17] International Organization for Standarization ISO/IEC 25012, “Report: Software product quality requirements and evaluation (SQuaRE) - Data quality model,” 2006, Source : JTC 1/SC7 WG06.

[18] G. Shanks and B. Corbitt, “Understanding data quality: Social and cultural aspects,” Proceedings of the 10th Australasian Conference on Information Systems 1999.

[19] A. Immonen, P. Pääkkönen, and E. Ovaska, “Evaluating the Quality of Social Media Data in Big Data Architecture,” IEEE Access 2015, 3, 1-1, 10.1109/ACCESS.2015.2490723.

[20] T. Dasu and T. Johnson, “Exploratory Data Mining and Data Cleaning,” John Wiley & Sons 2003, Canada, isbn:0471268518, 10.1002/0471448354.

[21] P. Missier, G. Lalk, V. Verykios, F. Grillo, T. Lorusso, and P. Angeletti, “Improving Data Quality in Practice: A Case Study in the Italian Public Administration,” Distributed and Parallel Databases (2003), 13, 135-160, 10.1023/A:1021548024224.

[22] M. Gamble and C. Goble, “Quality, Trust, and Utility of Scientific Data on the Web: Towards a Joint Model,” Proceedings of the 3rd International Web Science Conference, WebSci 2011, 10.1145/2527031.2527048.

[23] H. Hussain, O. Hussain, and E. Chang, “An overview of the interpretations of trust and reputation,” IEEE International Conference on Emerging Technologies and Factory Automation, 2007, ETFA, 826-830, 10.1109/EFTA.2007.4416865.

[24] C. Fox, A. Levitin, and T. Redman, “The notion of data and its quality dimensions,” Information Processing & Management 1994, 30, 9-19, 10.1016/0306-4573(94)90020-5.

[25] I. Rafique, P. Lew, M.Q. Abbasi, and Z. Li, “Information quality evaluation framework: Extending ISO 25012 data quality model,” World academy of science, Engineering and Technology 2012, 65, 523-528.

[26] L. Cai, and Y. Zhu, “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era,” Data Science Journal 2015, 14, 10.5334/dsj-2015-002.

[27] V. Peralta, “Data Quality Evaluation in Data Integration Systems,” Human-Computer Interaction [cs.HC], 2006, Université de Versailles-Saint Quentin en Yvelines - Université de la République d’Uruguay, HAL Id: tel-00325139.

[28] D. Loshin, “Enterprise Knowledge Management: The Data Quality Approach,” San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2001, isbn:0-12-455840-2.

[29] S. Pyne, B.L.S. Prakasa Rao, and S.B. Rao, “Big data analytics: Methods and applications,” Springer India 2016, isbn:978-81-322-3626-9, 10.1007/978-81-322-3628-3.

[30] J. Dean, “Big Data, Data Mining, and Machine Learning,” John Wiley & Sons 2014, Canada, isbn:9781118618042.

[31] V. Prajapati, “Big Data Analytics with R and Hadoop,” Birmingham: Packt Publishing 2013, isbn:978-1782163282.

[32] M. Talha, N. Elmarzouqi, and A. Abou El Kalam, “Quality and Security in Big Data: Challenges as opportunities to build a powerful wrap-up solution,” Journal of Ubiquitous Systems and Pervasive Networks 2019, 12, 09-15, 10.5383/JUSPN.12.01.002.

[33] J. Vitter, “Random Sampling with a Reservoir,” ACM Transactions on Mathematical Software 1985, 11, 37-57, 10.1145/3147.3165.

[34] P. Bernstein, J. Madhavan, and E. Rahm, “Generic Schema Matching, Ten Years Later,” 2011, PVLDB, 4, 695-701.

[35] E. Rahm and P. Bernstein, “A Survey of Approaches to Automatic Schema Matching,” VLDB J. 2001, 10, 334-350, 10.1007/s007780100057.

[36] E. Sutanta, R. Wardoyo, K. Mustofa, and E. Winarko, “Survey: Models and Prototypes of Schema Matching,” International Journal of Electrical and Computer Engineering, 2016, 6:3, 1011-1022, issn:2088-8708, 10.11591/ijece.v6i3.9789.

[37] E. Rahm, “Towards Large-Scale Schema and Ontology Matching,” Springer-Verlag Berlin Heidelberg 2011, 10.1007/978-3-642-16518-4_1.

[38] D. Aumueller, H. Do, S. Massmann, and E. Rahm, “Schema and ontology matching with COMA++,” Proceedings of the ACM SIGMOD International Conference on Management of Data, 2005, 906-908, 10.1145/1066157.1066283.

[39] T.N. Herzog, F.J. Scheuren, and W.E. Winkler, “Data Quality and Record Linkage,” Springer Science+Business Media, LLC, 2007, isbn:978-0-387-69502-0, 10.1007/0-387-69505-2.

[40] H. Newcombe and J. Kennedy, “Record linkage: making maximum Use of the discriminating power of identifying information,” Commun, ACM, 1962, 5:11, 563-566, 10.1145/368996.369026.

[41] G. Shankaranarayanan, M. Ziad, and R. Wang, “Managing Data Quality in Dynamic Decision Environments,” Journal of Database Management 2005, 14, 14-32, 10.4018/jdm.2003100102.

[42] J. de Bruin, “Python Record Linkage Toolkit,” 2018, https://recordlinkage.readthedocs.io/en/latest/index.html.

[43] M. Talha, A. Abou El Kalam, and N. Elmarzouqi, “Big Data: Trade-off between Data Quality and Data Security,” Procedia Computer Science 2019, 151, 916-922, 10.1016/j.procs.2019.04.127.

Semantic Architecture for Modelling and Reasoning IoT Data Resources based on SPARK

Ahmed Salama[1], Masoud E. Shaheen[2], Haytham Al-Feel[3]
Information Systems Department
Faculty of Computers and Information
Fayoum University, Fayoum, Egypt

Abstract—**Electronic Internet-of-Things is one of the foremost valuable techniques today. Through it, everything everywhere the globe became connected and intelligent, eliminating the wants to human-to-human interaction to perform tasks. This by changing all of those objects like humans, machines, devices and something around to be simply an internet Protocol (IP) to be expressed within the network environment through completely different sensors and actuators devices which might facilitate the interaction between all of them. These different types of sensors generate a large volume of various information and data. This type of sensor information created it generally useless because of the heterogeneity and lack of interoperability of it that represents it in unstructured form. So, investing from semantic internet techniques might handle these main challenges that face the IoT applications. Hence, the main contribution behind this research aims to boost the performance and quality of sensors information retrieved from IoT resources and applications by using semantic web technologies to resolve the matter of heterogeneity and interoperability and then convert the unstructured sensor data to structured form to realize the next level of investing of sensors employed in IoT applications. Also, the aim through this research to improve the performance of the tremendous amount of information that represents the demonstrated IoT information utilizing Big Data techniques such as Spark and its query language that's named SPARK-SQL as a streaming inquiry language for a colossal amount of information. The proposed architecture demonstrated that utilizing the semantic techniques to model the streaming sensors data improve the value of information and permit us to gather unused information. Moreover, the improvement by using SPARK leads to extend the performance of utilizing this sensor information in terms of the time retrieval of running queries, particularly when running the same queries utilizing the conventional SPARQL inquiry language.**

Keywords—Big Data; Internet of Things; Semantic Modelling; Semantic_Reasonin; Semantic_Rules; Sensors; Apache SPARK; SPARK_SQL

I. Introduction

Internet-of-Things is considered one of the hottest trends that formulate the progress of information technology development sector. Connecting every object via Internet Protocol (IP) facilitates the intercommunication between human users and machines in different aspects. In this context, there were various researches that focused on the physical side of the IoT applications without representing the importance of the information that is gathered from the resources of Internet of Things devices On this context, there have been various researches that centered at the physical aspect of the IoT packages without representing the importance of the information which might be collected from the sources of internet of things devices. IoT is divided into four architectural layers which started with the specified networked things, consisting of wireless sensors and actuators as layer 1, and layer 2 represents each structure of sensors data aggregation and virtual data conversion. Additionally, layer three overviews the role of IT structures in appearing preprocessing of information earlier than it saved into the storage repository. Finally, the extracted information is analyzed, controlled, and loaded directly to the conventional lower back-give up storage systems as shown in Fig. 1.

Hence, the aim of this research is to:

- Build a semantic modeled architecture. This proposed architecture could model the different information fetched from the IoT sensors and actuators such as humidity, temperature, and pressure. This could enrich the meaning of this data and solve the main issue of heterogeneity.
- Build a reasoner tool based on the Description Logic (DL) as one of the Artificial Intelligence languages that depend on semantic web technologies to infer a set of new rules based on a set of existing concepts and individuals after modeling this fetched information.
- Providing the proposed model with the SPARK ecosystem as a big data platform based on Hadoop. This enhancement will increase the performance of the queries performed semantically against the SPARQL query language. This enhancement will illustrate the strength factor that advantages the contribution to others.

The rest of the research is organized as follows: Section 2 presents the literature review that relates to the proposed works. Also, the background technologies which are used through the work are explained in Section 3. In addition to that, the proposed architecture is discussed in Section 4. On the other hand, the implementation processes of the work is presented through Section 5. Also, the results and the comparative study are presented in Section 6. In addition, evaluating the proposed architecture is explained through Section 7. Eventually, Section 8 concludes the paper and discusses the possible directions for future work.

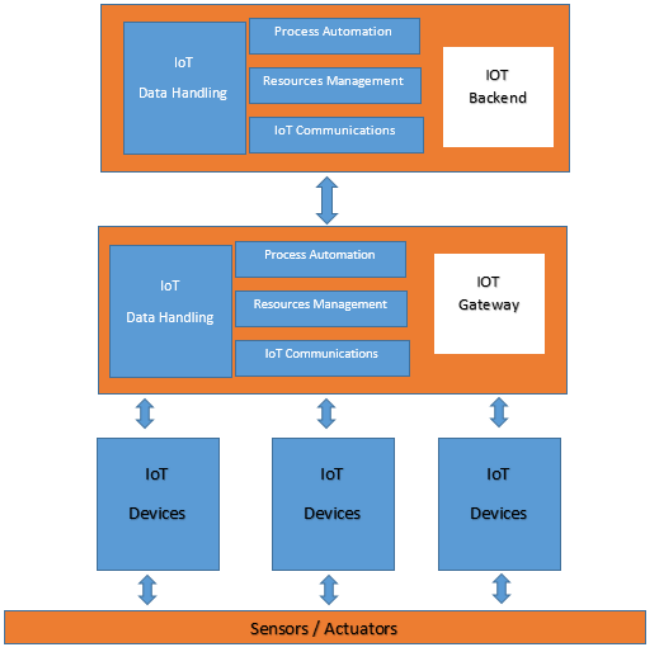


Fig. 1. IoT layered Architecture.

II. Literature Review

Through this section, we will concentrate on the onset of the most significant logical inquiries that show the significance of incorporating semantic web advancements with Web of-Things applications. The author of [1] concentrated on a methodology that supplements the depictions of Web of-Things assets with progressively nitty-gritty data separated from the demonstrated semantics of ontologies' ideas to improve their use and interoperability. Then again, there is another commitment centered on the portrayal models for sensor's information utilizing reasonable ontologies OntoSensor [2] that fabricate a cosmology based particular model for sensors by excerpting portions of SensorML [3] depictions. Also, the work presented in [4] presented a Sensor observation application via semantic web methodologies, which is called SemSOS, which enables users to perform sophisticated queries on information from the environment and data gathered from sensors. While the community of the World Wide Web Consortium (W3C) published one of the standard ontologies that integrate with IoT resources standard ontology named Semantic Sensor Network (SSN) [5] which describes sensors and their gathered data. Additionally, the SSN philosophy could deal with the heterogeneity issue of sensors when assembling their related information, however, it has a set of confinements in taking care of the transient or spatial information of sensors assets. Lamentably, the majority of the current IoT or sensor related ontologies speak to IoT gadgets just halfway for example detecting gadgets in SSN cosmology. Which is required for speaking to more extravagant data identified with IoT elements and their properties adjusts additionally with one of the difficulties of Semantic Web research as far as smart entities, detailed by Sabou M. in [6] for example the portrayal of a variety of data with respect to smart objects on the IoT. On the other hand, the IoT-A [7] considers a portion of the undertakings that stretch out the SSN ontology to simulate other IoT resources and services. But unfortunately, the IoT-A model appears to be increasingly unpredictable particularly for quick client adjustment and responsive conditions. In the same context, authors in [8] tried to examine how the IoT methods can be demonstrated utilizing web ontologies that enable them to straightforwardly convey the strategy usage. On the other hand, authors in [9] propose an IoT-Lite ontology, by launching of the semantic sensor organize (SSN) ontology to depict key IoT ideas permitting interoperability and revelation of tactile information in heterogeneous IoT stages by a lightweight semantics. In addition to that, authors In [10] established a more concluded method for collecting sensor information by using SASML (Sensors Annotation and Semantic Mapping Language) to annotate the corresponding relationship between the SSN ontology and its sensor data in the mapping file using the RDF Mapping Data Sensor (SDRM) algorithm.

In addition, the writers in [11] transmitted a major dataset of sensor metadata and measurements based on a set of measurements and observations standards that mapped semantic format such as the Resource Description Framework (RDF). Also, processing and handling the fetched sensors information that is stored into an ontology is one of the most important relevant achieves described in [12]. On the other hand, handling the huge amount of data semantically is issued and handled by using big data techniques such as Hive and Shark as discussed in [13].

III. Background Technologies

Firstly, the gathered data from sensors are collected using different network techniques either wired or wireless communications earlier than starting the processing phase of it in the proposed ontology. There are different technical and scientific steps, strategies and technologies used at some point in the processing phase of the proposed architecture together with Semantic web mapping, modeling, reasoning, and querying methods.

Semantic Web is considered one of the Knowledge Representation (KR) technologies that give plausibility to a better comprehension of encompassing situations. With a developing number of sensors and devices linked with the Internet, semantics play an ever-increasing number of basic roles as far as information fusion, interoperability, and understanding. It focuses on how to model different data types to be processed instead of presenting them only. In addition to that, the Semantic Web architecture has set basic languages such as XML, Resource Description Framework (RDF) and Web Ontology Language (OWL). The ontology can be considered as a Knowledge Organization and data modeling tool. Semantic Reasoning is considered the process of generating new inferences from a collection of given propositions condition. It is deeply related to the logical ontologies perspective provided such as OWL Description Logics (DL) [14].

The Description Logics are used as one of Knowledge Representation Languages (KR) that depends on the artificial intelligence tools and Semantic Web technologies to dedicate a piece of new information based on the given and relevant concepts of the terminological knowledge of applications. It also provides a logical shape Semantic Web is considered one of the Knowledge Representation (KR) technologies that give plausibility to a better comprehension of encompassing situations. With a developing number of sensors and devices linked with the Internet, semantics play an ever-increasing number of basic roles as far as information fusion, interoperability, and understanding. It focuses on how to model different data types to be processed instead of presenting them only.

In addition to that, the Semantic Web architecture has set basic languages such as XML, Resource Description Framework (RDF) and Web Ontology Language (OWL). The ontology can be considered as a Knowledge Organization and data modeling tool. Semantic Reasoning is considered the process of generating new inferences from a collection of given propositions condition. It is deeply related to the logical ontologies perspective provided such as OWL Description Logics (DL) [14]. The Description Logics are used as one of Knowledge Representation Languages (KR) that depends on the artificial intelligence tools and Semantic Web technologies to dedicate a piece of new information based on the given and relevant concepts of the terminological knowledge of applications. It also provides a logical shape for ontologies of Semantic Web.

In this context, semantic reasoner tools can be used to model the sensors' data within an ontology environment to infer new information which enriches the given pure and obvious data [15]. This reasoning techniques are formed into Rule-based layer which is defined on the top of the new Semantic Web architecture where different rule languages designed for handling the Semantic Web reasoning tasks such as the Semantic Web Rule Language (SWRL), REVERSE Rule Markup Language (R2ML), RuleML (Rule Markup Language), and Rule Interchange Format (RIF) [16]. On the other hand, levering from the huge amount of data collected from sensors is one of the main important challenges which could be enhanced through utilizing big data handling methods and techniques such as Hadoop and SPARK as big data ecosystems that could improve the performance of the retrieved data.

IV. Proposed Work Discussion

The aim behind this investigation is to construct a new semantic architecture to model different data retrieved from sensors systems that control the Internet-of-Things resources and applications. This new architecture aims to handle the issues of information heterogeneity that characterizes the IoT recourses, particularly when utilizing diverse sensor information for diverse utilization purposes. Moreover, we aim through the proposed demonstrate to use as it were the standard substances and ideas of the SSN ontology and after that give the fundamental ontology with the particular and required concepts and properties which are required for abstracting sensors' data of IoT applications for smart homes.

Hence, the model will enable us to use these different sensors data in different aspects without affecting the meaning or usability of each of them. Also, we are aiming through this research to improve this modeled ontology by extracting more knowledge and information from the gathered data of sensors. This could be performed by using the reasoning language such as Description Logics (DL) and Semantic Web Rule Language (SWRL) to infer new relations and rules that could enrich the value of the proposed model of IoT architecture. the new proposed architecture is organized into four main layers as follow as shown in Fig. 2:

- The First Layer represents the data collection phase in which we gather the data of different types of IoT sensors in real-time such as lighting, temperature, and air conditioner through the WI-FI protocols of IoT such as Arduino Node MCU as a preprocessing layer of the architecture.

- The Second Layer shows the data processing layer in which we model the collected data of sensors using semantic web modeling techniques such as Ontology Engineering languages such as OWL-FULL, OWL-DL, and mapping tools such as R2Q and D2R to map the sensor data to the semantic format to classify the different sensor data into relevant categories.

- The Third Layer focuses on how to enrich the meaning of the modeled sensor data via reasoning techniques such as Description Logics and SRWL language that process the given sensor data and then infer new relations between them to increase the robustness and homogeneity of data gathered from different sensors types.

- Eventually, the fourth and top layer represents the solution for handling the expected huge amount of sensor data when running the architecture on a large scale. This layer shows the mapping process of the modeled Sensor data to be stored on the Hadoop Distributed File system (HDFS) to be handled and processed by using big data techniques such as Apache SPARK and its query language that is named SPARK-SQL. This enables users to perform different sophisticated queries on the new inferred relations between different modeled sensor data in the proposed ontology which helps researchers to generate accurate statistics and reports regarding understandable, consistent and homogeneous IoT sensors' data by using SPARK-SQL query language to enhance the performance of data retrieval time.

We divide the development process of the proposed architecture into a set of phases as mention before. Through this phase, we used different programming technologies to develop it such as python, java, android, firebase as a mobile database management system and Semantic web tools for modeling and reasoning the IoT resources data. We used the IoT infrastructure techniques such as NodeMCU wireless module for configuring the IoT sensors and actuators Light Dependent Resistor (LDR) sensor for detecting the light mode, lm35 temperature sensor, GPS location sensor, and IR Transmitter sensor, all of these are developed across Arduino environment with integration of python programming language. Also, we use an ontology, RDF, Description Logics

(DL) [17], Semantic Web Rule Language (SWRL) for processing the collected sensor data [18].

- Firstly, we performed the data gathering step by fetching the used sensors data such as different values of temperatures, lighting modes, air conditioners degrees, location latitude and longitude, type of sensors and then stored them into real-time database storage such as Firebase to update them simultaneously.
- Secondly, the stored sensor data is then processed by mapping them from Relational Database into the proposed ontology as a real-time database using R2RML as a Mapping Language and RDFLib [19] as python package for asserting the mapped RDF graphs in the ontology. This can be done beside the integration with the standard SSN ontology to model the stored data into the Resource Description Framework (RDF) Format as a traditional framework for representing data in semantic format which help in solving the problem of heterogeneity and give the ability for constructing new relationships0020between different IoT-Resources to interact together efficiently producing the main objective of this research; the semantic model of IoT sensors' data as shown in Fig. 3 and Fig. 4.

After that, we used DL as an Artificial Intelligence knowledge Representation techniques to build a semantic web reasoning model as well as the SWRL Rule Language to enrich the meaning of ontology by generating a set of new inference rules based on the modeled data for both sensors resources actuators through the proposed semantic ontology and other integrated standard ontologies as shown in Fig. 5.

The work is implemented based on real data of a smart home that has different resources with different sensor's data. Where the above-mentioned Fig. 3 presented the hierarchical model of the proposed ontology modeled for the IoT domain in the case study. In addition to that, the set of reasoning rules was performed to infer new relations to enrich the meaning of the modeled ontology. Where a new semantic rule is defined for users who use IoT applications that depend on the modeled sensors data in their homes, where sensors detect the temperature of it. This rule could infer a piece of new information for the cases of low-temperature degrees, the sensors of air conditioners should detect off status, otherwise, it shows detect ON status as shown in Fig. 6.

On the other hand, we used the DL techniques that increase the meaning, integrating, and maintaining of the proposed sensors ontology and improve its knowledge representation to handle the issues raised in the traditional reasoning models as shown in Fig. 7.

It named also ALCQIO which provides the modeled ontology with only negation, conjunction, disjunction, and universal and existential restrictions which is called ALC. In addition to that it provides set of additional constructors that refer to number restrictions (Q); inverse roles (I), and nominals (O) as shown in Table I.

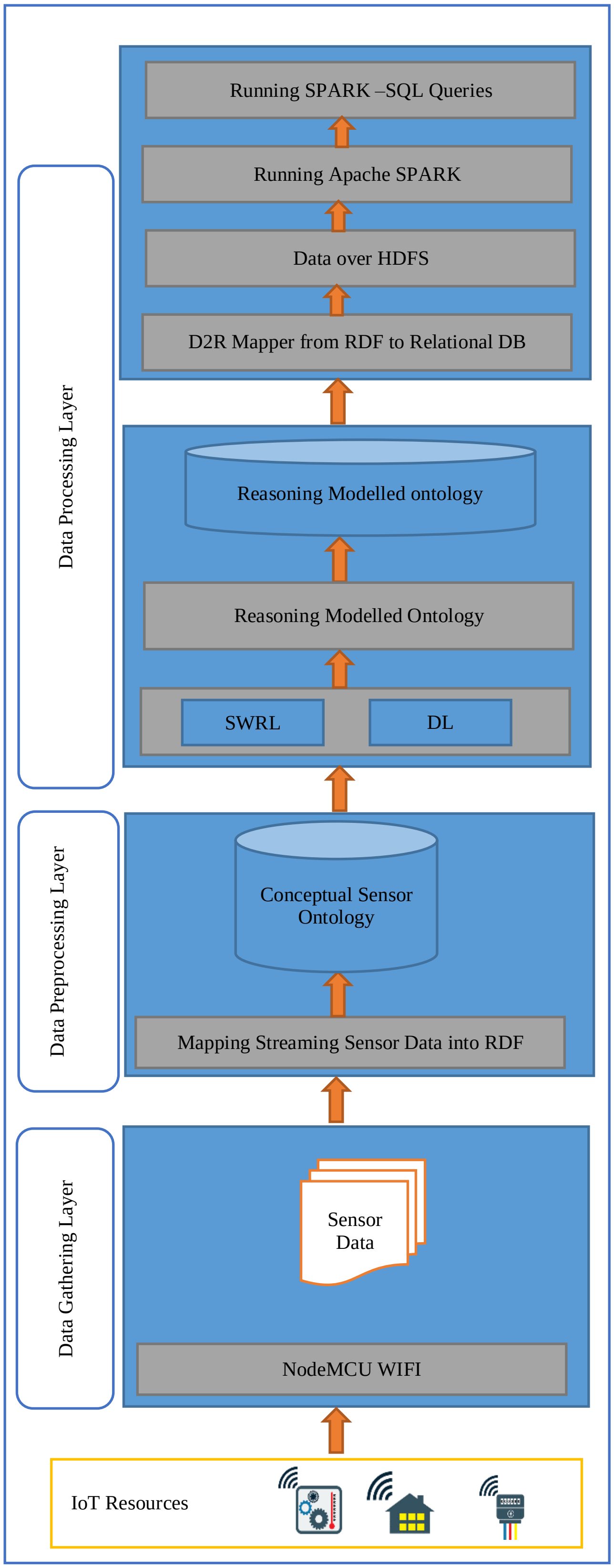
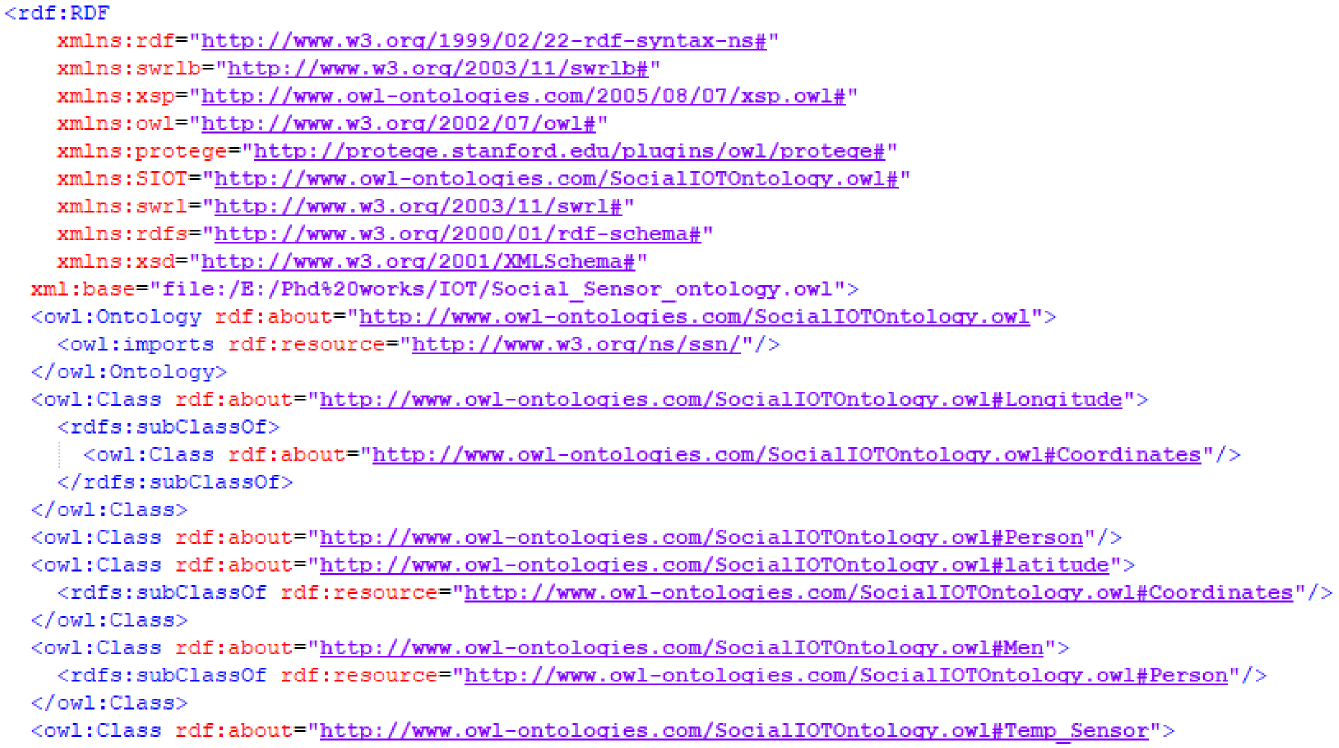

Running SPARK –SQL Queries
Running Apache SPARK
Data over HDFS
D2R Mapper from RDF to Relational DB
Data Processing Layer
Reasoning Modelled ontology
Reasoning Modelled Ontology
SWRL
DL
Data Preprocessing Layer
Conceptual Sensor Ontology
Mapping Streaming Sensor Data into RDF
Data Gathering Layer
Sensor Data
NodeMCU WIFI
IoT Resources

Fig. 2. Proposed Semantic Modelling and Reasoning Architecture.



```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:swrlb="http://www.w3.org/2003/11/swrlb#"
    xmlns:xsp="http://www.owl-ontologies.com/2005/08/07/xsp.owl#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
    xmlns:SIOT="http://www.owl-ontologies.com/SocialIOTOntology.owl#"
    xmlns:swrl="http://www.w3.org/2003/11/swrl#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xml:base="file:/E:/Phd%20works/IOT/Social_Sensor_ontology.owl">
  <owl:Ontology rdf:about="http://www.owl-ontologies.com/SocialIOTOntology.owl">
    <owl:imports rdf:resource="http://www.w3.org/ns/ssn/"/>
  </owl:Ontology>
  <owl:Class rdf:about="http://www.owl-ontologies.com/SocialIOTOntology.owl#Longitude">
    <rdfs:subClassOf>
      <owl:Class rdf:about="http://www.owl-ontologies.com/SocialIOTOntology.owl#Coordinates"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:about="http://www.owl-ontologies.com/SocialIOTOntology.owl#Person"/>
  <owl:Class rdf:about="http://www.owl-ontologies.com/SocialIOTOntology.owl#latitude">
    <rdfs:subClassOf rdf:resource="http://www.owl-ontologies.com/SocialIOTOntology.owl#Coordinates"/>
  </owl:Class>
  <owl:Class rdf:about="http://www.owl-ontologies.com/SocialIOTOntology.owl#Men">
    <rdfs:subClassOf rdf:resource="http://www.owl-ontologies.com/SocialIOTOntology.owl#Person"/>
  </owl:Class>
  <owl:Class rdf:about="http://www.owl-ontologies.com/SocialIOTOntology.owl#Temp_Sensor">
```

Fig. 3. Snap of Our Proposed Ontology in OWL DL Format.

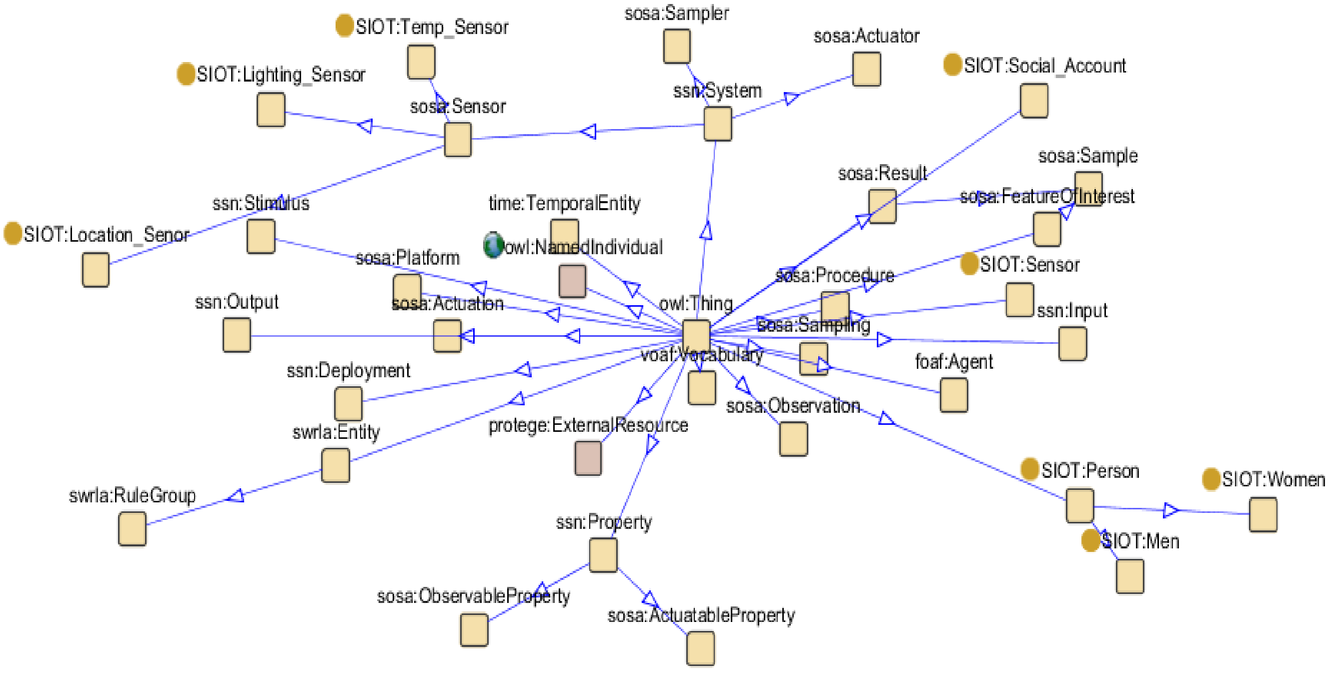
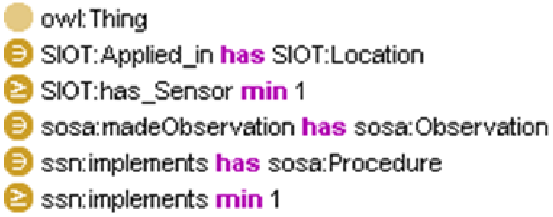
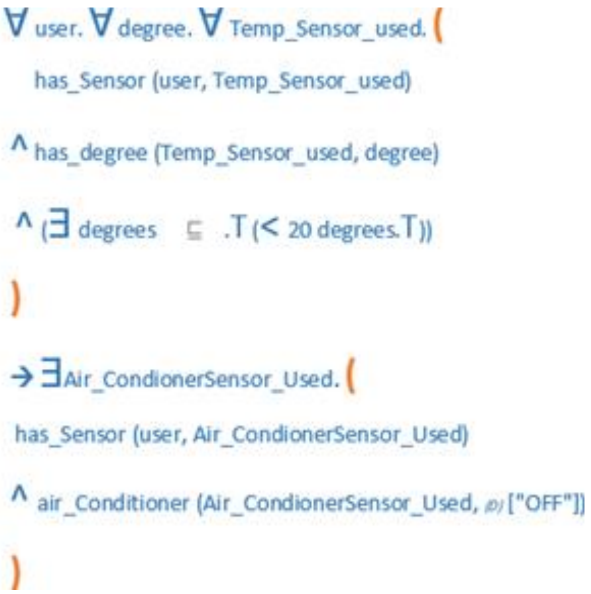


Fig. 4. Our Proposed Modelled Sensor Ontology.



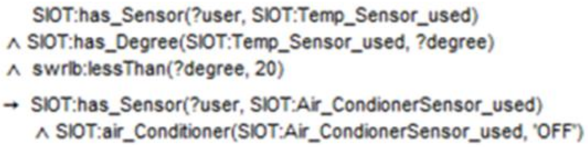
owl:Thing
SIOT:Applied_in has SIOT:Location
SIOT:has_Sensor min 1
sosa:madeObservation has sosa:Observation
ssn:implements has sosa:Procedure
ssn:implements min 1

Fig. 5. Snap Logical view from our Proposed Integrated Modelled Ontology.



∀ user. ∀ degree. ∀ Temp_Sensor_used. (
has_Sensor (user, Temp_Sensor_used)
^ has_degree (Temp_Sensor_used, degree)
^ (∃ degrees ⊑ .T (< 20 degrees.T))
)
→ ∃Air_CondionerSensor_Used. (
has_Sensor (user, Air_CondionerSensor_Used)
^ air_Conditioner (Air_CondionerSensor_Used, ["OFF"])
)

Fig. 6. SWRL Rule Reasoning Example.



SIOT:has_Sensor(?user, SIOT:Temp_Sensor_used)
∧ SIOT:has_Degree(SIOT:Temp_Sensor_used, ?degree)
∧ swrlb:lessThan(?degree, 20)
→ SIOT:has_Sensor(?user, SIOT:Air_CondionerSensor_used)
∧ SIOT:air_Conditioner(SIOT:Air_CondionerSensor_used, 'OFF')

Fig. 7. Using DL to Create New Semantic Rule Example.

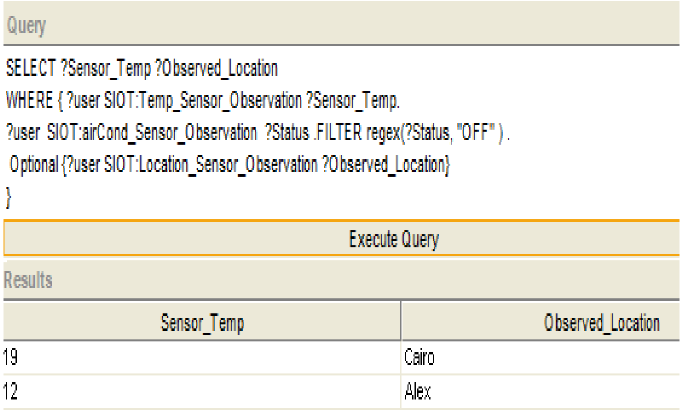
TABLE. I. THE SYNTAX OF DL (ALCQIO) REASONING

Symbol	Description
⊤	is a special concept with every individual as an instance
⊥	empty concept
Π	intersection or conjunction of concepts
¬	negation or complement of concepts
□	universal restriction
∃	existential restriction
≡	Concept equivalence

V. EXPERIMENTAL RESULTS

Our proposed model could deliver an accurate analysis for different data types collected from the Sensors and other resources of the IoT applications. Hence, we performed a SPARQL queries to generate an analysis form for the temperature degrees of locations in where sensors of air conditioners observed OFF, these insights will reflect to the rate of electricity consuming of these air conditioners and hence, provide IoT developers to keep in their mind this factor during designing and developing the IoT applications to be more efficient for their customers as shown in Fig. 8. On the other hand, we enhance the proposed architecture of the ontological model by applying the most recent techniques that handle the huge amount of data that are expected to be collected especially when applying this ontology on different aspects of IoT applications. Further, the Semantic model of IoT sensor data is formatted into RDF format which facilitates performing SPARQL queries on structure form of Subjects, Predicates, and Objects. Hence, When we propose to enhance this model with the big data techniques such as SPARK, we firstly map the RDF form of the modeled sensor' data into Relational Format which is the basic form of the SPARK-SQL query language, which is built on the top of Apache Spark. After mapping this data into Relational format, we store the data into the Hadoop Distributed File System (HDFS) to generate a dump of data over Hadoop.

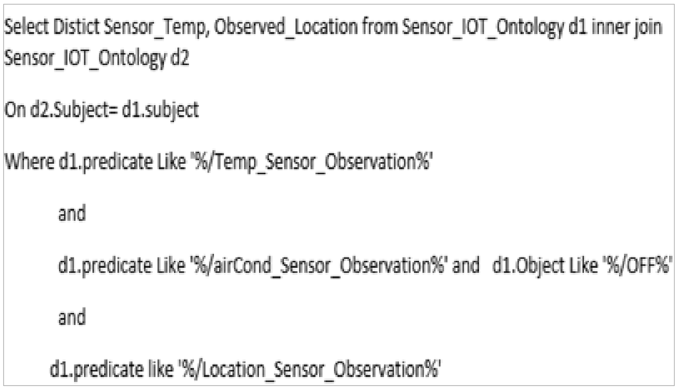
Finally, we performed set of different Semantic queries such as the mentioned query by using SPARK SQL [13] as shown in Fig. 9 query language instead of SPARQL to measure the rate of retrieve time of sensor data which refer to the better performance of running through SPARK-SQL as shown in Fig. 10 as a big data query language against the traditional SPARQL query language as shown in Table II that illustrate the retrieval time that results from performing the same query using the SPARQL and SPARK-SQL that reflect the better performance of SPARQL-SQL against SPARQL.



```
SELECT ?Sensor_Temp ?Observed_Location
WHERE { ?user SIOT:Temp_Sensor_Observation ?Sensor_Temp.
?user  SIOT:airCond_Sensor_Observation  ?Status .FILTER regex(?Status, "OFF" ) .
 Optional {?user SIOT:Location_Sensor_Observation ?Observed_Location}
}
```

Sensor_Temp	Observed_Location
19	Cairo
12	Alex

Fig. 8. SPARQL QUERY on Temp Degrees of Locations where it's Air Conditioner observed OFF in our Case Study.



```
Select Distict Sensor_Temp, Observed_Location from Sensor_IOT_Ontology d1 inner join
Sensor_IOT_Ontology d2

On d2.Subject= d1.subject

Where d1.predicate Like '%/Temp_Sensor_Observation%'

        and

        d1.predicate Like '%/airCond_Sensor_Observation%' and   d1.Object Like '%/OFF%'

        and

       d1.predicate like '%/Location_Sensor_Observation%'
```

Fig. 9. The Same Query Performed using SPARK SQL Query Language.

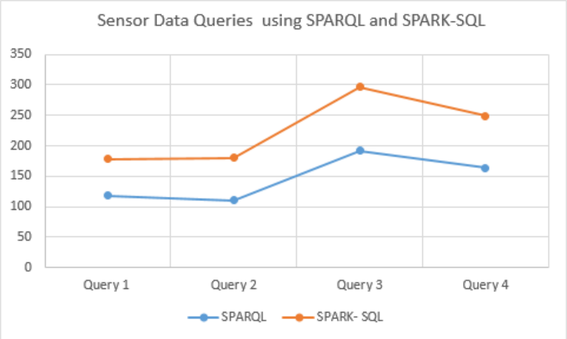


Fig. 10. The Rate of Retrieval Time of the Same Query using SPARQL and SPARK SQL.

TABLE. II. THE RATE OF RETRIEVAL TIME OF THE SAME QUERY USING SPARQL AND SPARK SQL

Query	SPARQL	SPARK- SQL
Query 1	118 seconds (Avg)	60 seconds (Avg)
Query 2	110	70
Query 3	192	105
Query 4	164	85

VI. ARCHITECTURE EVALUATION

Through this section, the proposed architecture had to be evaluated based on a set of evaluation criteria defined in [20] such as availability, accuracy, robustness, upgradeability, clearly defined functional layers, and Interoperability as shown in Table III.

TABLE. III. ARCHITECTURE EVALUATION CRITERIA

Evaluation Criteria	**Our SOIT Ontology**	**Reasons**
Availability:	Totally	Available for being operable in different mission
Accuracy	Totally	Totally accrue for modeling and retrieving the accurate values of sensors data based on the full interlinking between each concept in the ontology
Robustness	Totally	due to the full integration among the old existing ontology and the new proposed ontology to build robust Modelled IoT ontology
Upgradeability:	Totally	Totally support for concurrent upgrading for the different sensor data sources which is updated all the time
Interoperability:	Totally	Totally support for interoperability factor due to the robustness generated between different concepts in the ontology that lead to the same context while changing the reference of them
Reasoning	Totally	It totally advantages from other Modelled ontology for IoT resources because of the integration with the reasoning techniques that generate a set of new relations that we will use for building an expert system.
Performance	Faster than other	Our architecture seems to be faster than other ones because we provide the query processing with the big data techniques such as Hadoop and spark to increase the performance of the Retrieval Time (RT) of the performed queries against a huge number of sensors data.

VII. CONCLUSION AND FUTURE WORKS

Through this research, a new architecture to model the internet of things using Semantic Web techniques has been designed. We aimed through this research to tackle the problems of heterogeneity and lack of interoperability of data by modeling the data of all the devices and the resources of IoT in a new proposed ontology as well as the standard SSN ontology. In addition to that, we used the semantic Web rules techniques and Description Logics language, to build a reasoner system to enrich the meaning of the proposed modeled ontology and hence enable users to use the modeled sensors data into different aspects. After building this smarter modeled ontology, the sensor's data is enriched, the data homogeneity and interoperability is enhanced and increased. On the other hand, we enhanced the performance of the proposed architecture, especially when applying it on a huge dataset of IoT sensors that need to be handled using Big Data techniques.

Hence, through the work, we performed a set of sophisticated queries using both SPARQL and SPARK-SQL as a big data query language on the top of Apache SPARK. These experiments reflect the higher performance of query processing of SPARK-SQL against SPARQL on the sensor's data. Hence after performing this enhancement, the performance of executing different sophisticated queries is doubled using SPARK-SQL instead of using the semantic SPARQL query language. Eventually, the future works will be on how to use the proposed model of ontological IoT to build an expert system that helps users who have these IoT applications to control their applications through their social networks accounts automatically based on the immediate behavior on social media. This will help users to get better and more flexible access to the Internet of Things recourses.

REFERENCES

[1] S. Zander, N. Merkle, and M. Frank, "Enhancing the Utilization of IoT Devices Using Ontological Semantics and Reasoning," Procedia Computer Science, vol. 98, pp. 87-90, 2016.

[2] D. J. Russomanno, C. Kothari and a. O. Thomas, "Sensor ontologies: from shallow to deep models," in Proceedings of the Thirty-Seventh Southeastern Symposium on System Theory, Los Alamitos, CA, USA, 2005.

[3] O. Specification, "OpenGIS Sensor Model Language (SensorML) Implementation Specification," 2007.

[4] C. A. Henson, J. K. Pschorr, A. P. Sheth and K. Thirunarayan, "SemSOS: Semantic Sensor Observation Service," in International Symposium on Collaborative Technologies and Systems, Baltimore, MD, 2009.

[5] M. Compton, P. Barnaghi, L. Bermudez, R. GarcíA-Castro, O. Corcho, S. Cox, J. Graybeal, and M. Hauswirth, "The SSN ontology of the W3C semantic sensor network incubator group.," Web Semant.Sci. Ser. Agents World Wide Web, vol. 17, p. 25–32, 2012.

[6] K. Kotis and A. Katasonov, "Semantic interoperability on the internet of things: The semantic smart gateway framework.," International Journal of Distributed Systems and Technologies (IJDST), vol. 4, no. 3, pp. pp.47-69, 2013.

[7] W. Wang, S. De, R. Toenjes, E. Reetz, and Klaus, "In Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference," 2012.

[8] V. Caballero, S. Valbuena, D. Vernet, and A. Zaballos, "Ontology-Defined Middleware for the Internet of Things Architectures.," Sensors, vol. 19, no. 5, p. 1163, 2019.

[9] M. e. a. Bermudez-Edo, "IoT-Lite: a lightweight semantic model for the Internet of Things."," in Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data C, 2016.

[10] "A Method for Mapping Sensor Data to SSN Ontology," International Journal of u- and e-Service, Science and Technology, vol. 8, no. 6, pp. 303-316, 2015.

[11] M.Alexandra and M.Dunja, "A Framework for Semantic Enrichment of Sensor Data," Journal of Computing and Information Technology, vol. 3, p. 167–173, 2012.

[12] D. F. Barbieri, D. Braga, S. Ceri, E. D. Valle, and M. Grossniklaus, "Querying RDF Streams with C-SPARQL," SIGMOD Record, vol. 39, no. 1, p. 20–26, 2010.

[13] A. Salama, H. Alfeel and H. Mokhtar, “Bridging the gap for retrieving DBpedia data.",” In 2015 Fourth International Conference on e-Technologies and Networks for Development (ICeND), Lodz, Poland, 2015.

[14] L. Li, C. Zhou, J. He, J. Wang, X. Li and X. Wu, "Collective semantic behavior extraction in social networks," Journal of Computational Science- Elsevier, vol. 28, pp. 236-244, 2018.

[15] A. Basu, "Semantic Web, Ontology, and Linked Data," in Web Services: Concepts, Methodologies, Tools, and Applications, India, 2019, p. 22.

[16] S. Mehla and S. Jain, "Rule Languages for the Semantic Web," in Emerging Technologies in Data Mining and Information Security, 2018.

[17] L. Botha, T. Meyer, and R. Peñaloza, "A Bayesian Extension of the Description Logic $$\mathcal {ALC} $$."," In European Conference on Logics in Artificial Intelligence, Cham, 2019.

[18] S. Mehla and S. Jain, "Rule languages for the semantic web," in Emerging Technologies in Data Mining and Information Security, Singapore, 2019.

[19] wikipedia, "RDFLib," [Online]. Available: https://en.wikipedia.org/wiki/RDFLib. [Accessed 10 Nov 2019].

[20] N. Harrison and A. Paris, "Using Pattern-Based Architecture Reviews to Detect Quality Attribute Issues-An Exploratory Study," in Transactions on Pattern Languages of Programming III, Berlin, Heidelberg., 2013.

Performance Evaluation LoRa-GPRS Integrated Electricity usage Monitoring System for Decentralized Mini-Grids

Shaban Omary[1], Anael Sam[2]
Department of Communications Science and Engineering
The Nelson Mandela African Institution of Science and Technology (NM-AIST), Arusha, Tanzania

***Abstract*—The emerging Internet of Things (IoT) technologies such as Long-Range (LoRa), combined with traditional cellular communications technologies such as General Packet Radio Service (GPRS) offers decentralized mini-grid companies the opportunity to have cost-effective monitoring systems for the mini-grid resources. Nevertheless, most of the existing decentralized mini-grid companies still rely on traditional cellular networks to fully monitor electricity consumption information, which is not a feasible solution especially for the resource-constrained mini-grid systems. This paper presents the performance evaluation of the proposed LoRa-GPRS integrated power consumption monitoring system for decentralized mini-grid centers. Each mini-grid center consists of a network of custom designed smart meters equipped with LoRa modules for local data collection, while the GPRS gateway is used to transmit collected data from the local monitoring centre to the cloud server. Performance testing was conducted by using five electrical appliances whose power consumption data form the cloud server was compared to the same data collected by using a reference digital meter. The correlation between the two data sets was used as a key performance metric of the proposed system. The performance results show that the proposed system has a good accuracy hence providing a cost-effective framework for monitoring and managing of power resources in decentralized mini-grid centers.**

Keywords*—Internet of Things; LoRa; GPRS; decentralized mini-grid systems; electricity usage monitoring

I. INTRODUCTION

For many decades, conventional centralized grid extension has been a prevalent mode of electrification in several countries [1]. Nevertheless, in many developing countries, particularly Sub-Saharan Africa (SSA), decentralized mini-grid systems have emerged as alternative to electrification efforts since national grid extensions cannot be extended to all community areas due to high budget requirements [2]. This is also influenced by the nature of community settlements in many parts of the developing world, as the majority of people live in areas far from the national grid network. As a result, numerous countries have formulated Small Power Producers (SPPs) frameworks to encourage private companies to invest in decentralized off-grid systems [3].

One of the challenges facing the existing decentralized mini-grid systems is the lack of cost-effective infrastructure to monitor and manage electricity consumption from remote mini-grid centers [4]. While most off-grid companies manage and run several grid centers across the country from different regions, most of these centers are located in remote areas where access to broadband access technologies is still limited [5]. Most grid utilities still rely on traditional cellular networks to completely control and monitor energy utilization in decentralized grid centers. Nonetheless, each mini-grid center usually has a large number of consumer meters, making it not a cost-effective solution to rely entirely on cellular networks for electricity usage monitoring [6]. The mini-grid company might have to negotiate data plan agreements with cellular network providers to accommodate all the enormous data traffic, depending on the number of customers each mini-grid center serves [7]. This strategy is obviously inefficient since mini-grid companies do not have full control of the communication infrastructure across all the mini-grid centers [8].

Recent developments in the Internet of Things (IoT) provide an incentive for the implementation of heterogeneous networks to more effectively manage remote monitoring systems such as smart cities, weather data and home automation compared to the use of traditional cellular networks [9,10]. Low Power Wide Area Networking (LPWAN) is one of the latest IoT technologies providing long-range communication while using low power to transmit data, which is one of the most important criteria for remote monitoring systems. The most popular LPWAN technologies include Long Range (LoRa), Narrow Band IoT (NB-IoT) and Sigfox [11]. Among these technologies, LoRa is the most adopted because it promises ubiquitous connectivity in outdoor IoT applications, while keeping network management simple [12]. Having higher noise immunity and ability to transmit data over long distance while using license-free spectrum makes LoRa a cost-effective solution for private companies to create a wide area private network for embedded systems [13]. LoRa also has an adaptive data rate capability in the Physical Layer (PHY) which offers competitive advantages for private network deployment in urban areas compared to Sigfox and NB-IoT technologies [14]. By integrating with traditional cellular communication technologies such as General Packet Radio Service (GPRS), LoRa implementation offers a viable solution for the deployment of a cost-effective cloud-based monitoring system for remote, decentralized mini-grid centers especially in developing countries [15].

This paper presents a performance evaluation of the LoRa-GPRS integrated electricity usage monitoring system for decentralized mini-grid systems. The proposed system will help mini-grid companies to effectively and economically monitor and manage utilization of the available mini-grid resources located in different locations across the country.

The main contributions of this paper are:

- This study proposes a cost-effective electricity power consumption monitoring system for decentralized mini-grid systems by integrating LoRa and GPRS technologies. While most of the existing grid companies use Power Line Communication (PLC) as the grid monitoring technology, power stability issues makes this solution to be less practical particularly in developing countries. In this study, the combination of LoRa and GPRS technologies is used to monitor power consumption in grid centers. This approach minimizes operating costs since LoRa uses unlicensed spectrum and low power to transmit data over large distance. A single GPRS gateway is only used to send power consumption data to the cloud server.

- While other studies[16]-[20] used different metrics to assess LoRa technology efficiency in IoT applications, the majority evaluated multiple dependent variables at a time that was time-consuming and demanded thorough analysis. In this study, a simple but effective approach is used to assess the performance of the proposed system using power consumption data collected on the cloud server over a specific period of time as the performance metric. The data collected was compared to a data set collected locally by means of a standard reference meter. By using this approach the final collected data takes into account all other communication parameters involved in the system infrastructure.

The reset of the paper is organized as follows. Section II reviews some of the related works in LoRa technology and its performance evaluation in various IoT based monitoring applications. Section III describes materials and methods used in the proposed system. Section IV discusses the approach used for assessing the evaluation of the proposed system. Section V discusses the results and the main findings while conclusion for the study is given in Section VI.

II. Related Works

This section presents review of the recent works related to LoRa implementation in IoT applications, its performance evaluation and their implications. The researchers' methodologies and their main findings are explicitly discussed in order to highlight potential contributions that have been used as benchmark for the implementation of the system proposed in this study.

In [16], authors presented the analysis of LoRa technology performance experimental results by considering various performance metrics such as Transmission Power (TP), Spreading Factor (SF), coverage and data throughput. Using the European Industrial Scientific and Medical (ISM) band of 868 MHz, the SF of 12 and the default TP of 14 dBm (25 mW), the study assessed the robustness of LoRa technology with different configurations such as the use of 125 kHz and 250 kHz Channel Bandwidths (CB) and various coverage distances from 2 km to 30 km. The study found that LoRa technology has a good performance for applications that do not need heavy data usage, and is therefore an impressive new technology that can be used in smart metering systems like power, gas and water utilities.

Another study proposed a LoRa-3G hybrid communication architecture for deploying a monitoring system to optimize maintenance operations in electricity and gas distribution systems [17]. The authors carried out a performance test of the effect of different terrain variables on the LoRa communication range. This was accomplished by installing a LoRa transmission node at various locations and distances from the gateway. The test was conducted by varying SF from 7 to 12, using the European ISM 868 MHz band and 125 kHz CB while using a 2.7 dBi antenna gain for both gateways and nodes. The study found that the Line of Sight (LoS) is one of the most key factors for LoRa network performance. Therefore, when deploying LoRa network, it was proposed that transmission nodes should be located at high and isolated points to ensure LoS and avoid unnecessary blockages.

In-depth analysis of LoRa technology and its functional components has been presented [18]. Field tests and simulations were used to assess the physical and data link layers of LoRa network. The test was carried out via a testbed by positioning LoRa end devices at various outdoor locations, while the Cisco910 industrial router was placed indoor as a gateway. Using the default transmission power of 14 dBm, the test results show that high SF provides better coverage but at a cost of lower data rate. Nonetheless, in actual implementation of LoRa network protocols, SF is usually automatically adjusted to maintain the signal transmission quality. Researchers agreed that hardware resource constraints like power and memory requirements restrict the applicability of conventional cellular networks in the implementation of IoT. Therefore, the researchers concluded that the advent of LoRa technology promises communication requirements suitable for IoT applications, particularly smart metering systems, due to the potential to allow low power consumption and adaptive transmission levels relative to transmission ranges.

In another related study [19], authors provided a review of LoRa technology, its main technological interpretations and its comparison to other traditional approaches in IoT wireless technologies operating in the ISM spectrum. A proof-of-concept performance testing was carried out by installing a private LoRa network with in order to monitor temperature and humidity across the multi-store building. The researchers found that implementation of the LoRa network offers good performance in IoT applications compared with other wired and wireless technologies. The researchers also found that while LPWAN inherited the basic features of the legacy cellular networks such as user mobility and resource management, the use of a lighter control plane makes LPWAN a preferred candidate technology for low data-rate services such as in smart grid applications.

In [20], proposed a hybrid LoRa and Power Line Communication (LoRa-PLC) IoT sensor network for managing and improving energy metering at the campus using LoRa as an outdoor solution and Power Line Communication (PLC) as an indoor solution. In this study, several SX1272 LoRa modules were used as end-devices distributed around office rooms and inside electricity meters. The S7G2 starter kit was used as a gateway to relay messages from end-device sensors to the network server for data processing and storage. The LoRa network was configured to use a CB of 125 kHz and a fixed SF of 7. The test results showed that the gradual increase in obstacles and distance has a major impact on the reliability of LoRa network. The study, therefore, recommended that while LoRa technology could be used to provide indoor IoT applications, it promises more efficiency when used as indoor solution.

Performance assessment of the LPWAN in smart grid systems based on LoRa technology was presented [21]. The assessment was carried out using a LoRaSim system network simulator which was configured with varying LoRa PHY parameters, such as Transmission Power (TP), Carrier Frequency (CF), SF and CB. The simulator used Data Extraction Rate (DER) and Network Energy Consumption (NEC) as the performance indicators for the LoRa deployment. Using the performance results, the authors postulated that the implementation of LoRa technology in distributed smart grids is not a feasible option since a large number of LoRa gateways are needed to manage tremendous traffic from all end nodes. The main concern of the authors was whether the LoRa network can be robust enough to handle all monitoring information from Smart Distributed Grids (SDG) applications including video surveillance which needs high data throughput. However, the authors concluded that the performance results can be considered fairly satisfactory for managing small traffic smart grid applications such as monitoring consumer power consumption data.

The results and key findings from the above related works were used as a benchmark in this study to assess the efficiency of the proposed electricity usage monitoring system for decentralized mini-grids. Compared to previous studies, this study has put much emphasis on low-cost system design and stakeholder-driven system requirements by using available infrastructure and hardware to develop the system.

III. Proposed System

In this section, the main features of the proposed integrated LoRa-GPRS electricity usage monitoring system is described in detail. Finally, the methodology used to assess the performance of the system is explained.

A. Proposed System Architecture

The conceptual framework architecture of the proposed system is shown in Fig. 1. The system incorporates the application of IoT to monitor power consumption information of each smart meter and transmits data to the cloud server.

Consumer power consumption data is collected through a private LoRa network installed between smart meters and a local monitoring centre, which is then transmitted to the cloud server via a GPRS gateway. All decision making logic such as remote connection or disconnection of smart meters, bill processing and user authentication are handled on the cloud server.

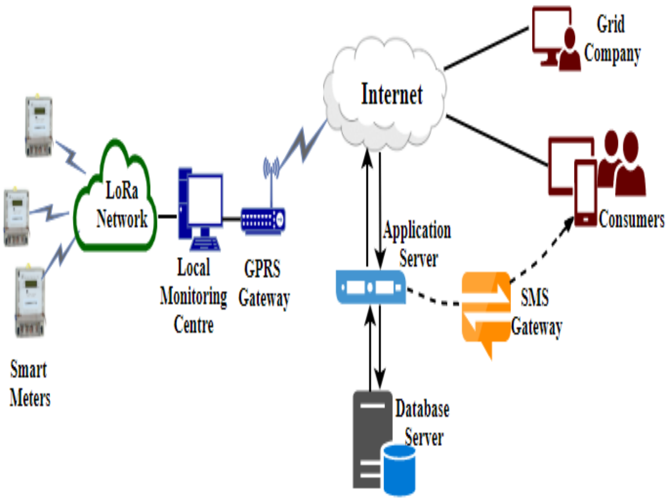


Fig. 1. Proposed System Architecture.

In addition, the Short Message Service (SMS) gateway was integrated into the system to provide consumers with usage notification feature when prepaid electricity units are about to get exhausted.

B. Smart Meter unit Configuration

The embedded part of the proposed system consists of the smart meter unit. The smart meter unit was designed to have three main features: the ability to accurately measure and record power consumption, the ability to be managed remotely by the utility company and a reliable communication link to the local monitoring centre. Fig. 2 shows the block diagram of the proposed smart meter unit indicating the main components.

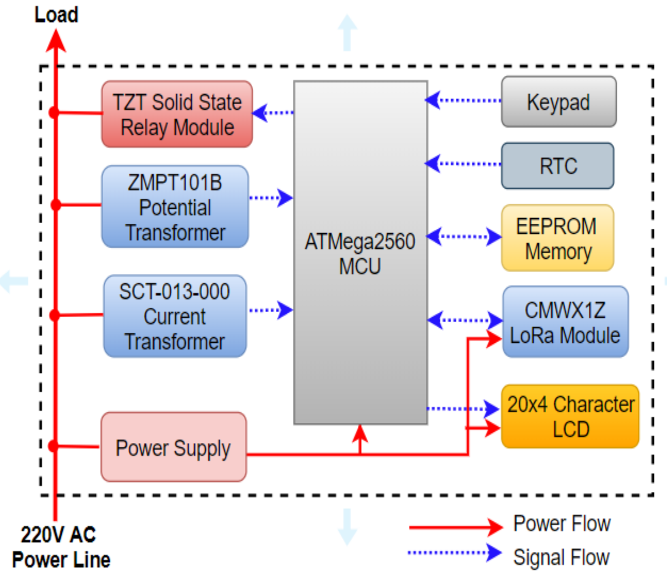


Fig. 2. Block Diagram of the Proposed Smart Meter unit.

Accurate power consumption measurement is achieved by using efficient but less expensive sensor modules available in the local market. The current sensor is made up of a non-invasive split-core SCT-013-000 current transformer which can withstand currents of up to 100 Amperes. This module has been selected because of its broad current range measurement and flexibility in cabling, as it does not require direct contact with the power cable [22]. Voltage sensor consists of a ZMPT101B single-phase voltage transformer which can withstand up to 250 AC voltage. This module was selected due to its high accuracy, and comes with a built-in trim potentiometer for adjusting the Analogue to Digital Converter (ADC) output.

The smart meter unit is remotely controlled using a TZT solid state relay module that connects or disconnects the consumer load from the main power supply. Additionally, in embedded systems and IoT applications, the Real Time Clock (RTC) is an integrated circuit (IC) which is used to keep the precise time and date of sensor measurements and microprocessor operations to synchronize with the absolute reference time [23]. RTC can be part of the microprocessor IC or an external module connected to the microprocessor via a serial interface. In this study, an external DS1307 RTC module was selected because it has a built-in backup battery that allows time tracking even when there is no primary power supply [24]. For the part of microcontroller unit, after thorough analysis, the Arduino-based ATmega2560 board was selected because of its low cost, ample processing capacity, it has a large number of integrated ADC inputs and is easily available on the local market.

C. LoRa-GPRS Network

Each smart meter unit was incorporated with a CMWX1ZZABZ LoRa module (shown in Fig. 3) which operates at the license-free band of 868 MHz and CB of 125 kHz [25]. The transmission power at this frequency band was limited to 14 dBm in accordance with the LoRa PHY specifications, which makes the signal power prone to attenuation from building blockages and other structures. In order to increase a robustness of the transmitted signal against interference, a high SF factor of 12 was used.

Additionally, LoRa gateway base stations have been installed at the top of a few poles as shown in Fig. 4 in order to boost the signal strength around the mini-grid centre. The modules used in this study belong to the class C LoRa devices, which have the ability to continuously send data to the gateway and listen to downlink messages all the time because they are powered by mains supply from the grid centre, therefore power consumption is not a big issue [26]. Fig. 5 shows a local monitoring center connected to a GPRS gateway to transmit power consumption data to the cloud server via the existing cellular communication network.

Fig. 6 shows the flowchart diagram illustrating the algorithm of the data acquisition unit between LoRa and GPRS modules on how power consumption data from each smart meter unit is collected and sent to the cloud server.

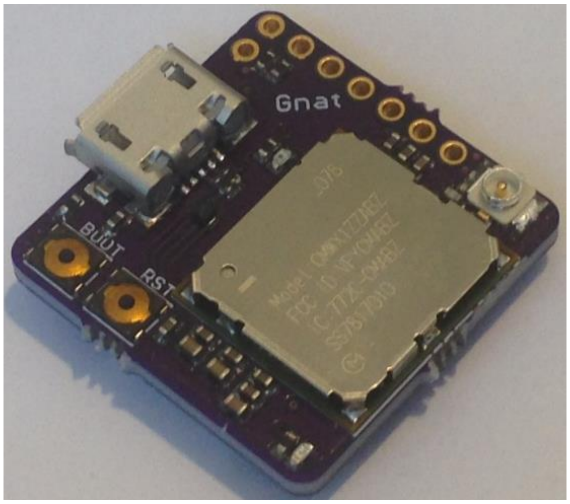


Fig. 3. CMWX1ZZABZ LoRa Module used for Data Transmission between Smart Meter units and Local Monitoring Centre.

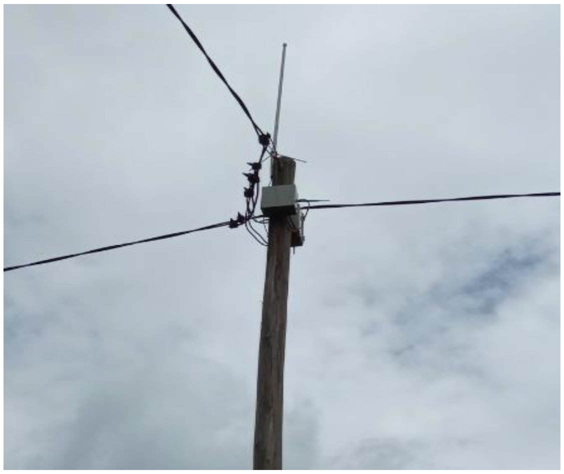


Fig. 4. LoRa Base Station Installed at the Pole.

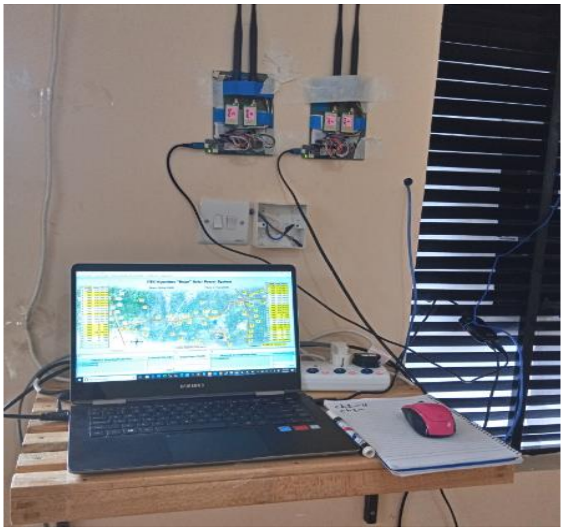


Fig. 5. Local Monitoring Centre and GPRS Gateway.

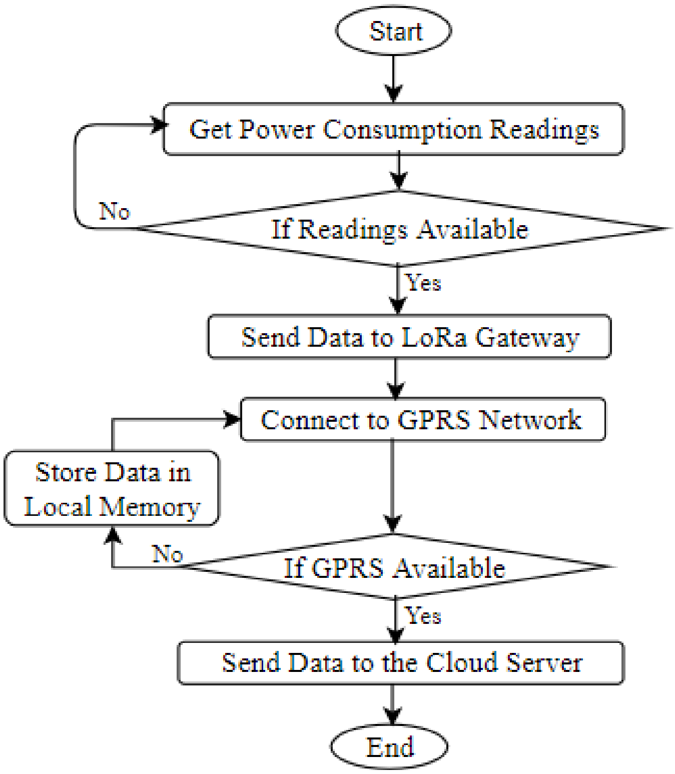

Start
Get Power Consumption Readings
No
If Readings Available
Yes
Send Data to LoRa Gateway
Connect to GPRS Network
Store Data in Local Memory
No
If GPRS Available
Yes
Send Data to the Cloud Server
End

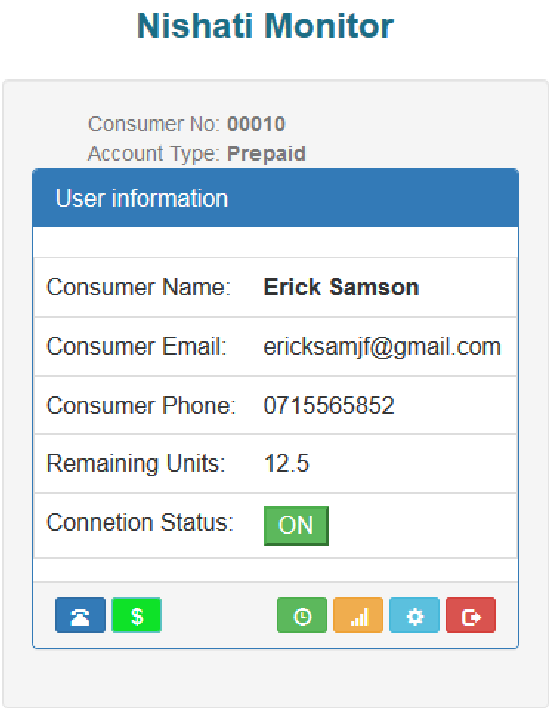
Fig. 6. Flowchart Algorithm of the Data Transmission unit.

D. Cloud Server Configuration

In order to minimize the burden of data processing overload between smart meters, the local monitoring center and the cloud server, all the basic computational tasks related to power consumption are carried out locally at the smart meter unit. The cloud server is only responsible for handling data storage, data presentation, usage monitoring and usage notification.

The cloud server was developed by using PHP Laravel and JavaScript frameworks while the database was developed by using MySQL database management system. Fig. 7 illustrates the consumer profile on the cloud server which displays basic information such as consumer contact information, status of the remaining electricity credits and the power supply connection status between smart meter unit and the mini-grid center. Links to other cloud server modules such as customer support, payment system, historical power usage and account settings are located at the bottom of the consumer profile interface.

Fig. 8 shows the historical power consumption information for the previous seven days in a graphical format for the same consumer account. The interface also shows the associated cost of the selected time frame. The cost is expressed in Tanzanian Shillings (TZS.), which is the official local currency of Tanzania where the study was conducted.



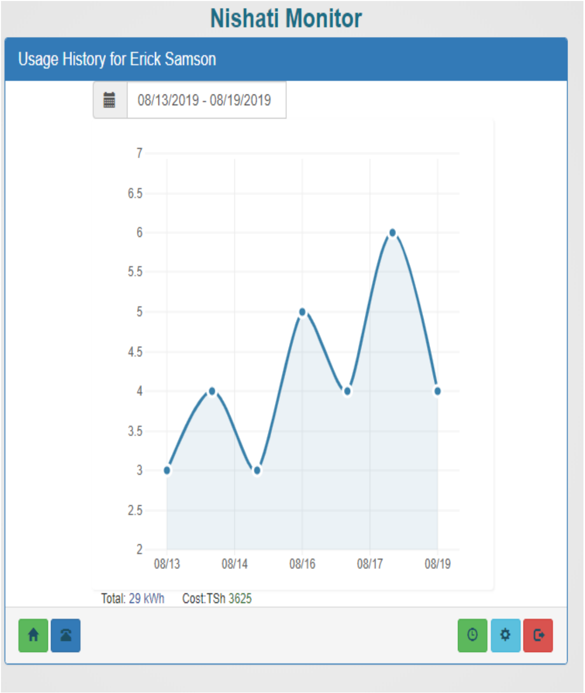
Nishati Monitor

Consumer No: **00010**
Account Type: **Prepaid**

User information

Consumer Name:	**Erick Samson**
Consumer Email:	ericksamjf@gmail.com
Consumer Phone:	0715565852
Remaining Units:	12.5
Connetion Status:	ON

Fig. 7. Cloud Server Consumer Profile user Interface.



Nishati Monitor

Usage History for Erick Samson

08/13/2019 - 08/19/2019

Total: 29 kWh Cost:TSh 3625

Fig. 8. Cloud Server Historical Power Consumption.

IV. Performance Evaluation of the Proposed System

In this section, the methodology used to evaluate the performance of the integrated LoRa-GPRS electricity usage monitoring system is described. The aim of this evaluation is to assess the accuracy of the communication infrastructure and whether the proposed system can record and transmit power consumption data with minimum packet loss. To achieve this, a smart meter prototype shown in Fig. 9 was used to measure and record power consumption of various electric appliances for a period of 20 days. Consumption data was then sent to the cloud server via a GPRS gateway for storage and historical access.

In order to validate the accuracy of the data collected on the cloud server, a DDSD101-KT1 digital meter shown in Fig. 10 was used as a reference for comparing the results [27]. This digital meter is a simple and portable power consumption meter that records power consumption in kilowatt-hour units within a 3% accuracy.

The performance evaluation was conduced using five common home appliances: light bulbs, laptop computer, electric kettle, a small fridge and a radio. These appliances were selected so that the system performance can be assessed on both types of load, linear and non-linear. Incandescent light bulbs, electric kettle and the radio are linear while a laptop computer and fridge are non-linear loads. The power ratings information of these appliances is shown in Table I. Each appliance was connected to the load side of the smart meter prototype and left to run over a certain period of time. The data collected on the cloud server was recorded and the results were compared to the same data set collected using the DDSD101-KT reference digital meter over the same time period with the same appliances.

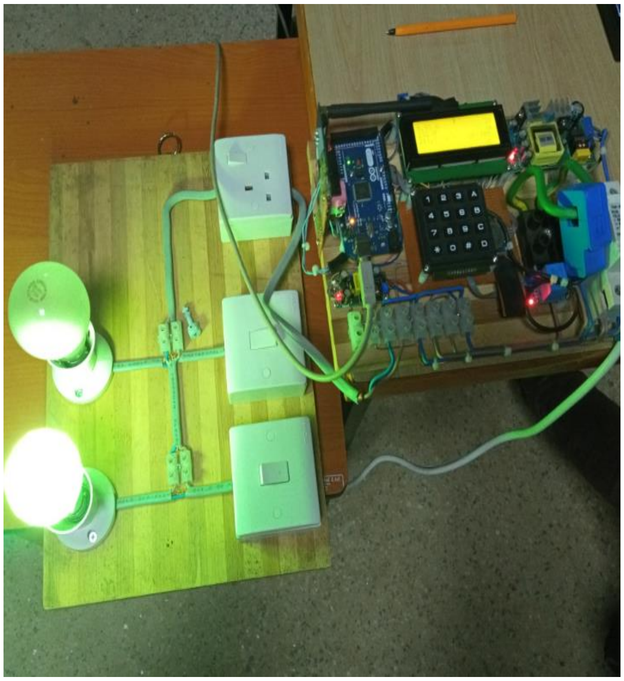


Fig. 9. Power Consumption Measurement on a Smart Meter Prototype.

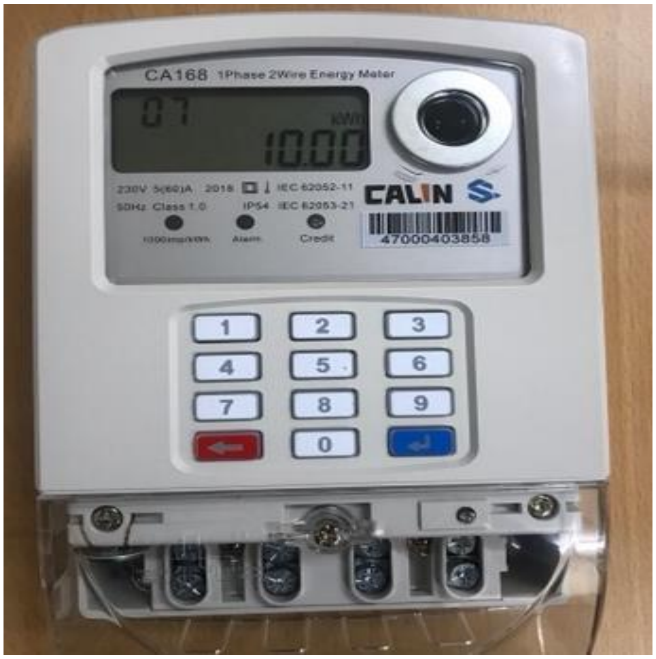

CA168 1Phase 2Wire Energy Meter
07
10.00
kWh
230V 5(60)A 2018 IEC 62052-11
50Hz Class 1.0 IP54 IEC 62053-21
1000imp/kWh Alarm Credit
CALIN
47000403858
1
2
3
4
5
6
7
8
9
0

Fig. 10. DDSD101-KT1 Digital Meter used as a Reference.

TABLE. I. POWER RATINGS OF THE SELECTED ELECTRIC APPLIANCES USED FOR THE PERFORMANCE TESTING

Appliance	Power Rating (Watts)
Incandescent Light Bulbs	220
Laptop Computer	60
Electric Kettle	1500
Fridge	620
Radio	160

V. Results and Discussion

Compared to other studies [16-21] which used various LoRa network parameters such as node-to-node data processing time, LoS and communication range as the key performance metrics, in this paper a simple approach was used to evaluate the performance of the proposed system. By assessing the accuracy of the power consumption data collected on the cloud server over a specific period of time as the performance metric, it ensures that all other communication variables are taken into account in the final data. In addition, certain performance metrics such as LoS and transmission range may not be extensively varied since smart meters are usually fixed at consumer houses, making parameters adjustment less flexible compared to other IoT applications such as in weather monitoring.

The results of power consumption data from both the cloud server and the digital reference meter are shown in Table II. From the table, the overall measurement performance of the integrated LoRa-GPRS electricity usage monitoring system is generally satisfactory. The proposed system produced a less than 5% absolute percentage error for all types of electrical appliances used during performance testing, which is within the acceptable standards.

TABLE. II. COMPARISION OF POWER CONSUMPTION RESULTS BETWEEN THE PROPOSED SYSTEM VERSUS THE DDSD101-KT DIGITAL METER

Appliance	Running Time (Hours)	Cloud Server Units (kWh)	DDSD101-KT1 Units (kWh)	Percentage Error (%)
Incandescent Light Bulbs	192.15	39.16	40.72	3.83
Laptop Computer	102.50	5.79	6.08	4.79
Electric Kettle	7.92	10.91	11.26	3.15
Fridge	22.43	13.03	13.52	3.66
Radio	17.0	2.35	2.43	3.42

The power consumption data recorded on the cloud server were slightly less than the same data recorded on the reference digital meter for all electrical appliances using during the test. This was attributed to the slight packet transmission delays between smart meter unit and the LoRa gateway which was largely caused by large high SF configuration in LoRa PHY layer [12]. High SF values increase the time on the air of the transmitted frames, although the signal strength against interference and communication range is increased.

To mitigate this type of systematic error, the LoRa modules must be configured with SF as low as 7, but this will have a negative impact on the signal transmission range between smart meter units and gateway base stations. A trade-off therefore needs to be made between signal quality and transmission range, but at the expense of the data throughput.

VI. CONCLUSION

In this paper, a performance evaluation of the cost-effective LoRa-GPRS integrated electricity usage monitoring for decentralized mini-grid systems was presented. Because of their decentralized nature, most mini-grid centers are located in remote areas with limited ICT infrastructure, hence it is typically not the best option to rely solely on public networks to monitor the smart meters. The combination of LPWAN and conventional cellular networks provides a better solution for the management of decentralized mini-grid centers. In this study, the performance evaluation of the proposed system was conducted by comparing the results obtained from the cloud server with the reference data collected locally using a standard digital meter.

This study found that the proposed system had satisfactory performance results promising its potential application for IoT monitoring in decentralized grids. Although the collected power consumption data were always slightly lower compared to the expected values, the percentage error was within the acceptable standards. To increase the accuracy of the data collected, LoRa modules need to be configured with lower SF values, which will also require additional cost of installing more LoRa gateways throughout the mini-grid centre. As a future work, the integrated mini-grid communication infrastructure needs to be improved in order to enable monitoring of more parameters such as power quality, power theft, Time of Use (ToU) pricing and event reporting.

ACKNOWLEDGMENT

This work was supported by the African Development Bank (AfDB), United Republic of Tanzania, through project No. P-Z1-IA0-016 and grant No. 2100155032816.

REFERENCES

[1] Hartvigsson E, Ehnberg J, Ahlgren E, Molander S. Assessment of load profiles in minigrids: A case in Tanzania. Proc Univ Power Eng Conf. 2015;15:1–5.

[2] Chaplin D, Mamun A, Protik A, Schurrer J, Vohra D, Bos K, et al. Grid Electricity Expansion in Tanzania: Findings from a Rigorous Impact Evaluation. Mathematica Policy Research. 2017.

[3] Ahlborg H, Hammar L. Drivers and barriers to rural electrification in Tanzania and Mozambique – grid extension, off-grid and renewable energy sources. 2011.

[4] Khadar Associate Professor AA, Ahamed Khan Professor J, MIT Madanpalli E, S Nagaraj Professor IM. Research Advancements Towards in Existing Smart Metering over Smart Grid [Internet]. Vol. 8, IJACSA) International Journal of Advanced Computer Science and Applications. 2017 [cited 2018 Oct 30]. Available from: www.ijacsa.thesai.org.

[5] Chattopadhyay D, Bazilian M, Lilienthal P. More power, less cost: Transitioning up the solar energy ladder from home systems to mini-grids. Electr J [Internet]. 2015;28(3):41–50. Available from: http://dx.doi.org/10.1016/j.tej.2015.03.009.

[6] Persia S, Carciofi C, Faccioli M. NB-IoT and LoRA connectivity analysis for M2M/IoT smart grids applications. 2017 AEIT Int Annu Conf Infrastructures Energy ICT Oppor Foster Innov AEIT. 2017.

[7] Nielsen JJ, Madueño GC, Pratas NK, Sørensen RB, Stefanovi´ C, Popovski P. What Can Wireless Cellular Technologies do about the Upcoming Smart Metering Traffic? IEEE Commun Mag - Commun Stand Suppl. 2015;(June):41–7.

[8] Cheng Y, Saputra H, Goh LM, Wu Y, Tower CS, Way F. Secure Smart Metering Based on LoRa Technology. 2018.

[9] Al-Fuqaha A, Guizani M, Mohammadi M, Aledhari M, Ayyash M. Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. IEEE Commun Surv Tutorials. 2015;17(4):2347–76.

[10] Yuan Z, Jin J, Sun L, Chin KW, Muntean GM. Ultra-Reliable IoT Communications with UAVs: A Swarm Use Case. IEEE Commun Mag. 2018;56(12):90–6.

[11] Sinha RS, Wei Y, Hwang SH. A survey on LPWA technology: LoRa and NB-IoT. ICT Express [Internet]. 2017;3(1):14–21. Available from: http://dx.doi.org/10.1016/j.icte.2017.03.004.

[12] Adelantado F, Vilajosana X, Tuset-Peiro P, Martinez B, Melia-Segui J, Watteyne T. Understanding the Limits of LoRaWAN. IEEE Commun Mag. 2017;55(9):34–40.

[13] Mekki K, Bajic E, Chaxel F, Meyer F. A comparative study of LPWAN technologies for large-scale IoT deployment. ICT Express [Internet]. 2019;5(1):1–7. Available from: https://doi.org/10.1016/j.icte.2017.12.005

[14] Li S, Raza U, Khan A. How Agile is the Adaptive Data Rate Mechanism of LoRaWAN? 2018 IEEE Glob Commun Conf GLOBECOM 2018 - Proc. 2018;

[15] Cattani M, Boano CA, Römer K. An experimental evaluation of the reliability of lora long-range low-power wireless communication. J Sens Actuator Networks. 2017;6(2).

[16] Petäjäjärvi J, Mikhaylov K, Pettissalo M, Janhunen J, Iinatti J. Performance of a low-power wide-area network based on lora technology: Doppler robustness, scalability, and coverage. Int J Distrib Sens Networks. 2017;13(3).

[17] Campo G, Gomez I, Calatrava S, Martinez R, Santamaria A. Power Distribution Monitoring Using LoRa : Coverage Analysis in Suburban Areas. 2018;233–8.

[18] Augustin A, Yi J, Clausen T, Townsley WM. A Study of LoRa : Long Range & Low Power Networks for the Internet of Things. 2016;1–18.

[19] Zanella A, Zorzi M. L ong -R ange C ommunications in U nlicensed B ands : T he R ising S tars in the I o T and S mart C ity S cenarios. 2016;(October):60–7.

[20] Ghiasimonfared A, Righini D, Marcuzzi F, Tonello AM. Development of a Hybrid LoRa / G3-PLC IoT Sensing Network : an Application Oriented Approach. 2017;(October):503–8.

[21] Haidine A, Aqqal A, Dahbi A. Performance Evaluation of Low-Power Wide Area based on LoRa Technology for Smart Metering. 2018 6th Int Conf Wirel Networks Mob Commun. 2018;1–6.

[22] Abubakar I, Khalid SN, Mustafa MW, Shareef H, Mustapha M. Calibration of ZMPT101B voltage sensor module using polynomial regression for accurate load monitoring. ARPN J Eng Appl Sci. 2017;12(4):1076–84.

[23] Barai GR, Krishnan S, Venkatesh B. Smart metering and functionalities of smart meters in smart grid - A review. 2015 IEEE Electr Power Energy Conf Smarter Resilient Power Syst EPEC 2015. 2016;138–45.

[24] ElectronicWings. Sensors Modules Real Time Clock Rtc Ds1307 Module | Sensors Modul... [Internet]. 2018 [cited 2019 Nov 10]. Available from: https://www.electronicwings.com/sensors-modules/real-time-clock-rtc-ds1307-module.

[25] Hackaday. Project | Hackable CMWX1ZZABZ (LoRa) Devices | Hackaday.io [Internet]. 2018 [cited 2019 Nov 11]. Available from: https://hackaday.io/project/35169/logs.

[26] Eric B. LoRa — LoRa documentation [Internet]. 2018. 2018 [cited 2019 Dec 19]. Available from: https://lora.readthedocs.io/en/latest/#.

[27] Wasion. DDSD101-KT1 Single Phase Electric Energy Meter [Internet]. 2019 [cited 2019 Dec 21]. Available from: http://www.wasion.com/EN/pro/484.html.

An Intellectual Detection System for Intrusions based on Collaborative Machine Learning

Dhikhi T[1]
Research Scholar, Saveetha School of Engineering
Saveetha Institute of Medical and Technical Sciences
Assistant Professor, SRM Institute of Science and Technology, India

Dr. M.S. Saravanan[2]
Professor, Saveetha School of Engineering
Saveetha Institute of Medical and Technical Sciences
India

***Abstract*—The necessity for safety of information in a network has inflated due to the impressive growth of web applications. Several methods of intrusion detection are used to detect irregularities which depend on precision, detection frequency, other parameters and are anticipated to familiarize to vigorously varying risk scenes. To accomplish consistent abnormalities detection in a network many machine learning algorithms have been formulated by researchers. A technique based on unsupervised machine learning that use two separate machine learning algorithms to identify anomalies in a network viz convolutional autoencoder and softmax classifier is proposed. These profound models were skilled as well as evaluated on NSLKDD test data sets on the NSLKDD training dataset. Using well-known classification metrics such as accuracy, precision and recall, these machine learning models were assessed. The developed intrusion detection system model experimental findings showed promising outcomes in anomaly detection systems for real-world implementation and is compared with the prevailing definitive machine learning techniques. This strategy increases the detection of network intrusion and offers a renewed intrusion detection study method.**

Keywords—Intrusion detection; machine learning; deep learning; convolutional autoencoder; softmax classifier; NSL-KDD dataset

I. Introduction

Detection of intrusion is to track device or network anomalies. This analyze both known and unknown attacks. Many approaches are processed to find the anomalies. As information is valuable resource the security is a crucial thing, thereby small complication intrusion detection system is a demanding assignment. Detection of intrusion detects external intrusions and monitors unauthorized inner user operations by recognizing and reacting to malicious network communication and computer utilization behaviour. IDSs plays an active role in network monitoring and was usually used as a network security element in latest years. Moreover, aims to detect intrusions by studying the process and features of intrusion conduct, thus allowing a process of invasion Two vital intrusion detection technologies be present, viz, anomaly detection and misuse detection [1]. Intrusion detection systems are classified as two: Network IDS, Host IDS [20]. The source of information accommodates IDS audit information. IDS triggers alarm by evaluating that audit information as it detects intrusion or attack [23].

Feature selection specifies the selection of the appropriate feature subset from extra dimensional quality depend on different calculation parameters, thereby achieving a model. In this research methods based on machine learning for intrusion detection is focused on [15]. Machine learning methods can be classified into i) supervised techniques ii) semi supervised learning iii) unsupervised learning methods. In this research multiple supervised learning methods for IDS is explored with regard to their performance metrics viz false alarm rate (FAR), accuracy, recall, F1 measure, time taken to train and test each classifier. NSLKDD database includes only selected dataset records that furnish a great research of different intrusion perception method for machine learning. NSLKDD incorporate 41 input together with class names [19]. In addition, the archive in the NSLKDD trained and tested sets is fair. This strength makes it inexpensive to execute the entire set of research without randomly selecting a small part. Accordingly, the assessment aggregation of various study job is coherent as well as similar. This excludes repetitive train records, because classifiers are not prone to ever-increasing records.

Rest of the paper is separated into five sections. Section II presents salient works associated to IDS. Section III offers the planned framework of Convolutional Softmax IDS and mention the different steps involved in the model. Section IV discusses on the evaluation criteria of the performance. Section V examines on the outcomes of the research along with comparison of results. Followed by Section VI describes the conclusion and next presents the references.

II. Existing Work

R. Vinayakumar [1] tells a profound DNN, a sort of deep learning model, is being studied for creating flexile and efficient IDS to categorize unanticipated and uncertain cyber-attacks. This sort of research promotes the identification of the finest algorithm that can operate efficiently to detect potential cyber-attacks. On several publicly accessible benchmark malware databases, a thorough analysis of DNN experiments and other classical machine learning classifiers is shown. It is confirmed by strict experimental testing that DNNs compared with classical machine learning classifiers, perform well. Shone [6] provides an original methodology of deep learning to detect interruptions that demonstrates that deep learning grouping is build using stacked NDAEs. This was used to evaluate the expenditure of the normal KDD Cup and

NSLKDD datasets in graphics processing unit enabled Tensor Flow. Moreover, measured the preparation time required for the stacked NDAE model, as well as a DBN model to examine the KDD99 dataset furnishing large accuracy. The well-known methods of machine learning were evaluated by I. Ahmad et al [10]. Support mechanism of vectors and machine of extreme learning. To evaluate the interruption detection system, the NSL datasets are considered. It is found in their assessment result that ELM is enhanced accurate. Al-Qatf [8] suggested a STL IDS approach that is effective in-depth training for learning features and dimensionality employing auto encoder machine to restructure an illustration of a novel function in an unsubstantiated way.

Naseer [12] explored the appropriate anomaly-based strategy to IDS produced on multiple profound ANN like convolutionary neural, regular neural systems and auto encoders which are competent on NSLKDD dataset. These are done on a GPU-related test bed that uses theano-backed keras. Evaluations were conducted using metrics of the organisation viz. Receiver operating attribute, curved region, accurate curve, mean average accuracy, conventional ML technique classification. M. H. Ali [13] implemented a Fast Learning Network knowledge model to support particle swarm optimization (PSO). This is useful in identifying entrant and KDD99 data set is endorsed. The scheme developed is associated with a nice variety of meta-heuristic schemes for tutoring both the extreme learning scheme and the FLN classification scheme. Within the testing precision of the training, PSO-FLN has defeated various teaching methods. P. Tao [14] recommends fresh inherited operation hinge on the features of GA and SVM algorithms, FWP-SVM. The stated technique reduces rate of SVM mistake that use a genetic algorithm option approach to modify the fitness algorithm. SVM's distinctive weights and limitations are simultaneously optimized, enabling optimum subset of features. The result of this article defines the right favorable rate of escalation and decreases the velocity of mistake. Q. Zhang [15] utilized fuzzy depends on the kernel – a set of KDD 99 data set for IDS validation and analysis. These blurred classifiers operate upon discrete, noise data's inaccuracy and vagueness, thus performing well in terms of effect and accuracy in reduction. The function selection techniques were commonly in use laterally with classifiers for network interruption identification.

H. Peng [16] exploited improved choice of features, FACO merged ant colony optimization algorithm for set of features. To improve the cataloging of separate classifiers, FACO is introduced. This optimization algorithm is an algorithm for simulation optimization that creates a detailed directed graph over n features, imitating ants ' scavenging behaviour. In addition, excess features are allocated to reduce the instance difficulty in grouping algorithms as well as enhance traffic allocation efficiency. Z. Wang [18] article assess different algorithms for intrusion catching domains using deep learning approaches and define different element application models for attack algorithms. Research indicates that the most commonly used highlights show their greater contribution to the exposure of intrusion detection created by the intense understanding and thus warrant additional consideration. Nisioti[17] provides comprehensive overview of an unattended and a crossed disturbance recognition approaches, examining their spatial potential. It characterizes the importance of highlighting construction techniques and also discuss actual IDS's should progress connection and attribution of the fundamental place. Moreover, suggested three innovative components related to communication on the outbound network. Haipeng Yao [4] introduces a multilevel model for IDS called multilevel semi-supervised ML (MSML). A notion of "pure cluster" is implemented in the module and implemented a semi-supervised hierarchical k-means algorithm. The "unknown pattern" and cluster-based technique is described in pattern discovery module. The updating module model offers a retraining mechanism. To evaluate MSML, the KDDCUP99 dataset is used. Experimental findings indicate that MSML is superior corresponding to general precision, F1-score, and unknown pattern recognition capacity to other current intrusion detection models.

III. Proposed Research

The system is planned to corporate trust unsupervised machine learning algorithms to boost the system's accuracy and efficiency. This model as shown in Fig. 1 compromises of distinct stages preprocessing, normalizing data, unifying data, feature extraction, classification, training and testing dataset. Two machine learning algorithms are implemented for training and testing dataset. A combination of both algorithms is also implemented for train and test datasets that improves the performance parameters [5]. The proposed approach uses collaborative supervised algorithms that offer an effective deep learning method. Thus, advances the performance of the model associated to the prevailing methods. This scheme combines convolutional autoencoder and softmax classifier for feature extraction and classification, respectively.

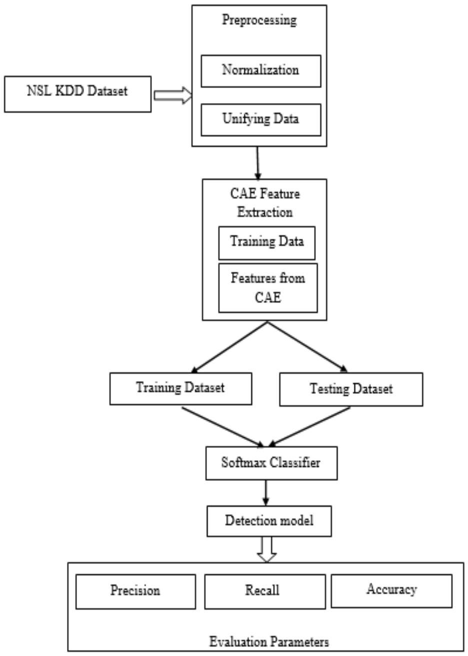


Fig. 1. Proposed Framework.

Preprocessing: The system's achievement is directly in proportion to the data set's accuracy, so the collection of data is a significant task. KDD 99[22] is utilized for anomaly detection valuation, occupies specific inspection data that consider a broad range of simulated intrusions. NSL-KDD [7] is a data set that is planned to resolve some of KDD99's key difficulties. The attacks in NSLKDD are classified mainly into four types as in Table I [11]. The protocol types in the dataset are shown in the Table II.

Preprocessing can be performed to remove symbolic characteristics in the procedure of identification. These types of symbolic data cannot be processed by the classifier to improve the efficiency of detection advancement. Pre-processing phase minimize data to a great extent as possible without loss of information and requires specific scheduling, preparation and testing. This helps to provide IDS with appropriate and effective information on computation, filtering fake rates and improving detection rate and to find patterns of attack and show suitable kinds of information for policy making by administrators.

In our approach the non -numeric values are changed to numerical values. Every attribute in the dataset are transformed into numeric values. In preprocessing, normalization of data is done. The intention of normalization is to alter numeric column values in the dataset to a popular scale without distorting value range distinctions [2]. Every dataset does not involve standardization for machine learning. It is only needed when there are distinct ranges of characteristics. Numerous NSL-KDD dataset features have wide ranges of maximum to minimum value, with a maximum value 58,329 also a minimum value 0. These features of dataset are normalized using min-max normalization and thereby maps the range from 0 to 1 using the equation (1)

$$v' = (v - min_F / max_F - min_F)\ (new\_max_F - new\_min_F) + new\_min_F \quad (1)$$

where v denotes the data point, min_F is the minimal value for all data and max_F is the upper limit value for all data factors. $new\_ min_F$ and new_max_F are the newly mapped minimum and maximum value, respectively.

Feature extraction: Feature Extraction [21] is method selecting and combining variables into features, effectively reducing the volume of data to be handled while still processing the original data set accurately and completely. Extraction of features can also decrease the quantity of redundant data for a particular assessment. The tests are directed to understand the effectiveness of performance and validate the efficiency of features mined from the two class and multiclass approach based on the NSL-KDD dataset. Training (NSLKDDTrain+) and testing (NSLKDDTest+) data are used separately for training and testing, respectively.

A. Autoencoder

An autoencoder (AE) neural network is an unsupervised machine learning algorithm that uses backpropagation to set goal values equal to the inputs. They are used in a smaller representation to reduce the size of the inputs and will recreate it from the compressed data if anyone wants the original data. AE exploits a balanced structure shown in Fig. 2 that consist of an encoder that constrict input into a fewer bits that comprise the actual information and a decoder part skilled to renew the input from the encoder's extracted features, each has a neural network with multiple hidden layers that are generally positioned evenly. It holds an unseen layer which studies the latent depiction of the input vector with smaller dimensions in a different feature space. The hidden layer of autoencoder, called bottleneck has lesser nodes than the input and the output layer. Then AE is called undercomplete. This is a method for deciding which aspects of observed data are appropriate information and which aspects can be rejected. Training task in an under-complete AE allows it to capture the utmost substantial features of bottleneck layer training data in order to recreate the input at the output layer. This is achieved by minimizing the loss function L(x, g(f(x))) which penalizes the difference between g(f(x)) and x. At this time, the data output of the hidden layer units is the maximum low-dimensional representation of the original data and contains all the information in the original data. AEs are created from numerous layers that link the outputs of the previous layer to the inputs of the next layer. Autoencoders will compress data just like they were educated on. Compared to the original inputs, the decompressed outputs will be reduced. Training specialized algorithm instances that will perform well on a particular type of input is simple. Upon receipt of normal data, the AE will produce similar outputs. With abnormal data, the AE must produce substantially dissimilar outputs and can therefore distinguish the abnormal data.

TABLE. I. CATEGORY OF ATTACKS

Category	**Attacks**
DoS	Neptune, Smurf, Pod, Land, Back, Udpstorm, process-table, mail bomb, Teardrop, Apache.
U2R	Buffer overflow, perl, rootkit, spy, Ps, Http tunnel, sql attack, worm, snmp guess, load module, Xterm.
R2L	Guess-password, ftp-write, Multihop, Warezmaster, Warezclient, snmpgetattack, Named, Xlock, Xsnoop, Send-mail, Imap, Phf.
Probe	Port-sweep, IP-sweep, Satan, Mscan, Nmap, saint.

TABLE. II. PROTOCOL WISE DISTRIBUTION IN NSL KDD DATASET

Protocols	TCP	UDP	ICMP
Count	18880	2621	1043

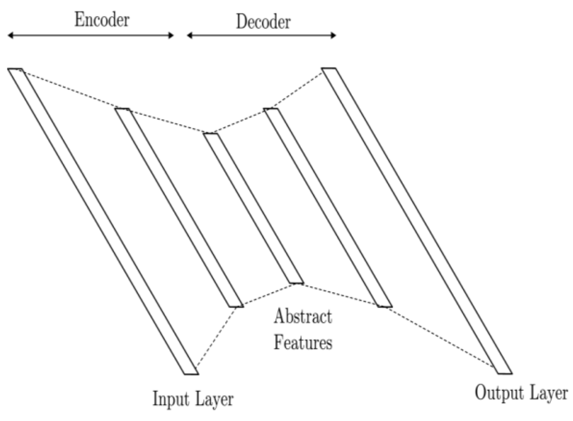


Fig. 2. Basic Structure of Autoencoder.

The method of encoding the hidden layer:

$$H = g_{\theta 1} (X) = \sigma (W_{ij}X + \phi_1) \quad (2)$$

The method of decoding the reconstruction layer from the hidden layer:

$$Y = g_{\theta 2} (H) = \sigma (W_{jk}H + \phi_2) \quad (3)$$

where $X = \{x_1, x_2, . . . , x_n\}$ the input data vector, $Y =\{y_1, y_2, . . . , y_n\}$ is the input data's reconstruction vector and $H = (h_1, h_2, . . . , h_m)$ is the hidden layer's low dimensional vector output, $X \in R^n$, $Y \in R^n$ and $H \in R^m$ in which n is the input vector's dimension and m is the no. of veiled units. $W_{ij} \in R^{m\times n}$, the matrix of the weight relation between the input and the hidden layer. $W_{jk} \in R^{n\times m}$, weight matrix of the hidden layer and the reference layer. $\phi_1 \in R^{n\times 1}$ and $\phi_2 \in R^{m\times 1}$ are the input and hidden layer bias vectors respectively. $g_{\theta 1}()$ and $g_{\theta 2}()$ are stimulation feature of the hidden layer neurons and the output layer neurons correspondingly, whose functions are to map to [0,1] the network summation result.

B. Convolutional Autoencoder

CAEs are identical to AE, but the distinction is that all the input locations are shared by weights in the CAE [24], maintaining the spatial position like CNN. Convolution Neural Networks (CNNs) are for handling high-dimensional data with some spatial denotation and the data can be images, video, speech sound signals, text sequence of characters, or any other multidimensional information. The loss function is same as autoencoder as given in equation (4).

$$e(x,y,W)= \frac{1}{2N} \sum_{i=1}^{n} || x_i - y_i ||_2^2 + \lambda \|W||_2^2 \quad (4)$$

λ is the regularization parameter for the regularization term.

CAE comprehends convolutional, deconvolutional, pooling, and unpooling layers. Convolutional layer outlines the data of a filter into a scalar with different parameters. In the function map, it joins multiple input activations in a filter's fixed receptive field to a singly activation output. Low-level features of the input frames are extracted in the initial layers of convolution layer and high-level features in later layers and vice versa in deconvolution layers. The pooling layer was planned for entirely supervised feedforward manners and shows a constant factor in the latent representation. Moreover, permits composite representations but lessens the three-dimensional size of representations by reducing the number of parameters and computation. Unpooling layer accomplishes the inverse pooling operation, reconstructing the original size of individually quadrilateral sub-region.

CAEs are state-of - the-art tools to learn convolutional filters in an unsupervised manner and then can be tested to any input to extract features. Instead, these features can be used to perform any function requiring a compact representation of the data, such as classification. CAEs [9] scale fine to realistic high-dimensional data due to their convolutionary nature, as the numeral of parameters essential to yield an activation map is permanently the same inspite of the input size. The encoder selects features over convolution and pooling layers and the decoder rebuild the input over unpooling and reordered convolution layers. Each decoder layer equivalent to that in the encoder shall be located in the reverse sequence of the encoder layers. Initially the input data is transformed to binary image using character-level binary image transformation technique. Then the output of this is fed into two convolution and deconvolution layers, two pooling layers and unpooling layers for feature extraction.

Consider the message's maximum permissible length is X and any character that exceeds it is neglected. The message is therefore converted into 68x1x X. For the given kth character within the permitted characters, all its positions are found inside the data. Then, their respective channel k locations in the picture are set to 1. The steps involved in this transformation technique are initially the given data is converted into reverse order and transform each character into a vector with a specific length. Then transform a set of vectors into one dimensional image with the specified number of channels. Image matrix is converted into an array, rescale it between 0 and 1.

The latent representation of the k-th feature map for a mono-channel input x is given by

$$h^k = \sigma (x * W^k + b^k) \quad (5)$$

Where the bias is transmitted to the entire map, σ is an activation function, ∗ signifies the 2D convolution. The minimizing cost function is the mean squared error

$$E(\theta) =1/2n \sum_{k=1}^{n} (x_k – y_k)^2 \quad (6)$$

As the backpropagation algorithm is used for standard networks to measure the gradient of the error function with respect to the parameters. Convolution operations can effectively achieve this with the following formula.

$$\delta E(\theta)/ \delta W^k = x * \delta\, h^k + h^k * \delta y \quad (7)$$

δh and δy are the deltas of the hidden states and the reconstruction of the hidden states, respectively. Using stochastic gradient descent, the weights are then updated.

1) Classification: The output from extraction of the CAE features was transferred to a classifier to be categorized using two separate classification, the binary classification that tells attack or normal data and the five classification that includes four class of attacks and the normal. Softmax is a soft version of max function. This divides the whole (1) instead of choosing a maximum value with the highest element having the largest portion of the distribution, but other smaller elements do get some of it [3]. This softmax property which outputs a distribution of probabilities appropriate for probabilistic clarification in classification tasks. We use this as the last layer in neural networks because of the necessary property of softmax function outputting a probability distribution. To do this, the derivative or gradient must be measured and transferred back to the preceding layer through backpropagation.

$$\delta p_i/\delta\, a_j = (\delta\, ea_i / \sum_{k=1}^{n} ea_k) / \delta\, a_j \quad (8)$$

Cross entropy reveals the difference between the assumption of distribution of output and the original distribution. This is considered a loss function in neural networks that have output layer softmax activations. It is defined as

$$H(y, p) = - \Sigma_i y_i \log (p_i) \qquad (9)$$

Loss function tests how consistent the set of parameters in the training dataset is with respect to ground truth labels. The loss function has been established in such a way that good training data predictions are tantamount to having a small loss.

IV. Evaluation Discussion

The anticipated IDS framework is tested on the NSL-KDD dataset that consists of approximately 22,544 features and has huge quantity of network traffic information, marked as usual or abnormal. The performance assessment of collaborative unsupervised machine learning is finished using NSLKDD training and testing data. Train datasets were used to train the model of machine learning and test datasets were accustomed to evaluate the trained model of machine learning. There are four possible states for each activity observed, in terms of the performance metrics of an IDS. The measures of the assessment are considered and measured and can be described as:

True positive (TP): irregularity decently characterized as anomalousness.

False positive (FP): irregularity poorly characterized as anomalousness.

True negative (TN): regular data correctly characterized as unusual.

False negative (FN): irregularity inaccurately characterized standard.

Accuracy: say the exact classification fraction of all records in the test set as shown in (10).

Precision: say the right intrusion estimate fraction with predictable overall intrusions as in (11)

Recall: say the allowed intrusion estimate fraction separated by the full amount of valid intrusion possibilities in the test set in (12).

$$A=(TP+TN) / (TP+TN+FP+FN) \qquad (10)$$

$$P=TP/(TP+FP) \qquad (11)$$

$$R=TP/(TP+FN) \qquad (12)$$

ROC Curves summarize the trade-off for a predictive model using different probability thresholds between the true positive rate and the false positive rate. ROC curves depend on the true positive, true negative, false positive and false negative. RoC is a plot of False Positive Rate (FPR) of binary classifiers against True Positive Rate (TPR). Area under RoC Curve (AuC) is a measure of how well a binary classifier can accomplish label predictions. This shows the performance of a classical model for a binary classifier.

V. Outcomes

Performance evaluation is done on testing data using CAE and softmax classifier. The experiments were performed on the basis of the NSL-KDD dataset to check performance efficiency and verify the reliability of the low-dimensional characteristics obtained from our two-class and multi-class classification strategy. Moreover, it compares the performance with the existing methods and several recent approaches like SVM, KNN, STL IDS, CNN.

In Fig. 3, execution metrics such as precision, recall and accuracy of CAE- Softmax is compared on training dataset. Accuracy, Precision and recall for two class training data are 99.9, 99.5, 99.5 respectively and five class training dataset are 97.92, 99.39, 99 respectively. Assessment of the same on test data for accuracy, precision and recall is depicted in Fig. 4 with values 92,91,91.05 for two class categories respectively and 97,95,91 for five class categories respectively. So, after several models have been introduced and evaluated, results show that the CAE- Softmax model being proposed has better performance. The planned method is then analyzed to the present algorithms as in Fig. 5 to give a better accuracy. The graph outcome of types of protocols is shown in Fig. 6. Fig. 7 shows the ROC Curve for NSLKDD dataset. These show that the model reduces the false alarm levels to an acceptable level to retain total safety against serious attacks. This system provides high detection rate.

Table III shows the comparison of precision, recall, accuracy the projected model on training data for two class and multiclass. Table IV illustrates the same evaluation method of the model on the test data. The accuracy of the existing IDS algorithms with the CAE – Softmax IDS is compared and the values are given in Table V. The experimentation outcome shows the attacks in NSL KDD test data as in Table VI.

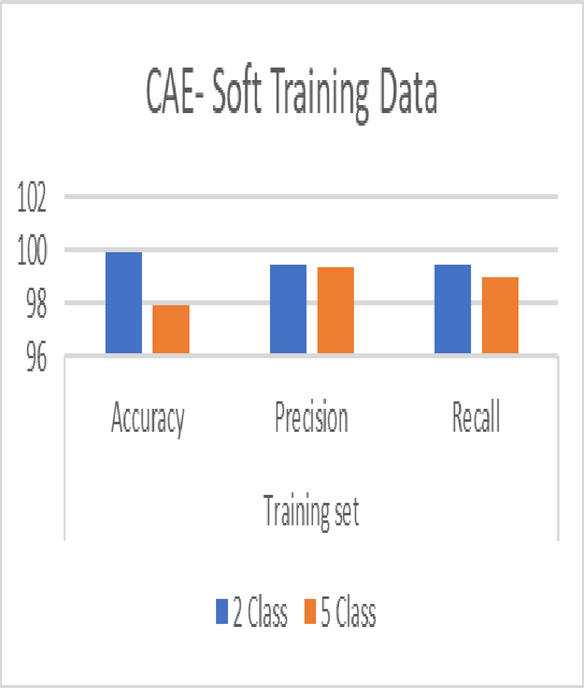


Fig. 3. Assessment of CAE-Soft IDS on Training Data.

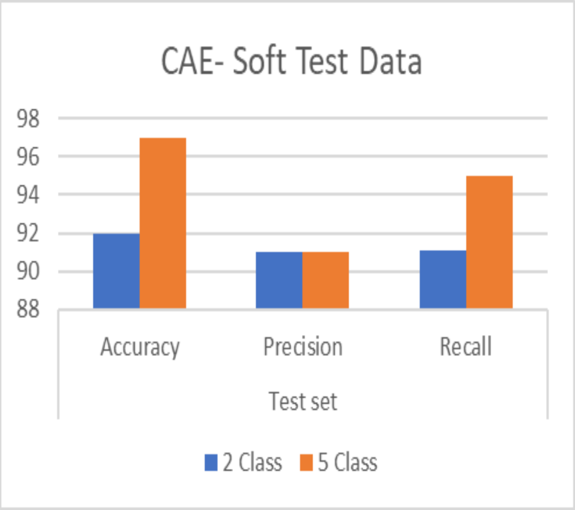


Fig. 4. Assessment of CAE-Soft IDS on Test Data.

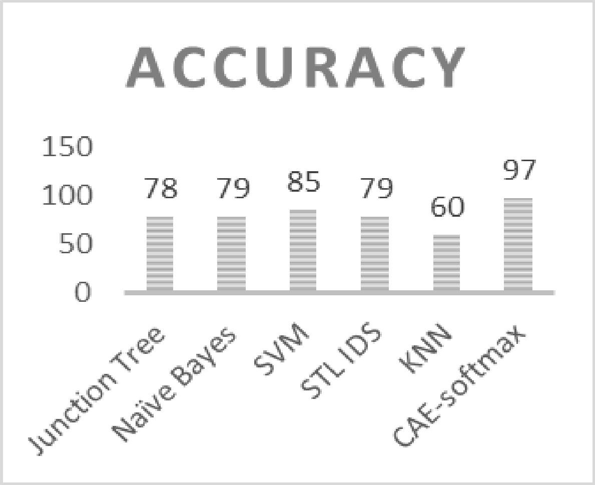


Fig. 5. Existing IDS Versus CAE-Softmax.

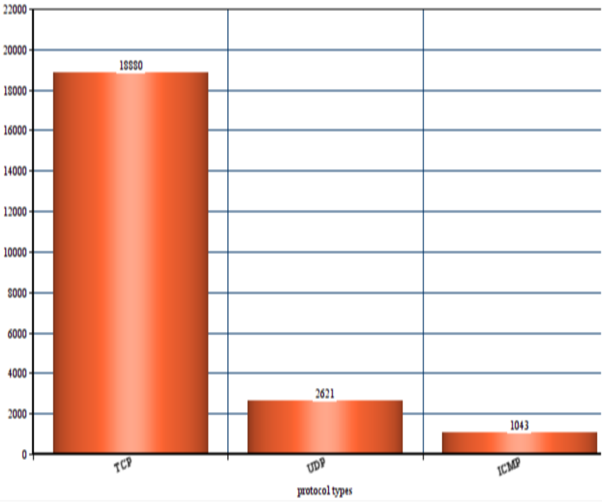


Fig. 6. Protocol Types in NSLKDD.

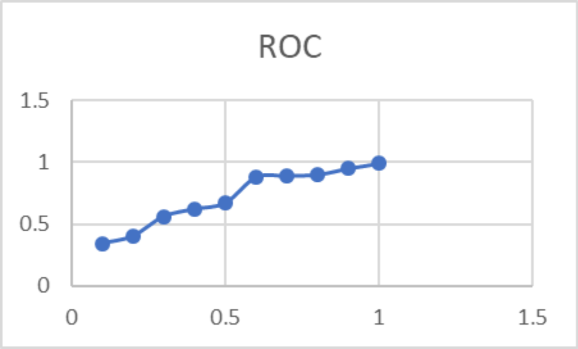


Fig. 7. ROC Cirve for NSLKDD Data.

TABLE. III. PERFORMANCE METRICS ON TRAINING DATA

	Training set		
	Accuracy	***Precision***	***Recall***
2 Class	99.9	99.5	99.5
5 Class	97.92	99.39	99

TABLE. IV. PERFORMANCE METRICS ON TEST DATA

	Test set		
	Accuracy	***Precision***	***Recall***
2 Class	92	91	91.05
5 Class	97	91	95

TABLE. V. COMPARISON OF ACCURACY OF ALGORITHMS

ALGORITHM	**ACCURACY**
Junction Tree	78
Naïve Bayes	79
SVM	85
STL IDS	79
KNN	60
CAE-softmax	97

TABLE. VI. COUNT OF ATTACKS IN NSL KDD

Attacks	**Count**
Normal	9711
DOS	7456
Probe	2421
U2R	200
R2L	2756

VI. CONCLUSION

The suggested CAE-Softmax IDS scheme is an enhanced method of intrusion that utilizes methods of machine learning to select and classify features. This technique is a pledge to reduce false positive as well as false negative. The above model analyzed the convolutional autoencoder, Softmax Classifier and existing IDS SVM, KNN, STL methods and outperformed present diverse methods in testing precision and training. By applying this to the actual network to implement it more effectively, further step can be taken. This can be applied to an improved efficiency for all class categories.

Furthermore, IDS outputs can be presented to any real time applications like investigations to construct timeline of an attack and associate attacks to find out the trespasser.

REFERENCES

[1] R. Vinayakumar, Mamoun Alazab, K. P. Somani, Prabaharan Poornachandran, Ameer Al-Nemrat, Sitalakshmi Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System" IEEE Access Volume 7, 2019 page. 41525 -41550.

[2] Dhikhi T, M. S. Saravanan "An Enhanced Intelligent Intrusion Detection System using Machine Learning" International Journal of Innovative Technology and Exploring Engineering, Vol 8, Issue 9, July 2019, pp. 2177-2181.

[3] Xin Ye And Qiuyu Zhu 'Class-Incremental Learning Based on Feature Extraction of CNN With Optimized Softmax and One-Class Classifiers' IEEE Access, Vol 7, 2019 pp. 42024-42031.

[4] Haipeng Yao, Danyang Fu, Peiying Zhang, Maozhen Li, and Yunjie Liu MSML: A Novel Multilevel Semi-Supervised Machine Learning Framework for Intrusion Detection System IEEE Internet Of Things Journal, Vol. 6, No. 2, April 2019.

[5] Gael Kamdem, Momo ZIAZET 'Convolutional Neural Network for Intrusion Detection System In Cyber Physical Systems', ResearchGate, May 2019.

[6] Nathan Shone, Tran Nguyen Ngoc, Vu DinhPhai, Qi Shi, "A Deep Learning Approach to Network Intrusion Detection", IEEE Transactions on Emerging Topics in Computational Intelligence, Vol. 2, No. 1, February 2018, pp. 41-50.

[7] MajdLatah, LeventToker, "Towards an efficient anomaly-based intrusion detection for software-defined networks" IET Netw., 2018, Vol. 7 Iss. 6, pp. 453-459.

[8] Majjed Al-Qatf, Yu Lasheng, Mohammed Al-Habib, And Kamal Al-9Sabahi "Deep Learning Approach Combining Sparse Autoencoder With SVM for Network Intrusion Detection" IEEE. Translations and content mining, VOLUME 6, 2018, pp. 52843-52856.

[9] Marco Maggipinto, Chiara Masiero,Alessandro Beghi,Gian AntonioSusto 'A Convolutional Autoencoder Approach for Feature Extraction in Virtual Metrology' , Procedia Manufacturing, Volume 17, 2018, Pages 126-133.

[10] Iftikhar Ahmad, Mohammad Basheri, Muhammad Javed Iqbal, And Aneel Rahim" Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection" IEEE Transactions on Special section on survivability Strategies for Emerging Wireless Networks, Volume 6 May 2018 pp. 33789-33795.

[11] Congyuan Xu, Jizhong Shen, Xin Du, and Fan Zhang "An Intrusion Detection System Using a Deep Neural Network With Gated Recurrent Units" IEEE,Volume: 6, 2018, Page(s): 48697 – 48707.

[12] Sheraz Naseer, Yasir Saleem, Shehzad Khalid, Muhammad Khawar Bashir, Jihun Han, Muhammad Munwar Iqbal, And Kijun Han" Enhanced Network Anomaly Detection Based on Deep Neural Networks" IEEE Transactions on Special Section on Cyber-Threats and Countermeasures in The Healthcare Sector Volume 6, 2018 pp.48111-48246.

[13] Mohammed HasanAli, Bahaa Abbas Dawood Al Mohammed, Alyani Ismail, and MohamadFadliZolkipli "A New Intrusion Detection System Based on Fast Learning Network and Particle Swarm Optimization" IEEE Transactions, Volume 6, 2018, pp. 20255-20261.

[14] PeiyingTao, Zhe Sun, And Zhixin Sun "An Improved Intrusion Detection Algorithm Based on GA and SVM" IEEE Transactions on Special Section on Human-Centered Smart Systems and Technologies, Volume 6,2018 pp. 13624-13631.

[15] Qiangyi Zhang, YanpengQu, Ansheng Deng "Network Intrusion Detection Using Kernel-based Fuzzy-rough Feature Selection", IEEE International Conference on Fuzzy Systems,2018.

[16] HuijunPeng, Chun Ying, Shuhua Tan, Bing Hu, And ZhixinSun, "An Improved Feature Selection Algorithm Based on Ant Colony Optimization", IEEE Transactions Volume 6, 2018, pp. 69203-69209.

[17] Antonia Nisioti, AlexiosMylonas, Paul D. Yoo, and VasiliosKatos "From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods", IEEE Communications Surveys & Tutorials, VOL. 20, NO. 4,2018, pp. 3369-3388.

[18] Zheng Wang, "Deep Learning-Based Intrusion Detection with Adversaries", IEEE Transactions on Special Section on Challenges and Opportunities of Big Data Against Cyber Crime, Volume 6, 2018, pp.38367-38384.

[19] L. M. Ibrahim, D. T. Basheer, and M. S. Mahmod, ''A comparison study for intrusion database (KDD99, NSL-KDD) based on self organization map (SOM) artificial neural network,'' J. Eng. Sci. Technol., vol. 8, no. 1, pp. 107–119, 2013.

[20] F. A. B. H. Ali and Y. Y. Len, ''Development of host based intrusion detection system for log files,'' in Proc. IEEE Symp. Bus., Eng. Ind. Appl. (ISBEIA), pp. 281–285, Sep. 2011.

[21] S. Rifai, P. Vincent, X. Müller, X. Glorot, and Y. Bengio, ''Contractive auto-encoders: Explicit invariance during feature extraction,'' in Proc. 28th Int. Conf. Int. Conf. Mach. Learn. (ICML), 2011, pp. 833–840.

[22] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, ''A detailed analysis of the KDD CUP 99 data set,'' in Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl. (CISDA). Piscataway, NJ, USA: IEEE Press, 2009, pp. 53–58.

[23] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, ''Intrusion detection by machine learning: A review,'' Expert Syst. Appl., vol. 36, no. 10, pp. 11994–12000, 2009.

[24] Seungyoung Park, Myungjin Kim, Seokwoo Lee 'Anomaly Detection for HTTP Using Convolutional Autoencoders' IEEE Access, Volume: 6, Page(s): 70884 - 70901.

Dataset Augmentation for Machine Learning Applications of Dental Radiography

Shahid Khan[1], Altaf Mukati[2]
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology
Karachi, Pakistan

***Abstract*—The performance of any machine learning algorithm heavily depends on the quality and quantity of the training data. Machine learning algorithms, driven by training data can accurately predict and produce the right outcome when trained through enough amount of quality data. In the medical applications, being more critical, the accuracy is of utmost importance. Obtaining medical imaging data, enough to train machine learning algorithm is difficult due to a variety of reasons. An effort has been made to produce an augmented dental radiography dataset to train machine learning algorithms. 116 panoramic dental radiographs have been manually segmented for each tooth producing 32 classes of teeth. Out of 3712 images of individual tooth, 2910 were used for machine learning through general augmentation methods that include rotation, intensity transformation and flipping of the images, creating a massive dataset of 5.12 million unique images. The dataset is labeled and classified into 32 classes. This dataset can be used to train deep convolutional neural networks to perform classification and segmentation of teeth in x-rays, Cone-Beam CT scans and other radiographs. We retrained AlexNet on a subset of 80,000 images of the entire dataset and obtained classification accuracy of 98.88% on 10 classes. The retraining on original dataset yielded 88.31%. The result is evident of nearly a 10% increase in the performance of the classifier trained on the augmented dataset. The training and validation datasets include teeth affected with metal objects. The manually segmented dataset can be used as a benchmark to evaluate the performance of machine learning algorithms for performing tooth segmentation and tooth classification.**

Keywords—Data augmentation; Cone-Beam Computed Tomography; dental X-Rays; panoramic; dataset; classification; deep convolutional neural network; benchmark

I. Introduction

The machine learning, especially deep convolutional neural network has been playing a key role in the advancements in the medical imaging field [1]. Today, more than ever medical imaging applications are powered by artificial intelligence. Medical practitioners use computer-based systems for automatic diagnosis, analysis, planning and simulation of diseases and treatment planning [1].

As the name indicates, data-driven methods depend on the training data [2]. Popular image classification methods [3] [4] [5] are trained on millions of images to be able to produce necessary outcome. The large datasets used to train these models are produced manually by teams which consist of many human resources. Some datasets with millions of images are produced by crowed sourcing. Enormous resources and man hours are spent to produce such datasets. The accuracy of applications of convolutional neural networks is more critical in medical applications. The accuracy is directly related to the amount and quality of the training data. Generally, the more training data we provide, the better results we get. Unlike AlexNet [4] and Keras where most images are downloaded from the internet, it is hard to put together a quality training dataset consisting millions of dental images. In addition, the number of instances in each class in dental radiographs are naturally imbalanced. A normal subject has only 4 canines and 0 – 4 wisdom teeth as compared to 8 instances in each of the rest of classes. The absence of wisdom tooth is also common, the ratio of subjects with no wisdom tooth is even higher in the youth. To solve this issue of lack of training dataset and natural imbalance, data augmentation is proposed to generate synthetic radiography datasets for training deep convolutional neural networks to perform classification and segmentation in CBCT, X-Rays and panoramic radiographs.

We manually segmented 116 panoramic radiographs to obtain 2,910 individual teeth. Fig. 1 shows an instance from the source panoramic dental radiographs. We applied common image augmentation technique such as rotation, resizing, flipping and intensity transformation to produce a synthetic dataset of total 5.12 million images containing 160,000 images of each of 32 teeth of humans. The dataset is labeled by directory names. We retained AlexNet [4] on a subset from this dataset containing 80,000 images and obtained 98.88% classification accuracy on 10 classes. The dataset we produced is not only a useful resource for training deep neural networks for tooth segmentation, classification and labeling but it can also be used as a benchmark for evaluating the performance of deep learning models.

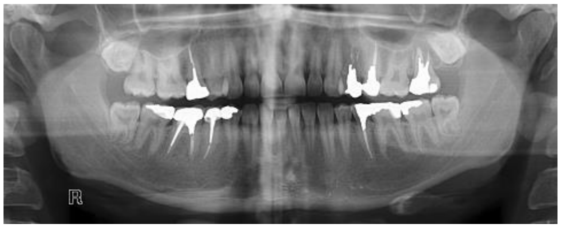


Fig. 1. Source Panoramic Dental Radiograph.

II. Related Work

To maximize the generalization capability of the deep learning models, we require a lot of training data. State-of-the-art deep neutral networks for example has millions of parameters which requires huge amount of data to achieve good results. In the field of image classification, researchers have been using a variety of dataset augmentation techniques to supplement the training and validation data. The augmentation of adding random noise in the original data points has been used in [2] to automatically generate augmented datasets. Classifiers trained on this generated dataset outperformed the classifier trained on the original network. In another example of dataset augmentation [6] in the field of medical image processing application of deep learning where a limited dataset of only 182 liver lesion cases is enlarged first using classical augmentation techniques and further enlarged using Generative Adversarial Networks (GAN). The reported that augmentation through GAN improved the classifier performance by nearly 10%. The classical as well as GAN based image dataset augmentation techniques helped in improving the image classification in underwater images. The scientists in the research [7] reported that they improved the classification confidence in the submarine and sonar image classification through classical and GANs based techniques. Image data augmentation also helps in improving classification confidence in classifying images from 3D volumetric radiographs. A novel approach [8] of image data augmentation is proposed in this research where low-resolution images are generated from high resolution volumetric CT scans. It is reported the significant improvement in classification has been achieved through their novel augmentation approach.

III. Materials and Methodology

The dataset used in [9] has been used as the starting point. This dataset contains 116 volumes of panoramic dental radiographs. We manually segmented all individual teeth present in the 116 volumes, of which 2910 individual teeth qualified for inclusion into our dataset. The 2910 teeth included the instances where there were external bodies such as implants and metallic fillings present. The dataset also included rotten and overlapping teeth.

Table I shows the distribution of 2910 images of individual teeth in 32 classes. It is evident that the ratio of wisdom teeth at number 1, 16, 17 and 32 is relatively low.

The distribution of the dataset in 10 classes is given in Table II.

Common image data augmentation techniques, as used in [10], such as rotation, resizing, intensity transformation, and flipping are applied in a sequence to produce exponential results.

A. Rotate

Each 400 x 400 image containing single tooth is first rotated to the right side one degree per iteration of a loop and the rotated image is captured as a unique image. The process is performed 10 times to produce 10 unique images. The original source image that has no rotation is used in the next step to perform the same operation in the opposite direction to produce 10 more unique images. It is considered to keep the rotation rate minimum to avoid the potential confusion between mandibular and maxillary teeth.

TABLE. I. Organic/Natural Data item Count in Dataset for each Tooth Type

Tooth	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Count	41	102	94	96	93	104	95	97
Tooth	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Count	99	95	97	90	91	90	96	51
Tooth	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Count	60	95	76	97	103	100	102	102
Tooth	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**
Count	99	98	107	104	99	76	91	70

TABLE. II. Distribution of Organic/Natural Dataset in 10 Classes

Class	**Count**	**Class**	**Count**
Maxillary Incisors	386	**Mandibular Incisors**	401
Maxillary Canine	201	**Mandibular Canine**	208
Maxillary Premolar	370	**Mandibular Premolar**	403
Maxillary Molar	382	**Mandibular Molar**	338
Maxillary Wisdom	91	**Mandibular Wisdom**	130

B. Resize

The second manipulation is done by resizing. The resizing is performed on each of the images that we produced in the previous step. To perform the resizing, each 400 x 400 image is enlarged 1 pixel per iteration diagonally. The operation is performed on the original image in reverse order by producing additional 10 images by reducing the size 1px per iteration.

C. Intensity Transformation

The next manipulation is achieved by increasing and decreasing grey level values of the pixels. The input for this step is the collection of all images that we produced in the last step. Grey level augmentation is very significant because the radiography equipment manufacturers do not follow common standard. To produce augmented data of various contrasts, we increased grey level by a random number between 10 and 40, a total of ten times on each input image. We then performed the same process for another ten times by decreasing the grey level by a random number between 10 and 40.

D. Horizontal Flip

In the final step, we horizontally flipped all the images that we obtained in the last step. This only doubles the amount of data that we have obtained so far. We did not perform the vertical flip because it will create a confusion between maxillary and mandibular teeth.

Fig. 2 shows a set of organic teeth followed by instances of augmented images.

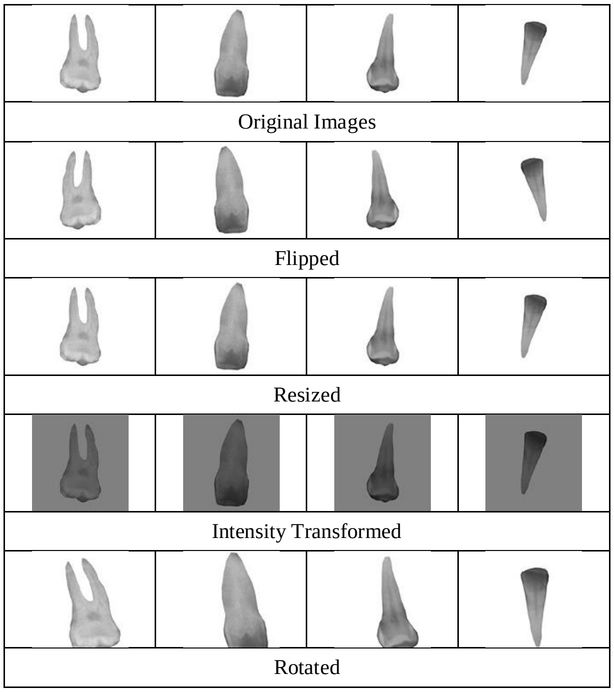


Fig. 2. Organic and Augmented Teeth Images.

IV. Experiment and Evaluation

We retrained AlexNet on the original dataset on 10 classes which yielded classification accuracy of 88.31%. Although the performance on original dataset is relatively low but for a very small dataset of only 2910 images and having imbalanced dataset, it is still favorable. The retraining of AlexNet with similar parameters and factors on the augmented and balanced dataset yielded 98.88% classification accuracy. The results are evident that the classifier trained on the augmented dataset outperformed the classifier trained on the original dataset with a considerable margin. In this experiment, we used a subset of 80,000 images from the dataset. Example images are shown in Fig. 2. We divided this subset into training and validation subsets by random sampling with ratio of 70% for training and 30% for validation. The training process converged after 1200 iterations. As training options, we used "Stochastic Gradient Descend with Momentum" at learning rate of 0.0001 and mini batch size of 4. Fig. 3 and 4 show the training progress.

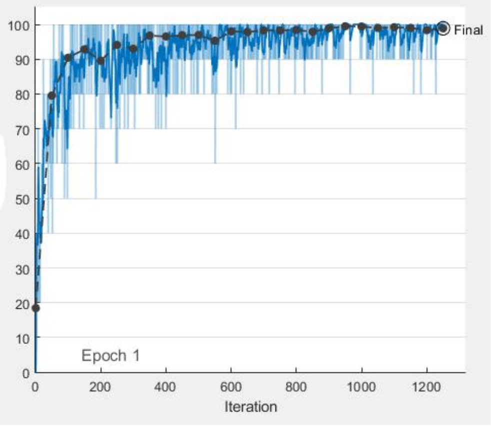


Fig. 3. Training Progress – Accuracy (%).

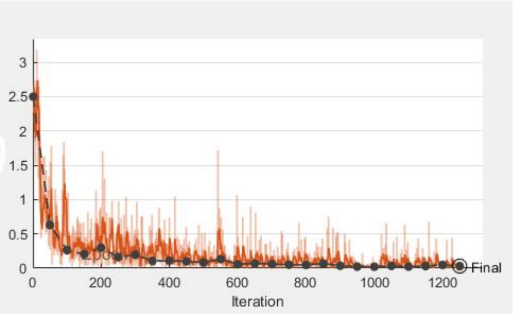


Fig. 4. Training Progress – Loss.

A copy of the pretrained neural network and a subset of the dataset can be downloaded from the website https://shahidsci.com/dataset for evaluation purpose.

V. Limitation

The source dataset in our work is from a single imaging facility. Although the outcome of source dataset is evident that several different radiography imaging machines have been used to produce this source dataset, but it is worthwhile to have a diverse dataset that is produced on a broader range of equipment. Further, the source radiography is conducted in single locality, which limits the diversity of the dataset.

VI. Conclusion

An effort has been made to solve the deficiency and imbalance of dental radiography data through proposed data augmentation method. We produced a massive dataset of 5.12 million images from dental radiographs. This dataset can be used to train ever-craving deep convolutional neural networks to perform segmentation and classification on CBCT dental images, x-rays and other dental radiographs. The dataset can be downloaded from the website mentioned in the previous section. We performed transfer learning from AlexNet on a subset of 80,000 images from the dataset and obtained accuracy of 98.88% on 10 classes which on nearly 10% more than the accuracy of the classifier trained on the original dataset.

VII. Future Work

In the future work, we plan to apply augmentation through Generative Adversarial Networks (GANs) alongside the classical augmentation techniques that we used. It is anticipated that it will further increase the classification confidence in the training of the classifiers. We also plan to diversify the dataset by adding more original X-Ray and CBCT radiography images from different parts of the world. We also plan to train a novel machine learning model on the produced dataset that can be used to classify teeth in radiography images from a verity of different sources.

References

[1] X. Xu, C. Liu, and Y. Zheng, "3D Tooth Segmentation and Labeling using Deep Convolutional Neural Networks," IEEE Trans. Vis. Comput. Graph., vol. XX, no. XX, pp. 1–13, 2018.

[2] K. Fujita, "Data Augmentation using Evolutionary Image Processing," 2018 Digit. Image Comput. Tech. Appl., pp. 1–6, 2018.

[3] C. Szegedy et al., "Going Deeper with Convolutions," in Computer Vision and Pattern Recognition (CVPR), 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[5] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in International Conference on Learning Representations, 2015.

[6] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “Synthetic data augmentation using GAN for improved liver lesion classification,” Proc. - Int. Symp. Biomed. Imaging, vol. 2018-April, no. Isbi, pp. 289–293, 2018.

[7] Y. Xu, Y. Zhang, H. Wang, and X. Liu, “Underwater image classification using deep convolutional neural networks and data augmentation,” 2017 IEEE Int. Conf. Signal Process. Commun. Comput. ICSPCC 2017, vol. 2017-January, pp. 1–5, 2017.

[8] M. Farhadi and A. H. Foruzan, “Data Augmentation of CT Images of Liver Tumors to Reconstruct Super-Resolution Slices based on a Multi-Frame Approach,” ICEE 2019 - 27th Iran. Conf. Electr. Eng., pp. 1783–1786, 2019.

[9] A. H. Abdi, “Automatic segmentation of mandible in panoramic x-ray.”

[10] Y. Miki et al., “Classification of teeth in cone-beam CT using deep convolutional neural network,” Comput. Biol. Med., vol. 80, pp. 24–29, 2017.

An Optimal Prediction Model's Credit Risk: The Implementation of the Backward Elimination and Forward Regression Method

Sara HALOUI[1]
PhD- Student
ENCG Kenitra, Ibn Tofail University -Kenitra, MOROCCO
Research Laboratory: Management Sciences of Organizations

Abdeslam El MOUDDEN[2]
Research- Professor
ENCG Kenitra, Ibn Tofail University -Kenitra, MOROCCO
Research Laboratory: Management Sciences of Organizations

Abstract—**The purpose of this paper is to verify whether there is a relationship between credit risk, main threat to the banks, and the demographic, marital, cultural and socio-economic characteristics of a sample of 40 credit applicants, by using the optimal backward elimination model and the forward regression method. Following the statistical modeling, the final result allows us to know the variables that have a degree of significance lower than 5%, and therefore a significant relationship with the credit risk, namely the CSP (Socio-occupational category), the amount of credit requested, the repayment term and the type of credit. However, by implementing the second method, the place of residence variable was selected as an impacting variable for the chosen model. Overall, these features will help us better predict the risk of bank credit.**

Keywords*—*Credit risk; prediction; optimal model; backward elimination; statistical modeling

I. Introduction

Generally in the banking environment, among a variety of risks to which a bank may be exposed, credit risk remains the biggest and most dangerous, its control and evaluation are essential steps to continually improve the performance of banks in the financial market [1]. This necessarily involves the implementation of instruments and devices to anticipate and predict this type of risk. In this sense, setting up credible mechanisms of banking risk management to ensure the stability of the international banking system, was the main goal behind the enactment of prudential rules, commonly referred to as the Basel Accords, which are generally applicable to all banks with significant international activity [2]. For the bank, the credit risk management is therefore a matter of survival. Moreover, logistic regression is the most widely used model in the development of credit scoring model [6].

In Morocco, the failure of banks' customers is increasing, since outstanding debts record the same upward trend every year, nevertheless no bank would resist at such conjunctures. Prior management of credit risk is fundamental since it takes into account the assessment and prevention of this risk.

In this article, we want to show the significant link between credit risk and the socio-economic, marital, cultural and demographic variables of the credit applicants. We would like to point out that the central issue to which we are trying to provide empirical answers is the following: "How can the prediction of bank credit risk be improved?"

To answer the problematic of our research, we will review generalities about credit risk, overview of the inventory of outstanding debts in Morocco would also be necessary in our research, and we will of course discuss prudential and banking regulation as the international banking environment is governed by the Basel agreements, essentially the bank credit component. In the end, this paper will use a statistical model to predict the risk of bank credit.

II. Literature Review

A. Credit Risk : A General View

1) The conceptual framework: As soon as an economic agent gives credit to counterparty, a risk relationship is established between the creditor and his debtor, the latter can indeed, with good or bad faith, do not pay his debt on the agreed date. The risk of meeting a commitment to settle a debt is the credit risk [13]. Hence, the credit transaction might create the risk that a debtor cannot honor its commitments [15].

According to the Bank of International Settlements (2011) [5], the credit risk lies in the fact that the counterparty may not fulfill its obligations according to the agreed upon contractual conditions. A financial asset is considered unpaid when a counterparty has not made the payment due at the contractual maturity.

It can also manifest itself in counterparty defaults, failures in commitments or concentration of bad debts. In general, credit risk is a particular risk from a lending transaction, and the occurrence of a negative event affects the debt on which the debtor is engaged. It is one of the main causes of bank failure.

2) Credit risk assessment: To mitigate credit risk, it is highly recommended for banks to assess repayment capacity and guarantees, to select operations taking into account profitability and costs (financing, operational, cost of risk, return on equity), take into account the economic and legal environment, and finally monitor bad debts and provisioning [11].

Banks have to set up a credit policy framed by a committee of commitments which sets the objectives (type of clientele, of credit, sectors and geographical areas), credit terms (rates, margins, guarantees) and delegations of power. The credit processing procedures go through the study of the demand (taking of information and evaluation of solvency), monitoring (detection of the risk of insolvency) and internal control of counterparty risk. The bank has real expertise in assessing the counterparty risk of individuals (consumer loans or mortgages) or companies (loans financing cash or investment) [14]. However, it is necessary to point out that it is common to use the term counterparty risk to refer exclusively to the credit risk while this is not the case.

B. Credit Risk and Basel Agreements

The bank is one of the most regulated economic sectors in the world. Otherwise, the State of the banking sector is indicative for the condition of the entire financial market and, by extension, of the entire economy [16]. To prevent the scale of banking crises, control devices have been introduced, mainly by Basel Committee. Their purpose is to help retail and investment banks control their credit or market risks through a prudential approach that combines risk measurement with a minimum equity allocation [14].

Cieply (2018) [7] concluded that the purpose of prudential regulation is to reduce the probability of bankruptcy of banks. For this, in their normal business, banks are required to meet management standards. They aim is to contain each of the major risks to which banks are exposed, particularly the risk of illiquidity and insolvency.

So, each of the three Basel agreements entails regulatory constraints imposed on the banking institutions that they are expected to respect, in order to maintain their financial stability.

1) Basel accord I: The Basel accord I, created in 1988, sets up a system to better control the measurement of credit risks. A minimum ratio of 8% is then imposed between a capital of a bank and the risks it bears on the market or the credit risks it takes with its customers [9].

This regulatory constraint is in the form of a ratio called Cooke which, according to Cieply (2018) [7], require credit institutions to constantly comply with a ratio of at least 8% between their own funds and the commitments of credit weighted against their risk. The weighting was based in Basel I on the nature of the counterparty by following a purely institutional criterion.

2) Basel accord II: Dhafer and Cesbron (2012) [9] argue that in the face of a financial system that has become more complex, in particular because of the growing importance of globalization, the Basel Committee strengthened its regulation. The new device is based on three pillars: 1st pillar – the capital requirement with a ratio of 8%, 2nd pillar – the establishment of a more comprehensive prudential supervisor procedure with, inter alia, the introduction of an internal risk management model, 3rd pillar – the need for better, transparent and uniform communication, which strengthens market discipline.

3) According to dhafer (2012) [8], the purpose of pillar II is twofold: On the one hand, to encourage banks to develop techniques for managing their risks and their level of capital and, on the other hand, to enable regulators to increase regulatory capital requirements if necessary. This need must be applied in two ways:

- Stress testing: Banks must prove during simulations of extreme situations, the validity of its own funds in case of economic crisis.
- Back testing: The banks must prove the validity of their statistical methods over quite long periods (5 to 7 years).

4) Basel accord III reinforcement of the basel II system: In 2010, the Basel Committee published the Basel Accord III in order to meet the Basel II limits and to prevent future crises [12].

In the same context, Dhafer and Cesbron [9] argue that following the 2008 crisis, the Basel Committee reacted and took a number of steps to strengthen the “resilience” of the banking sector. It is then a matter of consolidating the solvency of banks, to develop a greater liquidity monitoring, to improve the ability of banks to absorb shocks, resulting from financial and economic stress, and finally reduce and control the risks of overflow to the real economy.

These agreements focus on four points that are: the redefinition of own funds, the establishment of a precautionary capital or mattress and counter-cyclical measures, setting up ratios and covering certain risks [17].

C. Statement of Outstanding Debts of Banks in Morocco

In recent years, Moroccan banks have turned to the household segment to boost credit activity. However, the interest in this category is not without its repercussions: the aggravation and rapid evolution of the outstanding debts of individuals is one of the remarkable effects.

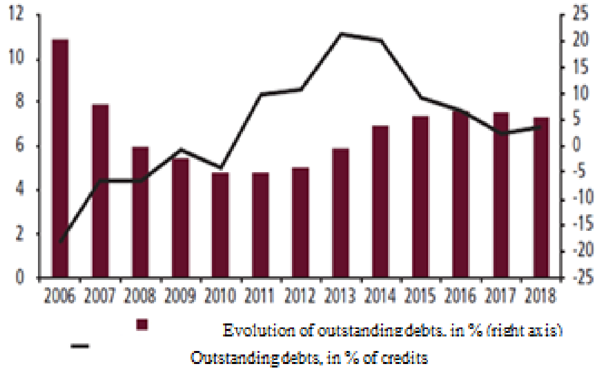
The monetary statistics provided by Bank Al-Maghrib under its report on financial stability (2018) [3] revealed as described in Fig. 1 that at the end of 2018 the outstanding debts recorded an increase of 3.7% against 2.3% a year earlier to reach 65.3 billion dirhams. This increase primarily concerns loans granted to households, which is more than 2.7 billion dirhams compared to 2017. On the other hand, the outstanding debts of companies fell by 1%. Moderately, the rate of outstanding debts posted is 7.3% against respectively 7.5% and 7.6% in 2017 and 2016.

According to the latest figures published by the BAM 2019 [4], at the end of June 2019, Table I shows that households are left with almost 27.5 billion DH of arrears at banks, given that the outstanding receivables total an amount of 67.7 billion dirhams and represent 7.5% of the total loans outstanding for the same period.

TABLE. I. BREAKDOWN OF BANK CREDIT AND OUTSTANDING DEBTS

	Outstanding at the end of June 2019*	Monthly variation (in %)	Annual growth rate (in %)					
			June18	**Sept.18**	**Dec.18**	**March19**	**May19**	**June19**
Bank credit	900,4	3,5	1,8	2,9	3,2	5,1	4,4	5,1
Households credit	338,2	-0,2	3,7	4,9	5,3	5,7	5,8	5,4
Consumer credits	55	0,1	5,7	6,0	6,2	6,0	4,9	4,1
Real estate loans	219,1	-0,2	4,8	4,6	6,0	5,2	5,0	3,8
Housing loans	211,6	-0,1	4,3	4,6	5,7	5,6	5,7	4,9
Loans to real estate developers	6,8	-4,1	16,1	10,3	16,2	-2,3	-12,4	-19,8
Outstanding debts	67,7	0	2,9	3,4	2,5	5,5	3,7	5,7
Private non-financial corporations	39,5	1,9	3,9	3,2	2,5	0,7	0,1	2,8
Households	27,5	-1,9	1,2	3,9	2,7	13,7	11,3	11,3
Ratio of outstanding debts			7,5	7,7	7,5	7,7	7,8	7,5
*In billions of DH								

Source: BAM, Août 2019, p.36



Source: BAM, 2018, p.75

Fig. 1. Evolution of Outstanding Debts of Banks.

The results recorded in relation to the situation of outstanding debts born by Moroccan banks, invite us to shed light on these findings, to think in a thorough way, and to further develop appropriate tools and mechanisms to assess this risk that could put banks in critical conditions.

III. MODELING OF CREDIT RISK

A. Work Methodology

Our empirical approach imposes the adoption of a methodology that is based on sampling, data analysis. These empirical results will be presented in statistical tables with their interpretations.

1) Sampling: To find the optimal model, we took a sample of 40 credit applicants using the simple random sampling method to estimate and predict credit risk through the backward elimination and forward regression method. The data is collected through the questionnaire method distributed to 40 respondents who are credit applicants. We choose these two statistical methods in order to search the optimal model in our scientific research.

Table II presents the composition of the sample of defaulting and non-defaulting customers.

TABLE. II. BREAKDOWN OF DEFAULTING AND NON-DEFAULTING CUSTOMERS IN THE SAMPLE

	Sample	**Percentage**
Defaulting customer	**13**	**32.5%**
Non-defaulting customer	**27**	**67.5%**
Total	**40**	**100%**

2) Characteristics of the sample: The independent variables retained with their modalities for the analysis of the data are declined below:

- Age: The age of the customer [Under 30years, 30 to 39 years, 40 to 49years, 50 years and over].
- Gender: The gender of the client [Man, Woman].
- Etatmatri: The marital status of client [Single, Married, Divorced, Widower].
- Milieuderés: The place of residence of client [Urban, Rural].
- Zonegéo: The geographical area where the customer lives [Casablanca-Settat, Rabat-Salé-Kénitra, Fès-Meknès, Marrakech-Asfi, Tanger-Tétouan-Alhoceima, Oriental, others].
- CSP: The socio-occupational category ofthe client [Employee of a small company, employee of a large company, civil servant, tradesman and entrepreneur, liberal professions].
- Income (Revenu): The income received by the client [less than 4000dh, from 4000 to 6000dh, from 6000 to 10000dh, Greater than 10000dh].
- Nbredoscréd: The number of credit files available to the client [0, 1, 2, 3].
- Montcrédsolli: The amount of credit requested [5000 to 10000dh, 10000 to 20000dh, 20000 to 50000dh, 50000 to 100000dh, 100000 and over].

- Duréerem: The repayment term of the credit [12months, 12-36months, 36-60months, more than 60months].
- Typecrédit: The type of credit desired [Consumer credit, Real estate credit].
- Degrérat: The difference between the amount requested and the amount awarded [=0, >0].

3) Descriptive statistics: First, we present the descriptive statistics relating to the explanatory variables in Table III as follow:

From the statistics above associated with the explanatory variables, we can observe a strong dispersion of the observations.

4) Hypotheses to test: Two hypotheses are to be tested by statistical modeling [10]:

H1: There is a significant relationship between the credit risk and the demographic, marital, cultural and socio-economic characteristics of credit applicants.

H2: Some variables may be important and impacting in predicting credit risk.

5) The variables of the problem: The variable to be explained and the explanatory variables adopted for the treatment of the problematic posed above is given as:

$$Y = b_0 + b_1X_1 + \cdots + b_kX_k + \varepsilon, k = 12$$

Y = the variable to be explained (RISQCREDIT)

The explanatory variables:

X_1 = the age of the client (Age)

X_2 = the gender of the client (Sexe)

X_3 = the marital status of the client (Etatmatrimonial)

X_4 = the place of residence of the client (Milieuderés)

X_5 = the geographical area where the client lives (Zonegéo)

X_6 = the socio-occupational category of the client (CSP)

X_7 = the income received by the client (Revenu)

X_8 =the number of credit files available to the client (Nbredoscréd)

X_9 = the amount of credit requested (Montcrédsolli)

X_{10} = the repayment term of the credit (Duréerem)

X_{11} = the type of credit desired (Typecrédit)

X_{12}= the difference between the amount requested and the amount awarded (Degrérat)

$\boldsymbol{b_i}$ = Coefficients representing the linear combination of the predictor and the constant

ε = The error

The dependent variable is credit risk. It is a dichotomous binary variable denoted “RISQCREDIT” such as:

RISQCREDIT = 0 if the client is solvent and repay his credits at maturity.

RISQCREDIT = 1 if the client is insolvent and will not repay his credits at maturity.

This makes it possible to highlight the degree of significance of the independent variables with respect to the dependent variable.

TABLE. III. EXPLANATORY VARIABLES

Explanatory variables	Mean	Median	Max	Min	Std. dev	Skewness	Kurtosis	Jarque-Bera	probability
Age	1.325	1.000	3.000	0.000	1.071484	0.210487	1.819931	2.616306	0.270319
Gender	0.450	0.000	1.000	0.000	0.503831	0.201008	1.040404	6.669387	0.035625
Marital Status	0.950	1.000	3.000	0.000	0.845804	0.866483	3.474737	5.380910	0.067850
Place of residence	0.250	0.000	1.000	0.000	0.438529	1.154701	2.333333	9.629630	0.008109
Geographical area	1.225	1.000	5.000	0.000	1.310461	1.238532	3.851849	11.43583	0.003287
Socio-occupational Category	1.775	2.000	4.000	0.000	1.270726	0.202029	2.179973	1.392843	0.498365
Income	1.750	2.000	3.000	0.000	1.031553	-0.195169	1.858760	2.424654	0.297504
Number of Credit files	0.575	0.000	3.000	0.000	0.747217	1.234518	4.202353	12.56966	0.001864
Amount of credit requested	2.625	3.000	4.000	0.000	1.314368	-0.790170	2.605819	4.421424	0.109623
Repayment term of credit	2.325	3.000	3.000	0.000	0.828576	-0.934844	2.912953	5.838853	0.053965
The type of credit	0.250	0.000	1.000	0.000	0.438529	1.154701	2.333333	9.629630	0.008109
The degree of rationing	0.325	0.000	1.000	0.000	0.474342	0.747265	1.558405	7.186359	0.027511

Source: These statistics were prepared using EViews

B. *Empirical Results*

1) The selection of the optimal model: Now, we proceed to the tests of the choice of the most significant explanatory variables in relation to the variable to be explained, and we will do this through the method of elimination of non-significant variables at the threshold of 5% one by one in order to make a successive correction of the proposed model. In this context, we use the backward elimination method and the forward regression.

- The Backward Elimination method

The initial model adopted using the different variables that is supposed to be explanatory is given as follows in Table IV:

- Income variable is the least significant. We eliminate it and we continue to re-estimate the equation. The new model then takes the new form (Annex 1).
- The variable to be eliminated this time is the marital status variable. Then the new model is obtained in Annex 2.
- Then, we eliminate the variable AGE. Hence the new model is in Annex 3.
- The variable to be eliminated now is the number of credit files available to the client (NBREDOSCRED). And the re-estimated model obtained in Annex 4.
- Then, the variable associated to the place of residence (MILIEUDERES) is eliminated. The new model takes the new form in Annex 5.
- And this time, the variable to eliminate is Gender. The new model (Annex 6) takes the new form.
- Then, we eliminate the variable associated with the degree of rationing (DEGRERAT) which means the difference between the amount requested and the amount awarded. And the re-estimated model becomes as indicated in Annex 7.
- And then the variable associated with the geographical area (ZONEGEO) is eliminated. Hence, the following and the last model adopted is given in Table V:

In this new final model, the remaining variables, CSP, MONTCREDSOLLI, DUREEREM, and TYPECREDIT are significant at the threshold of α=5%. This is the optimal model obtained by the backward regression method.

TABLE. IV. THE INITIAL MODEL ADOPTED

Dependent Variable: RISQCREDIT				
Method: Least Squares				
Date: 07/29/19 Time: 21:01				
Sample: 1 40				
Included observations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
AGE	0.018147	0.092667	0.195827	0.8462
SEXE	0.114491	0.160077	0.715222	0.4804
ETATMATRIMONIAL	0.014819	0.107447	0.137916	0.8913
MILIEUDERES	0.112921	0.206375	0.547163	0.5886
ZONEGEO	0.063002	0.060039	1.049345	0.3030
CSP	0.088344	0.065254	1.353843	0.1866
REVENU	0.011269	0.143172	0.078711	0.9378
NBREDOSCRED	-0.046702	0.137017	-0.340850	0.7358
MONTCREDSOLLI	-0.281958	0.153218	-1.840240	0.0764
DUREEREM	0.215460	0.145782	1.477959	0.1506
TYPECREDIT	0.410711	0.249898	1.643511	0.1115
DEGRERAT	0.213677	0.208225	1.026186	0.3136
R-squared	0.324790	Mean dependent var		0.325000
Adjusted R-squared	0.059529	S.D. dependent var		0.474342
S.E. of regression	0.460007	Akaike info criterion		1.528173
Sumsquaredresid	5.924969	Schwarz criterion		2.034837
Log likelihood	-18.56346	Hannan-Quinn criter.		1.711367
Durbin-Watson stat	1.881081			

Source: These estimates were prepared using Eviews

TABLE. V. THE LAST MODEL ADOPTED

Dependent Variable: RISQCREDIT				
Method: Least Squares				
Date: 07/29/19 Time: 21:47				
Sample: 1 40				
Included observations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
CSP	0.131344	0.050025	2.625593	0.0126
MONTCREDSOLLI	-0.364113	0.092067	-3.954881	0.0003
DUREEREM	0.374530	0.098955	3.784852	0.0006
TYPECREDIT	0.462116	0.206267	2.240377	0.0313
R-squared	0.216425	Mean dependent var		0.325000
Adjusted R-squared	0.151127	S.D. dependent var		0.474342
S.E. of regression	0.437031	Akaike info criterion		1.277016
Sumsquaredresid	6.875868	Schwarz criterion		1.445903
Log likelihood	-21.54031	Hannan-Quinn criter.		1.338080
Durbin-Watson stat	1.836085			

Source: These estimates were prepared using EViews

TABLE. VI. FORWARD REGRESSION METHOD

	RISQCREDIT
RISQCREDIT	1.000000
AGE	-0.213150
SEXE	0.230673
ETATMATRIMONIAL	-0.086280
MILIEUDERES	0.338983
ZONEGEO	0.085593
CSP	-0.003190
REVENU	-0.615730
NBREDOSCRED	-0.251392
MONTCREDSOLLI	-0.539792
DUREEREM	-0.340877
TYPECREDIT	-0.030817
DEGRERAT	0.202279

Source: These estimates were prepared using EViews

- Forward Regression method

The forward regression consists in selecting one by one the explanatory variables according to the highest correlation coefficient recorded with the variable to be explained.

Table VI presents the correlation coefficients connecting the dependent variable to the independent variables. The used method prompts us to use the highest correlation coefficient recorded with the credit risk (Y).

The highest coefficient correlation we have selected is the variable place of residence (MILIEUDERES); we then check the meaning of the correlation coefficient:

$$\hat{t} = \frac{0.338983}{\sqrt{\frac{1-0.338983}{40}}} = 16.677$$

Table VII gives us a new model with the variable selected.

Regarding the forward regression, the place of residence variable (MILIEUDERES), even if it was eliminated by the backward elimination method, since it is not significant at the 5% level, it was selected by forward regression since it scored the highest correlation with the variable to be explained. This occurred just after we assumed that the addition of this variable is significant, which means that it is significantly correlated with credit risk.

TABLE. VII. THE NEW MODEL WITH THE EXPLANATORY VARIABLE SELECTED

DependentVariable: RISQCREDIT				
Method: Least Squares				
Date: 07/30/19 Time: 20:50				
Sample: 1 40				
Includedobservations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.233333	0.082540	2.826909	0.0075
MILIEUDERES	0.366667	0.165080	2.221143	0.0324
R-squared	0.114910	Meandependent var		0.325000
Adjusted R-squared	0.091618	S.D. dependent var		0.474342
S.E. of regression	0.452091	Akaike info criterion		1.298839
Sumsquaredresid	7.766667	Schwarz criterion		1.383283
Log likelihood	-23.97677	Hannan-Quinn criter.		1.329371
F-statistic	4.933476	Durbin-Watson stat		1.770529
Prob(F-statistic)	0.032376			

Source: These estimates were prepared using Eviews

2) Analysis of results: Overall and according to tests done on Eviews, the obtained results showed that the explanatory variables, among the twelve variables are significant at the 5% level, namely the CSP (socio-occupational category), the MONTCREDSOLLI (amount of credit requested), the DUREEREM (repayment period), and the TYPECREDIT (type of credit). This is based on the optimal model and the Backward Elimination, which consists of eliminating each non-significant explanatory variable. This means that the remaining variables, which are socio-economic characteristics related to the profile of each applicant for credit, have an influence on credit risk. More explicitly, these variables predict the default of credit applicants at maturity and can prove their impact on the probability of repaying debts or not. While the Forward Regression method gave us a new result, which is the variable "place of residence". We can therefore conclude that there is a significant relationship between credit risk and the socio-economic, marital, cultural and demographic characteristics of credit applicants, and that some independent variables can be impacting and predictive of credit risk, therefore, we confirm the two hypotheses (H1) and (H2) that have been tested and verified through modeling.

IV. CONCLUSION

Credit risk is a particular risk that can be one of the main causes of bank failure in the event of actual occurrence. Thus, this article concludes that the two methods of statistical modeling (Backward Elimination and Forward Regression), carried out on Eviews, could lead to different results, and proposed new models.

The purpose of this research was to select the explanatory variables that predict the likelihood of credit risk for banks among credit applicants. To do this, two statistical methods are used to perform this modeling on a sample of 40 borrowers of bank credits.

The above results lead us to assume that it is difficult to bypass the credit risk in Morocco which represents the variable to explain in our case. However, it is important to note that the explanatory variables selected by the models, such as the type of credit, the repayment period, the socio-occupational category, the amount of credit requested and finally the place of residence, represent an influence tools on the probability of repayment of credits.

REFERENCES

[1] Ahmed, A., Seyoum, A., Kedir, H., &Kedir, S. (2015). Credit Risk Management OfMfis Found In Ethiopia. European Scientific Journal, 11(31).

[2] Angima, C.B., Mwangi, M., Kaijage, E., &Ogutu, M. (2017). Actuarial Risk Management Practices, Underwriting Risk and Performance of P & C Insurance Firms in East Africa. European Scientific Journal. 13(22). doi: 10.19044/esj.2017.v13n22p207.

[3] Bank Al-Maghrib (2018). Rapport sur la stabilité financière, Number 6.

[4] BAM (2019). Revue mensuelle de la conjoncture économique, monétaire et financière. August.

[5] Bank of International Settlements (2011). « Réforme de la réglementation financière : Réalisations, risques et perspectives », 81eAnnual Report, Bâle Suisse, June, 228p.

[6] Chen, H., & Xiang, Y. (2017). The study of credit scoring model based on group lasso. Procedia computer science, 122, pp.677-684.

[7] Cieply, S. (2018). Quel avenir pour la relation banque-entreprise ?, Caen, EMS editions, 99p.

[8] Dhafer, S. (2012). L'impact de la réglementation de bale III sur les métiers des salariés des banques, étude thématique, Université Lille Nord de France, Septembre, 50p.

[9] Dhafer, S., & Cesbron, C. (2012). L'impact de la réglementation de bale III sur les métiers des salariés des banques, thematicstudy, University Lille Nord de France et SKEMA et cabinet MEDIATION Observatoire des métiers, des qualifications et de l'égalité professionnelle entre les femmes et les hommes dans la banque, pp.20-23.

[10] Ghassan, H., & Raiss, N., & El-Moudden, A. (2008). Testing the Effect of the Land Tax on Tourism Investment.Munich Personal RePEc

Archive (MPRA), n°56384, pp.1-11. June. (Online at http://mpra.ub.uni-muenchen.de/56384/).

[11] Greuning, H.,&BrajovicBratanovic, S. (2004). Analyse et gestion du risque bancaire : Un cadre de référence pour l'évaluation de la gouvernance d'entreprise et du risque financier, 1ère édition, Paris, Editions ESKA.

[12] Hologne, A.-L. (2014). Le stress financier impacte-t-il les banques ? Cas de Belfius et d'Axa Bank Europe, Mémoire de recherche, Université catholique de Louvain, Louvain School of management, 126p.

[13] Kharoubi, C., & Thomas, P. (2016). Analyse du risque de crédit : Banque & Marchés, 2ème édition, Paris, RB édition, 164p.

[14] Pierandrei, L. (2015). Risk Management, Gestion des risques en entreprise, banque et assurance, Paris, Dunod, 320p. 2015.

[15] Siqani, S. H., &Sekiraca, E. (2016). The Impact of the Internal Audit in Reducing Credit Risk in Commercial Banks in Kosovo. European Scientific Journal, 12(4). doi: 10.19044/esj.2016.v12n4p268.

[16] Sumna, P. (2013). Credit risk dynamics in Czech Republic. European Scientific Journal, 9(16).

[17] Yota, R. (2016). Le test de l'effet médiateur de la prise de risque et des effets modérateurs de la réglementation prudentielle et de la taille de la banque, Doctoral Thesis in Managment Sciences, supported on 12 Decembre 2016, University of Artois, 253p.

ANNEXURES

ANNEX. I. ELIMINATION OF THE VARIABLE INCOME AND THE RE-ESTIMATION OF THE EQUATION

Dependent Variable: RISQCREDIT				
Method: Least Squares				
Date: 29/07/19 Time: 21:10				
Sample: 1 40				
Included observations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
AGE	0.018242	0.091057	0.200337	0.8426
SEXE	0.116106	0.156013	0.744207	0.4627
ETATMATRIMONIAL	0.016870	0.102438	0.164682	0.8703
MILIEUDERES	0.111442	0.201966	0.551787	0.5853
ZONEGEO	0.062670	0.058856	1.064805	0.2957
CSP	0.089524	0.062408	1.434497	0.1621
NBREDOSCRED	-0.040635	0.111317	-0.365034	0.7177
MONTCREDSOLLI	-0.274525	0.118566	-2.315371	0.0279
DUREEREM	0.213862	0.141866	1.507494	0.1425
TYPECREDIT	0.404838	0.234376	1.727297	0.0948
DEGRERAT	0.205480	0.177195	1.159626	0.2557
R-squared	0.324640	Meandependent var		0.325000
Adjusted R-squared	0.091758	S.D. dependent var		0.474342
S.E. of regression	0.452056	Akaike info criterion		1.478394
Sumsquaredresid	5.926280	Schwarz criterion		1.942836
Log likelihood	-18.56789	Hannan-Quinn criter.		1.646322
Durbin-Watson stat	1.879857			

ANNEX. II. Elimination of the Variable Marital Status and the Re-Estimation of the Equation

DependentVariable: RISQCREDIT				
Method: Least Squares				
Date: 07/29/19 Time: 21:16				
Sample: 1 40				
Included observations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
AGE	0.024276	0.081999	0.296049	0.7692
SEXE	0.116249	0.153460	0.757520	0.4546
MILIEUDERES	0.111904	0.198645	0.563335	0.5774
ZONEGEO	0.060792	0.056797	1.070345	0.2930
CSP	0.091816	0.059843	1.534282	0.1354
NBREDOSCRED	-0.039854	0.109398	-0.364304	0.7182
MONTCREDSOLLI	-0.275347	0.116524	-2.362996	0.0248
DUREEREM	0.216841	0.138408	1.566677	0.1277
TYPECREDIT	0.405861	0.230464	1.761064	0.0884
DEGRERAT	0.208043	0.173625	1.198231	0.2402
R-squared	0.324009	Meandependent var		0.325000
Adjusted R-squared	0.121212	S.D. dependent var		0.474342
S.E. of regression	0.444666	Akaike info criterion		1.429329
Sumsquaredresid	5.931822	Schwarz criterion		1.851549
Log likelihood	-18.58658	Hannan-Quinn criter.		1.581990
Durbin-Watson stat	1.860989			

ANNEX. III. Elimination of the Variable AGE and the Re-Estimation of the Equation

Dependent Variable: RISQCREDIT				
Method: Least Squares				
Date: 29/07/19 Time: 21:20				
Sample: 1 40				
Included observations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
SEXE	0.122524	0.149736	0.818267	0.4195
MILIEUDERES	0.103856	0.193859	0.535730	0.5960
ZONEGEO	0.066633	0.052470	1.269930	0.2136
CSP	0.096744	0.056629	1.708374	0.0976
NBREDOSCRED	-0.030639	0.103322	-0.296544	0.7688
MONTCREDSOLLI	-0.273460	0.114625	-2.385691	0.0233
DUREEREM	0.222274	0.135152	1.644624	0.1102
TYPECREDIT	0.395921	0.224624	1.762591	0.0878
DEGRERAT	0.190603	0.160905	1.184570	0.2452
R-squared	0.322034	Meandependent var		0.325000
Adjusted R-squared	0.147075	S.D. dependent var		0.474342
S.E. of regression	0.438073	Akaike info criterion		1.382246
Sumsquaredresid	5.949152	Schwarz criterion		1.762244
Log likelihood	-18.64493	Hannan-Quinn criter.		1.519642
Durbin-Watson stat	1.861077			

ANNEX. IV. Elimination of the Variable Number of Credit Files and there-Estimation of the Equation

DependentVariable: RISQCREDIT				
Method: Least Squares				
Date: 07/29/19 Time: 21:27				
Sample: 1 40				
Included observations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
SEXE	0.126886	0.146873	0.863918	0.3941
MILIEUDERES	0.116494	0.186402	0.624960	0.5364
ZONEGEO	0.067668	0.051603	1.311329	0.1991
CSP	0.094739	0.055417	1.709561	0.0970
MONTCREDSOLLI	-0.274303	0.112945	-2.428643	0.0210
DUREEREM	0.214715	0.130821	1.641287	0.1105
TYPECREDIT	0.415985	0.211118	1.970389	0.0575
DEGRERAT	0.177909	0.152880	1.163717	0.2531
R-squared	0.320111	Meandependent var		0.325000
Adjusted R-squared	0.171385	S.D. dependent var		0.474342
S.E. of regression	0.431785	Akaike info criterion		1.335079
Sumsquaredresid	5.966028	Schwarz criterion		1.672855
Log likelihood	-18.70158	Hannan-Quinn criter.		1.457208
Durbin-Watson stat	1.846679			

ANNEX. V. Elimination of the Variable Place of Residence and the Re-Estimation of the Equation

DependentVariable: RISQCREDIT				
Method: Least Squares				
Date: 29/07/19 Time: 21:34				
Sample: 1 40				
Includedobservations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
SEXE	0.134660	0.144987	0.928771	0.3598
ZONEGEO	0.069575	0.051034	1.363297	0.1820
CSP	0.107824	0.050834	2.121104	0.0415
MONTCREDSOLLI	-0.308620	0.097783	-3.156157	0.0034
DUREEREM	0.248393	0.118100	2.103230	0.0432
TYPECREDIT	0.438383	0.206123	2.126799	0.0410
DEGRERAT	0.184605	0.151089	1.221827	0.2304
R-squared	0.311812	Meandependent var		0.325000
Adjusted R-squared	0.186687	S.D. dependent var		0.474342
S.E. of regression	0.427780	Akaike info criterion		1.297211
Sumsquaredresid	6.038847	Schwarz criterion		1.592765
Log likelihood	-18.94421	Hannan-Quinn criter.		1.404074
Durbin-Watson stat	1.899551			

ANNEX. VI. Elimination of the Variable Gender and the Re-Estimation of the Equation

DependentVariable: RISQCREDIT				
Method: Least Squares				
Date: 07/29/19 Time: 21:38				
Sample: 1 40				
Includedobservations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ZONEGEO	0.073479	0.050758	1.447632	0.1569
CSP	0.116151	0.049936	2.326006	0.0261
MONTCREDSOLLI	-0.337580	0.092491	-3.649873	0.0009
DUREEREM	0.294141	0.107121	2.745872	0.0096
TYPECREDIT	0.467207	0.203362	2.297416	0.0279
DEGRERAT	0.190543	0.150649	1.264817	0.2145
R-squared	0.293823	Meandependent var		0.325000
Adjusted R-squared	0.189974	S.D. dependent var		0.474342
S.E. of regression	0.426914	Akaike info criterion		1.273015
Sumsquaredresid	6.196701	Schwarz criterion		1.526347
Log likelihood	-19.46029	Hannan-Quinn criter.		1.364612
Durbin-Watson stat	1.773236			

ANNEX. VII. Elimination of the Variable Degree of Rationing and the Re-Estimation of the Equation

DependentVariable: RISQCREDIT				
Method: Least Squares				
Date: 07/29/19 Time: 21:42				
Sample: 1 40				
Included observations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
ZONEGEO	0.074019	0.051189	1.445979	0.1571
CSP	0.116358	0.050361	2.310457	0.0269
MONTCREDSOLLI	-0.364881	0.090704	-4.022743	0.0003
DUREEREM	0.347684	0.099241	3.503426	0.0013
TYPECREDIT	0.491187	0.204203	2.405380	0.0216
R-squared	0.260596	Meandependent var		0.325000
Adjusted R-squared	0.176093	S.D. dependent var		0.474342
S.E. of regression	0.430557	Akaike info criterion		1.268993
Sumsquaredresid	6.488267	Schwarz criterion		1.480103
Log likelihood	-20.37986	Hannan-Quinn criter.		1.345324
Durbin-Watson stat	1.779325			

Digital Twins Development Architectures and Deployment Technologies: Moroccan use Case

Mezzour Ghita[1], Benhadou Siham[2]
Engineering research laboratory (LRI), System Architecture Team (EAS) National and high school of electricity and mechanic (ENSEM) Hassan II University, Research Foundation for Development and Innovation in Science and Engineering, Casablanca, 8118, Morocco

Medromi Hicham[3]
Engineering research laboratory (LRI), System Architecture Team (EAS), National and high school of electricity and mechanic (ENSEM) Hassan II University Research Foundation for Development and Innovation in Science and Engineering, Casablanca, 8118, Morocco

***Abstract*—With the initiation of the fourth industrial revolution and the advent of information and communication technologies that reinforces the development of advanced technological solutions which engage data sciences, artificial intelligence, and cyber physical systems many long-established research concepts have been revived with in-depth applications within manufacturing plants. Thus currently the interest is turning more and more towards technologies and approaches that can combine between the virtual world and its increased capacities in computer sciences and processing, and the physical world with its complex systems and constantly evolving requirements. A relevant concept in this context is the concept of digital twins. Digital twins as defined by their founder Dr Michael Grieves are virtual replicas of a physical system that evolves within a virtual environment in order to mirror their real counterparts' life cycle and evolvement within the physical environment for applications in numerous domains. This paper's aim is to present a literature review of digital twin concept, its different development and deployment architectures, and its potential of application across Moroccan industrial ecosystems.**

Keywords—Digital twins; industry 4.0; digital twins challenges and opportunities; Moroccan industrial context

I. Introduction

On the threshold of the third industrial revolution that has enabled automation to cross the doors of industrial plants in different sectors, production systems are forced today to improve not only the diversity of their product, but also the flexibility of their processes [1]. This flexibility is now sought throughout the whole supply chain moreover throughout the whole project life cycle regardless of its nature or field of application. Today's industrial world is entering a new era that requires from simulation, modeling and data management processes to be flexible, intelligent and at a real time [2], which is a challenging quest for manufacturing operations management existing methods [3]. Some of the main challenges that encounter current manufacturing and management approaches on the weir of fourth industrial revolution are related to data acquisition and processing [4], systems design and manufacturing [5], Simulation processes [6], hosting networks and communication architectures [7], plants hardware

Infrastructures adaptability, modeling [8], agility to stakeholders requirements [10], and finally performances enhancement [9]. This fourth industrial revolution has taken into application within Morocco in particular and the whole world numerous concepts and advanced technologies such as artificial intelligence, big data analysis, machine learning and concerning simulation approaches digital twins [10]. Digital twin technology represents a huge advantage for production systems and process engineering throughout products life cycle [11]. A lot of companies around the world have integrated this technology within their supply chain. Today with the emergence of digitalization and its integration within national manufacturing development plan Moroccan industrial first founded ecosystems such as OCP ecosystem decided to take this international companies path through smart and connected factories [12]. However integrating these developed technologies within plants and production systems within Moroccan context comes to be extremely challenging and requires a structured and strongly founded road map and architecture development and deployment strategies [13]. This work was developed within this context. The first section of this paper presents a state of the art of digital twin concept development throughout literature review and industrial use cases; the second section introduces a comparative study of digital twin different existing architectures. The third section based on this literature review and the application feedback discuses digital twin deployment within Moroccan industrial context. and the fourth section illustrates an application of one digital twins architectures for the development of a turbo alternator group digital twin Finally the last section summarizes the whole paper and presents future works and perspectives.

II. State of the Art

A. A Literature Review of Concept Development

The concept of digital twins was first introduced by Michael Grieves in one of his presentation about Product Life Cycle Management in 2003 at the University of Michigan [13], he defined it as a virtual equivalent or a replica of the physical product that requires three main parts to be established a physical product in a physical space, a virtual product in a virtual space and finally the connection between them [14]. Before introducing the term Digital Twin , the concept was referred to by Grieves as the Information Mirroring Model [11]. The National Aeronautics and Space Administration NASA used for years twining concept for its Apollo program in order to manage maintenance engineering operations and tests. During the design phase two replicas of the system were developed, the first one was dedicated to tests and simulations after and during the mission whereas the second one was especially dedicated to the mission. In 2010 NASA gave the definition of digital twins for predictive maintenance of space vehicles. It was introduced as a vehicles multidisciplinary simulation that make use of operational data , sensors data , functioning histories and systems physical models in order to manage space vehicles health. In 2011 the US Air Force used digital twin concept and technologies in order to management aircraft structural life prediction process [15]. In 2012 US Air force in collaboration with NASA published a paper that represented the results of a study which addressed the digital twin as a performant alternative to traditional simulation techniques for vehicles safety and reliability [16]. The aerospace industry was the first industry to introduce digital twins into its processes and products development and life cycle management through numerous industrial applications , and then other industries and fields for example automotive industry [17], integration of digital twins for real time monitoring of fleet management systems [18], urban mobility and sustainable development within smart cities [19], [20], and last but not least medicine and health care started developing and studying this technology challenges and advantages for their performances enhancement [21], [22]. Digital twin modules development and applications enhancement through literature and industry were interconnected with the development of information technologies IT and communication means and complex systems management and engineering [23]. Recently research community came with an extended four dimension vision of Grieves initial digital twin framework development [24]; this vision included four main components that are models, data, connections and services. The first dimension which is models refers to the different conceptual models that draw the first lines of the digital twin design. Within advanced simulation process mainly development project of digital twin, modeling plays a primordial role, in the sense that it's the starting point for launching a simulation. Modeling enables simulator and project stakeholders to understand the complex system their working on, its behaviors, states and interactions with its environments. The first step for establishing a simulation program is the modeling of real system components and driving functions .Modeling create a bridge between the real system its different parameters, constraints, data and the simulation program [25] simulation process start with an elaborated study of the real system, the establishment of its conceptual model, then development of simulation program and finally analysis of simulation results. The second dimension shed light on data of the digital twin which is a crucial aspect that enables digital twin not only to interact with its virtual space but also with the physical asset and its physical space, with project stakeholders and finally with external environment changes and constraints. Digital twin can be alimented with data from the physical systems such as functioning histories, field real time measurements and parameters performances, with external decisions imposed by stakeholders and due to the integration of machine learning to its personal knowledge base and memory [26]. The third-dimension concerns communication technologies and protocols between systems, spaces and models. Notable effort has been done in the recent years concerning this aspect in order to improve real-time monitoring of equipments and assets, data acquisition and processing, and last but not least models interoperability and simulation synchronization time [27]. The last dimension within the new proposed vision is services intended from the established digital twin development and deployment project. Services refers to digital twin features in order to meet stakeholders and context requirements [28]. Digital twin has proved through its different industrial applications its efficiency in handling a multitude of functions such as product health management (PHM), state prediction, energy consumption optimization [29], real time monitoring and so many others applications within different sectors that will be detailed in further parts.

B. Digital Twin Development and Deployment Architectures within Industry 4.0

Throughout literature and different industrial use cases a multitude of architectures had been proposed in order to create a framework for digital twins' development and implementation. These architectures varied depending on digital twins' field of applications, intended services and benefits and related technologies and concepts. This part focuses on the analysis of these architectures background, key concepts, challenges, opportunities and cyber physical data lifecycle management. The analysis of the different proposed architectures and the review of reference architectures for intelligent and autonomous systems [26] enabled us to conclude on a common reference representation for digital twin development and implementation architectures which is based on three layers. The first layer consists on industrial layer that includes field complex physical systems which can be a production line, an equipment or the whole industrial installation. This layer composition depends strongly on digital twin scope and physical twin environment. The second layer is application layer that focuses on digital surrogates' main cyber components and features. Finally the third layer is communication layer that enables cyber physical interactions. Based on our literature review and our analysis of a large number of application use cases in various fields, we have tried to nourish these different layers and to conclude on different existing digital twins' implementation and development architectures. Fig. 1 summarizes the proposed digital twin reference architecture and figure 2,3,4,5,6 and 7 put into effect this representation within the fourth industrial ecosystem.

1) Development architecture based on the fusion of digital twin concepts and big data principles: The fusion between big data and digital twin can result on the development of cyber physical data that can allow smart manufacturing approaches and benefits from highly developed big data techniques to improve and enlarge digital twin applications and services [27]. Some of the advantages that digital twin can brings to big data field concerning data variability is digital twin ability to generate new information's through virtual models and real time connection to operational and legacy data. In the sense that surrogates virtual models can through their interactions with the physical space and real systems performances generate a variety of predictive models directly related to the virtual twin. These predictive models can help maintenance staff to detect root causes of some failures and prevent some complex systems unpredictable behaviors [21]. The other way round big data provides digital twins with some advanced data cleaning, mining and analysis techniques that can enable the improvement of product life cycle management [22]; asset health management [28], production planning and other manufacturing critical operations that will be discussed further in this paper. Fig. 2 concretizes this vision resulting architecture.

2) Development architecture based on Augmented Reality (AR) and Virtual Reality (VR): When combined with digital twin basic concepts augmented reality can bring a lot of advantages for different domains mainly predictive maintenance, systems design and operators training. Augmented and virtual realities are both concepts that establish continuous connections between physical and virtual world but also dynamic interactions between complex systems stakeholders, components and virtual twins. The use of augmented reality for digital twins deployment and exploitation within factories enables the application of numerous complex and critical tasks within the physical space with virtual assets. The fusion of digital twins and augmented reality was applied for numerous applications in the industrial field. For example maintenance operations management [29], distant process control, and within education field for attaining learning objectives [30]. These large set of different and diversified applications had proven the efficiency of this combination. Fig. 3 represents the proposed architecture for this amalgamation.

3) Devellopement architecture based on Industrial Internet of Thing concepts (IIoT): Digital twin and industrial internet of things (IIOT) combination can presents some notable benefits for cyber physical systems development framework [31]. Industrial IoT enables systems large scale connection as well as external connections [32]. IIOT advanced communication capabilities are crucial for unleashing digital twin solutions potential. On the one hand Industrial IoT architectures encourages exchange between different levels and elements its enable reliable and safe interactions with physical smarts objects and systems. On the other hand digital twin concept through exploiting data gathered with IoT gateways and components can give ergonomic virtual representations for users and offers for systems data analytics smart, embedded, and performant algorithms and tools [33]. Fig. 4 represents the different layers of the proposed architecture within this framework.

4) Development architecture based on complex systems engineering: The emergence of industry 4.0 concepts and visions has bring a lot of new requirements and functions to engineering and manufacturing[34]. With the spread of new paradigms such as cyber physical systems, digital twins that is a key concept for virtual and physical world fusion has gained a lot of interests [30]. The development and implementation of digital twins within factories despite the different technologies offered by advanced technologies and solutions vendors encountered a lot of challenges concerning digital twin effective deployment within manufactories [31] hence comes the importance of developing an organized and well founded framework for digital twins. Complex systems engineering has been a key tool for complex physical systems development [35]. A lot of methods and languages had been developed in order to exploit this method effectively. These tools offer a modeling and simulation dynamic framework for digital twins One of these languages is Systems Modeling Language SYSML that offers numerous diagrams for digital twin functional and behavioral development and that have been used in some digital twins applications [32]. By using complex system engineering digital twin external and internal stakeholders will interact actively in order to develop an autonomous and performant system that fits their technical and non technical needs and requirements [13]. Fig. 5 illustrates the architecture resulting from this collaboration.

5) Development architecture based on cloud services: Cloud computing and cloud services has gained a lot of popularity among industrials from different sectors lately [33]. Today cloud platforms provide in addition to storage and computing services a lot of others features that can foster manufactories management operations [36]. Lately digital surrogates development and deployment has been added to this broad portfolio of services [28]. The main idea behind this contribution from Oracle point of view for example is to ensure better visibility, accurate prediction, efficient documentation, real time communication management and last but not least integration of disparate heterogeneous systems [37]. A lot of cloud computing services vendors had proposed architectures for digital twin development and deployment such as Amazon, Microsoft Azure and Oracle. These architectures were developed with the combination of cloud computing, IoT concepts as well as augmented Reality for some use cases [38]. For example Amazon through AWS IoT Core offers assets shadows service [39]. The shadow service enables users to connect their devices to the cloud and create their devices twins within a dedicated platform in Amazon cloud. The shadow will record physical devices status on a JSON format available for consultation and shares it for analysis with other cloud services and clients. The service gives client the ability to configure different shadows parameters such as access, documentation, updates, desired states and many other options. Fig. 6 concretizes this vision.

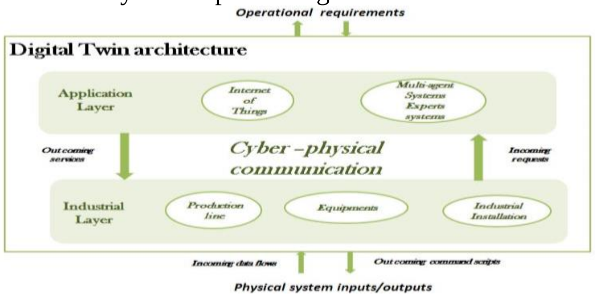


Fig. 1. Digital Twin Development and Deployment Architecture Layer.

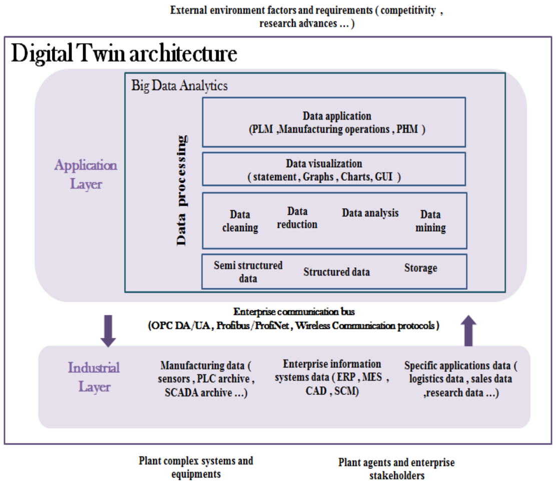

External environment factors and requirements (competitivity ,
research advances ...)
Digital Twin architecture
Application Layer
Big Data Analytics
Data processing
Data application
(PLM ,Manufacturing operations , PHM)
Data visualization
(statement , Graphs , Charts, GUI)
Data cleaning
Data reduction
Data analysis
Data mining
Semi structured data
Structured data
Storage
Enterprise communication bus
(OPC DA/UA , Profibus / ProfiNet , Wireless Communication protocols)
Industrial Layer
Manufacturing data (sensors , PLC archive , SCADA archive ...)
Enterprise information systems data (ERP , MES , CAD , SCM)
Specific applications data (logistics data , sales data ,research data ...)
Plant complex systems and equipments
Plant agents and enterprise stakeholders

Fig. 2. Digital Twin Architecture based on Big Data.

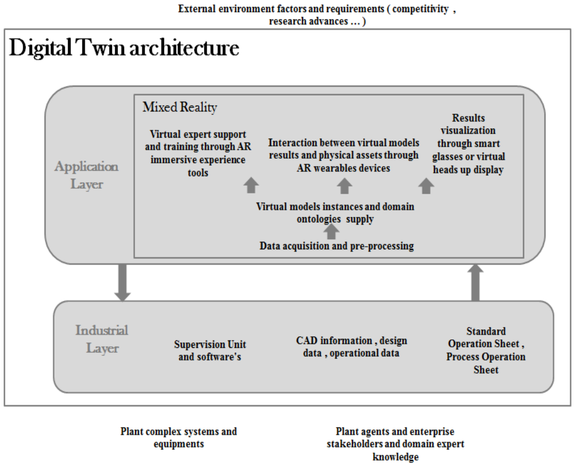

External environment factors and requirements (competitivity ,
research advances ...)
Digital Twin architecture
Application Layer
Mixed Reality
Virtual expert support and training through AR immersive experience tools
Interaction between virtual models results and physical assets through AR wearables devices
Results visualization through smart glasses or virtual heads up display
Virtual models instances and domain ontologies supply
Data acquisition and pre-processing
Industrial Layer
Supervision Unit and software's
CAD information , design data , operational data
Standard Operation Sheet , Process Operation Sheet
Plant complex systems and equipments
Plant agents and enterprise stakeholders and domain expert knowledge

Fig. 3. Digital Twin Architecture based on Mixed Reality.

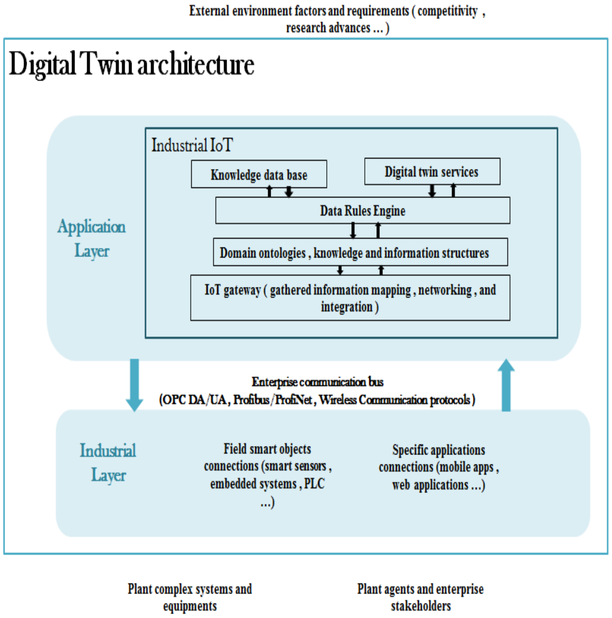

External environment factors and requirements (competitivity ,
research advances ...)
Digital Twin architecture
Application Layer
Industrial IoT
Knowledge data base
Digital twin services
Data Rules Engine
Domain ontologies , knowledge and information structures
IoT gateway (gathered information mapping , networking , and integration)
Enterprise communication bus
(OPC DA/UA , Profibus / ProfiNet , Wireless Communication protocols)
Industrial Layer
Field smart objects connections (smart sensors , embedded systems , PLC ...)
Specific applications connections (mobile apps , web applications ...)
Plant complex systems and equipments
Plant agents and enterprise stakeholders

Fig. 4. Digital Twins Architecture based on Idustrial Internet of Things.

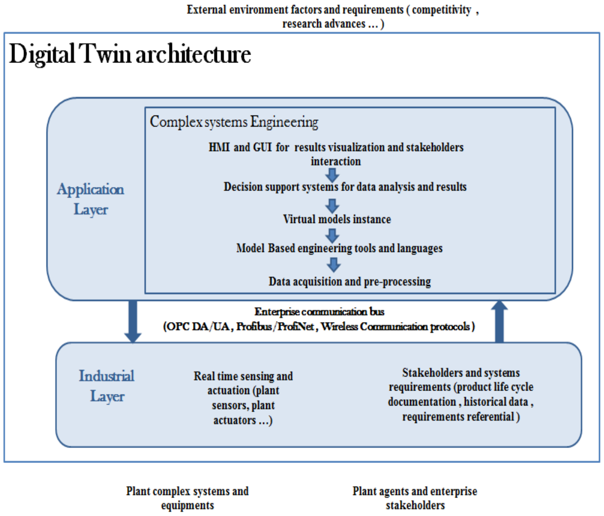

External environment factors and requirements (competitivity ,
research advances ...)
Digital Twin architecture
Application Layer
Complex systems Engineering
HMI and GUI for results visualization and stakeholders interaction
Decision support systems for data analysis and results
Virtual models instance
Model Based engineering tools and languages
Data acquisition and pre-processing
Enterprise communication bus
(OPC DA/UA , Profibus / ProfiNet , Wireless Communication protocols)
Industrial Layer
Real time sensing and actuation (plant sensors, plant actuators ...)
Stakeholders and systems requirements (product life cycle documentation , historical data , requirements referential)
Plant complex systems and equipments
Plant agents and enterprise stakeholders

Fig. 5. Digital Twin Architecture based on Complex System Engineering.

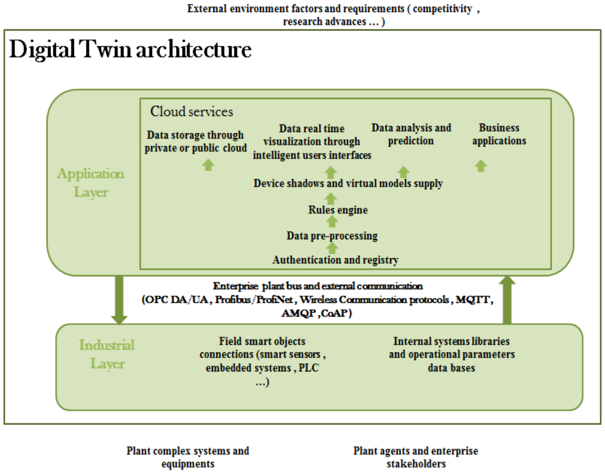

External environment factors and requirements (competitivity ,
research advances ...)
Digital Twin architecture
Application Layer
Cloud services
Data storage through private or public cloud
Data real time visualization through intelligent users interfaces
Data analysis and prediction
Business applications
Device shadows and virtual models supply
Rules engine
Data pre-processing
Authentication and registry
Enterprise plant bus and external communication
(OPC DA/UA , Profibus / ProfiNet , Wireless Communication protocols , MQTT , AMQP ,CoAP)
Industrial Layer
Field smart objects connections (smart sensors , embedded systems , PLC ...)
Internal systems libraries and operational parameters data bases
Plant complex systems and equipments
Plant agents and enterprise stakeholders

Fig. 6. Digital Twin Architecture based on Cloud Services.

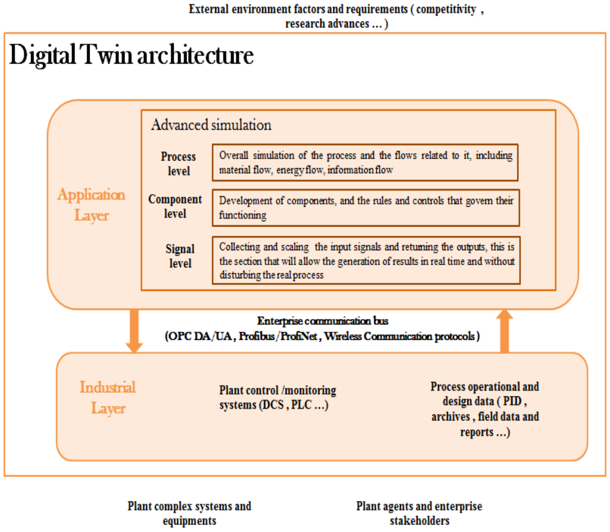

External environment factors and requirements (competitivity ,
research advances ...)
Digital Twin architecture
Application Layer
Advanced simulation
Process level
Overall simulation of the process and the flows related to it, including material flow, energy flow, information flow
Component level
Development of components, and the rules and controls that govern their functioning
Signal level
Collecting and scaling the input signals and returning the outputs, this is the section that will allow the generation of results in real time and without disturbing the real process
Enterprise communication bus
(OPC DA/UA , Profibus / ProfiNet , Wireless Communication protocols)
Industrial Layer
Plant control /monitoring systems (DCS , PLC ...)
Process operational and design data (PID , archives , field data and reports ...)
Plant complex systems and equipments
Plant agents and enterprise stakeholders

Fig. 7. Digital Twins Architecture based on Advanced Simulation Techniques.

6) Development architecture based on advanced simulation techniques: A lot of research organism such as NASA and US Air force exploited digital twin concept with the use of simulation for manufacturing process and prototyping operations enhancement, predictive maintenance, product life cycle management [14]. Through its combination with existing simulation tools or by developing a whole new framework such as Digital Thread concept as proposed by US Air Force [40]. This new simulation framework will be based on knowledge management, vertical and horizontal integration and the integration of all interactions with external ecosystem [41]. Currently a lot of simulation software vendors provide technologies that are founded on the basis of digital twin concept [32]. Whether it's through simulation software's, other simulation tools and their combination with data management tools, modeling techniques and information and communication techniques digital twin represents numerous opportunities for real time simulation techniques development and deployment within manufactories [42]. Fig. 7 illustrates digital twin architecture based on advanced simulation techniques and technologies.

C. Digital Twin Development and Deployment Technologies

There are numerous companies in different fields that developed and deployed digital twin concept through their manufactories in order to improve their performances. According to their core business, field of applications and services intended from the digital twin each of these companies gave it a particular definition. For example the first sector concerns use cases of companies that operate in the field of electric power generation such as Siemens [34] [35], Honeywell [36], Shneider Electric [37], General Electric GE [38],[39], the second ones are companies that offer digital and IoT services like Parametric Technology Corporation PTC [40] ,Amazon, IBM, DELL and the third one automotive entreprises like PACCAR, and aviation industry such as GE Aviation Systems LLC [41]. In addition to these technologies a lot of patents [42] and international reports concerning digital twins concept effective deployment and their application within manufactories and other fields were published in recent years [43], [44]. The thing that once again reinforces their position as en efficient tool and a promising concept for both research and industry.

D. Digital Twin Development and Implementation Related Standards

In recent years many standards have been developed to support the deployment and implementation of industry 4.0 and cyber and physical systems connections new technologies and tools such as digital twins[[45],[46],[47]]. The literature review in reference to the different architectures founded for the implementation of digital twins has enabled us to identify a number of significant standards and referential that relate to each of the layers of the architectures proposed. All these standards are not specific to digital twins' global development framework and do only address one of the aspects connected to the digital twins adopted architectures. In 2018 the Technical Committee ISO/TC 184/SC 4 proposed a standard development project specifically dedicated to the framework of digital twin development and implementation, this standard is currently being elaborated and its appearance is foreseen for the next few year [48]. Some of these standards are listed below and will be detailed furthermore. Standards description is divided into four categories industry 4.0 global framework standards; industrial layer related standards; application layer related standards and finally physical and digital communication related standards. This classification is inspired by the Reference Architecture Model for industry 4.0 (RAMI4.0) proposed segmentation for architecture 4.0 framework [49] and International standards Organization (ISO) International Classification for Standards (ICS) 35.240.35 standards family for IT application in industry. The first category deals with global framework standards firstly standards that concerns smart factory implementation guide lines and requirements it is introduced by IEC 62794 that defines a reference model for digital factory development and integration and IEC 62832 that highlights the general principles fir digital factory pillars integration through the automation architecture. And finally RAMI 4.0 that is a reference guide line for industry 4.0 [50].

Secondly this category includes standards on digital twin ecosystem global security development [31] mainly ISO 27000 families for information systems layers security and risks management , DIN SPEC that details IT security techniques based on domain expert definition, IEC 61850 for automation systems advanced protection and finally IEC 62443 for industrial automation and control systems security [51]. Industrial layer related standards list a number of standards that relates to raw data structure and exploitation references techniques, smart asset and device management under 4.0 architectures and finally knowledge management requirements throughout different affiliated norms respectively ISO/TS 18876-2:2003, ISO/IEC 30101:2014; IEC 62264 and ISO 30401:2018.Application layer introduces standards related to the different concepts and architectures for digital twin development mainly big data-based architecture through ISO/TEC 19395 that define best practices for data center resources monitoring and control and ISO/IEC WD 27045 for security concerns within big data field which is a critical element for digital twins' deployment within industrial plants. Industrial IoT based architecture introduces main standards of IoT implementation and global concepts for its understanding through ISO/IEC TR 22417:2017, IEEE P241, ISO/IEC NP 30166 and finally ISO/IEC 21823-1:2019. This parts also shed light on an important component of digital twin implementation that is visualization and which can be apprehended through human machine interaction standards and ergonomic standards.

Communication layer which is a crucial element for digital twin efficient exploitation introduces communication standards related to virtual and physical connection [27] mainly DIN SPEC that gives guide lines for AML and OPCUA integration, IEC 62601:2015 for wireless communication within industrial plants and RESTFUL -API related standards and referential as its one of the most used communication protocol for physical and virtual layers connection .A lot digital twins applications across literature and for some propose digital twins technologies used OPC UA and Automation ML [52].

III. Synthesis and Comparison

A. Synthesis

All over the literature and the various industrial use cases discussed along the previous section, a plethora of architectures have been proposed to create a framework for the development and implementation of digital twins, depending on the scope of application of digital twins, the services and benefits intended and the architectures associated technologies and concepts. This section summarizes the results of the analysis conducted on the main foundations of each of the architectures detailed in the previous section. Digital transformation has brought a lot of concepts to application within automation-based plants these concepts evolvement can contribute a lot for digital twins' deployment and development. Throughout their main components and key technologies that highlights the major impact of information and communication technologies and computer sciences on automation and industries performances improvement and efficiency. The focus on providing ecological, performant but less expensive technological solutions when it comes to cloud services, big data or IoT can be exploited to create and propose innovative digital twins' architectures and technologies.

B. Architectures Comparasion

Throughout literary review of digital twins we were able to identify a number of factors that occur in the process of adapting one of the architectures mentioned above to efficiently exploit the potential of digital twin concept [53]. Thus we distinguished two categories of factors for architecture suitability assessment. The first category concerned factors related to the contextual adaptability of the architecture in order to develop context aware digital twins. And the second category is constituted by factors related to architectures functional suitability [54] once the twin in question is deployed. The first category that concerns contextual adaptability includes five elements, multi-agent interoperability, project management, energy management, security management, and finally data life cycle management. The second category that is concerned with functional adaptability and it includes digital twin implementation architectures hardware and software layers functional suitability evaluation. This differentiation has many advantages for a digital twin development and deployment project, particularly in an evolving industrial context that is characterized by several requirements on different levels and that needs agile management methods [55],[56].

IV. Digital Twins Development and Implementation Architecture –Moroccan use Case

A. Opportunities and Threats

Morocco is currently in the midst of a revolution to integrate digital transformation in several sectors, primarily education, public administration and the industrial ecosystem through its actors[57]. This transition requires the collaboration of several entities, including not only research institutions but also the internal entities of the industrial tissue [58].

A crucial element for the integration of industry 4.0 concepts of which digital twins [59], and particularly to the Moroccan industrial context are change management strategies. Change management strategies should not only affect the internal structure of companies but also training systems and management practices at the strategic and operational levels of the country. Many studies in recent years have been conducted for the development of these aspects within Morocco with regards to the integration of industry 4.0 and its elements such as artificial intelligence, big data ,cloud computing [10] more essentially smart cities [60] and their numerous applications and challenges [61]. Inspired by these studies and internal and external reports on 4.0 technologies [57],[58], [62] and the implementation of digital twins at the international level, we were able to gather a set of strengths and weaknesses related to Moroccan context and opportunities and threats related to the external context. This analysis concerned the adaptability of current context for implementing the various digital twin architectures concluded from the state of the art on concept development and its relationship with the evolvement of industry 4.0. This study has allowed us to identify some major development axes for the integration of digital twins within Morocco that will be further strengthened through several other context studies and surveys. The various strengths and opportunities that we have been able to conclude offered us a basis for the development of a high-performing digital twin architecture that could use experience feedback from international experiences on this subject to develop an adaptability mechanism specific to the Moroccan context Thus this mechanism main role is to link between computer centric development methods and human centric development methods to propose a generic framework for the development of context aware digital twins. Table I summarizes the results of SWOT analysis. These analyses were established in order to assess the potential of existing digital twins' development and deployment architectures.

B. Application use Case

1) Global context: This project is a part of OCP‘s Mining Digitization Initiative. since the emergence of industry 4.0 technologies OCP have been known as one of the first Moroccan enterprises that tried to integrate these technologies within their plants [12]. Digital twin's technology that is one of the most spread industry 4.0 technologies offers numerous applications and features for operation management improvement, advanced process control and collaborators training techniques evolvement. Thus, it considered as one of the pillars for industrial field digital transformation from installation and processes automation to the development of industrial plants cyber physical systems. The project first use case presented through this paper consisted on the development of a digital twin for the turbo alternator group of the thermal power plant in OCP‘s Jorf fertilizer Company 5 (JFC5) utilities facility with the collaboration of the different stakeholders. The main purpose of this project was the development and improvement of the group performances and the development of a simulation platform 4.0 improvement of the group performances and the development of a simulation platform 4.0 specific to the OCP group, throughout refining industrials needs and evaluating the proposed digital twin technology adaptability to context requirements.

TABLE. I. SWOT ANALYSIS FOR DIGITAL TWINS ARCHITECTURES DEPLOYMENT

SWOT analysis	**Strengths**	**Weaknesses**	**Opportunities**	**Threats**
Development architecture Based on the fusion of twin digital concepts and Big data management	There is a great interest in data management with all its aspects Availability of a solid and developed software and hardware infrastructure Deployment of several research projects on the development of internal knowledge management Feedback from internal application cases on problems encountered	Confidentiality requirements Delays in the time required to implement solutions related to the architecture Challenges of interoperability with existing data management and analysis systems	Publication of several standards structuring the development of this type of architectures Numerous and diverse bibliographic sources and application cases Several contextual parameters and requirements such as hardware energy consumption and software security are already taken into account	Cyber security risks National regulations and standards
Development architecture based on Industrial Internet of Things	Digitization is a strategic national priority The initiation of different knowledge capitalization and data management projects -Feedback and a microscopic and macroscopic view of the landscape within morocco Networking infrastructure in progress The availability of advanced ICTs	Difficulties of integration -Need to carry out several risk studies to assess the potential of the solution, in the field of security in in particular High short-term costs Resilience to change Interoperability with legacy systems and Ergonomics concerns Hostile implementation environments	Field expert facility implementation within Morocco -Partnership with several research organizations dedicated to the development of IIoT solutions Accessible experience feedback Development of the IIoT solutions standards framework Field in continuous evolution in Morocco and competence development within the sector	Dependency towards solution providers' Insufficient regulation and national guidelines Security-related risks Accessibility risks (network, equipment, etc.) Worldwide rapid growth
Development and deployment architecture Based on Cloud Computing CC and Cloud Services CS	High energy saving compared to other technologies especially for large enterprises that needs a lot of computing and storage resources Flexibility and portability Reducing hardware costs and implementation delays	Increased security requirements and favoring the use of private cloud technologies High dependency to cloud provider Lack of national regulation directly related to cloud computing activities management Reticence with regard to data sharing Insufficient mastery of the concept Needs of high performance networks and communication technologies	National awareness of the potential of the cloud and its services Worldwide regulation and consolidation of the cloud computing industry Existence of national feedback and internal case study on cloud exploitation Approval of several standards projects on cloud computing	Cybersecurity concerns High dependency to cloud provider Lack of national regulation directly related to cloud computing activities management
Development and deployment architecture based on advanced simulation techniques	Structured control and monitoring architecture Maturity 3.0 and automation under development Control and monitoring systems open to modifications High-performance field communication protocol Availability and accessibility to the expertise of different suppliers Development of advanced automation data management	Difficulty of adaptability and interoperability with legacy control command systems Difficulty in accessing the histories of existing data Complex systems and modelling-related difficulties Adaptability and mastering challenges	Domain in a state of very high speed evolution Development of high-performance computing and storage tools -Variety of proposed technologies and extension to several cases of industrial environments	Security risks with high levels of criticality Reticence to data sharing Energy consumption constraints Compliance with regulations
Development and deployment architecture Based on Complex systems engineering	Multidisciplinary and constantly collaborating engineering and research teams to explore advances related to the different core businesses of companies Low investment required for the hardware layer to implement the solution Development of the knowledge management component and data management techniques within companies Capitalization of expertise and communication is becoming one of the strategic objectives of the country recently	Difficulty in modeling and uniformizing models for large structures based on automation layers provided by different vendors Confidentiality and IT security constraints Reluctance towards public data sharing Resiliency to change Difficulty in capitalizing knowledge and accessing data for existing installations Adaptability of internal infrastructure Long product lines development times	Availability of scientific and industrial feedback -Development of Model Based Engineering and its tools National Digitalization Strategy Development of a high-performance open source platforms for data modeling and visualization Training increasingly oriented towards the use of advanced modelling and simulation tools Development of network and telecommunications infrastructure throughout the country's provinces	Difficulty in keeping up with developments in the field at the international level Conformity with national regulations Changeable market requirements and needs

TABLE. II. ARCHITECTURES COMPARATIVE STUDY

Criteria	Sub criteria	Factors for assessment	A1	A2	A3	A4	A5	A6
Context adaptability	Energy management	Hardware and software energy consumption regards	🟧	🟩	🟥	🟧	🟩	🟩
		Compliance to energy management strategies within the company	🟧	🟩	🟧	🟧	🟩	🟩
		Energy management indicators and possible assessment	🟩	🟩	🟧	🟧	🟩	🟩
	Data management	Compliance with existing data structures and representations (supported forms, supported representations, supported data volume)	🟥	🟧	🟧	🟩	🟩	🟧
		Compliance with national data sharing process regulations and laws	🟧	🟧	🟧	🟩	🟩	🟧
		Performant data acquisition and preprocessing platforms	🟩	🟩	🟩	🟩	🟧	🟧
		Data modeling (unified semantics, unified interfaces)	🟧	🟧	🟧	🟩	🟧	🟧
	Multi-Agent interoperability	Legacy systems interoperability	🟥	🟧	🟧	🟩	🟧	🟥
		Agents and collaborators access	🟥	🟧	🟧	🟩	🟧	🟧
		Solution usability and ergonomic	🟧	🟧	🟩	🟩	🟩	🟧
		Solution trustworthiness and technology freedom of context risks	🟥	🟥	🟧	🟩	🟧	🟥
	Security management	Compliance of industrial layer security management mechanisms with project stakeholders' requirements and solution potential risks within the project context	🟥	🟥	🟧	🟩	🟧	🟥
		Compliance of communication layer security management mechanisms with project stakeholders' requirements and solution potential risks within the project context	🟥	🟥	🟧	🟧	🟩	
		Compliance of application layer security management mechanisms with project stakeholders' requirements and solution potential risks within the project context	🟧	🟧	🟧	🟩	🟩	🟧
	Project management	Human capital management	🟧	🟧	🟧	🟩	🟩	🟧
		Project costs and budget management	🟥	🟧	🟥	🟩	🟩	🟧
	Conformity management	Existence of international and national feedback, references and standards for the solution scope	🟧	🟩	🟩	🟩		🟩
		Existence of certified use cases on technology exploitation and deployment	🟧	🟧	🟧	🟧	🟩	🟧
		Compliance with national regulation and standards	🟧	🟧	🟧	🟩	🟩	🟧
		Compliance with enterprise internal standards	🟧	🟧	🟧	🟩	🟩	🟧
Functional suitability	Hardware suitability	Compliance with existing control command architecture	🟧	🟧	🟧	🟩	🟧	🟧
		Compliance with plant communication infrastructure	🟧	🟧	🟧	🟩	🟩	🟧
		Independency from vendors possibility	🟥	🟧	🟧	🟧		🟥
		Compliance with plant hostile environments (data acquisition hardware immunity	🟥	🟩	🟩	🟩	🟩	🟥
		Hardware implementation delays and costs	🟥	🟧	🟧	🟩	🟩	🟥
	Software functional suitability	Software appropriateness with regards to project stakeholders' requirements	🟧	🟩	🟩	🟧	🟧	🟧
		Software correctness with regards to project stakeholders' requirements	🟧	🟩	🟩	🟩	🟧	🟧
		Software completeness with regards to project stakeholders' requirements	🟧	🟧	🟧	🟩	🟩	🟧
		Software portability	🟧	🟩	🟩	🟩	🟩	🟩
		Software efficiency	🟧	🟧	🟧	🟧	🟩	🟧
		Software compatibility with legacy systems	🟥	🟥	🟧	🟩	🟧	🟥
		Software maintainability	🟥	🟧	🟧	🟩	🟩	🟧

2) Project stakeholders research organization: In order to create a bridge between research and industry, research organizations mobilize groups of engineers from different specialties who have acquired during their studies a set of basic concepts allowing them to easily integrate both industrial and research environments. The main purpose of these groups is to search for advances within enterprises core business context the mining sector and their principal mission is to understand the new proposed technologies and practices in order to improve them and adapt them to the context of their application. Thus, for this stakeholder, the project will serve as a springboard between research and development.

Industrial in the mining sector: For the industrial sector, the main objective is to improve its processes and consequently the quality of its supply chain and, in a more global vision, its competitiveness in terms of its DAS. To achieve this objective, it is essential to stay up to date with new technologies that have the same purpose. Supplier of simulation technologies and engineering team: Today on the market a wide range of companies offers several simulation technologies and software to meet industrialist needs. Supplier's and internal engineering team goal is to meet specific needs and requirements in terms of performances and creativity.

C. Turbo Alternator Group Digital Twin

1) Architecture selection and technlogy adaptability: The choice of digital twin development and deployment architecture was strongly related to context constraints, stakeholders' requirements and technology functional suitability. Table II represents the results of the six proposed digital twin architectures selection based on a set of criteria that were developed in collaboration with project engineering team and based on users initial set of digital twin functional and operational requirements and context adaptability factors. The first architecture (A1) evaluated through the criteria defined is IIOT based architecture, the second one (A2) is big data based architecture, the third one (A3) is mixed reality based architecture, the fourth one (A4) is simulation based architecture, then (A5) complex systems based architecture and finally the last one (A6) is cloud computing based architecture. The evaluation of the different criteria for the six architectures proposed was based on context parameters and stakeholders' requirements towards digital twin solution development and deployment. Criterion evaluation was through three levels the first level represents high compliance with required criterion, the second level is medium compliance; and the last level represents low or non-existent compliance.

2) Digital twin development and implementation: Based on the different models developed through the different foundations defined and operational data preprocessing we were able to supply SIMIT simulation three levels, signal level, component level and process level [63]. The simulation results and test conducted through the three levels enabled us to verify and validate the twin virtual components models. The models developed will be shaped through the record-keeping of the obtained simulation results and their correlation with the actual installation properties. This ongoing process, which links the digital twin to his real compatriot, will enable him to create his own training data set. Fig. 8 represents the twin developed user interface for simulation and tests. Fig. 9 introduces foundations for turbo alternator digital twin development based on architectures comparison results.

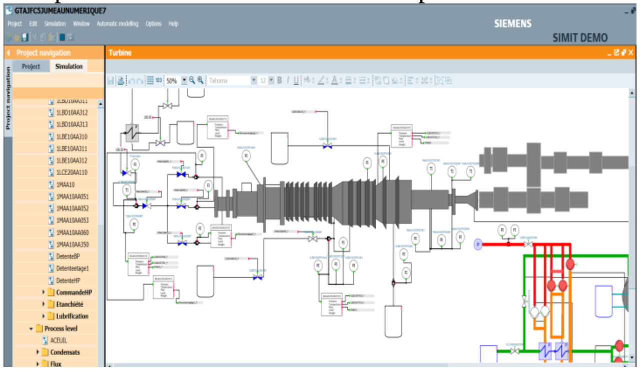


Fig. 8. Digital Twin Process Simulation and Tests Interface.

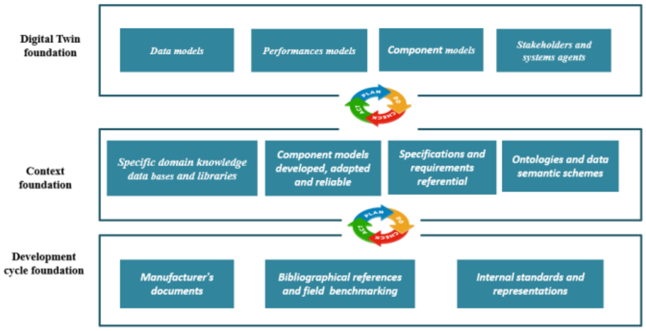

Digital Twin foundation
Data models
Performances models
Component models
Stakeholders and systems agents
PLAN
DO
CHECK
ACT
Context foundation
Specific domain knowledge data bases and libraries
Component models developed, adapted and reliable
Specifications and requirements referential
Ontologies and data semantic schemes
PLAN
DO
CHECK
ACT
Development cycle foundation
Manufacturer's documents
Bibliographical references and field benchmarking
Internal standards and representations

Fig. 9. Digital Twin Development Foundations.

V. Concluding Remarks and Future

Through this paper and our literature review on the field we tried to give an overview of the roots of digital twin concept, its different fields of application, its development and deployment architectures, and the technologies offered by different market vendors that mainly consist on the exploitation of the concept foundations. The use case of application we worked on enabled us to detect some of the limitations and challenges that can hinder digital twins deployment despite the large set of opportunities it represents for the development of a lot of research axes for example predictive maintenance and real time monitoring. These limitations that were highlighted previously by architectures comparison and SWOT analysis concerned some of the main problematics of new technologies implementation within manufacturing plants. Mainly these problematics concern hardware infrastructures maturity and their adjustment costs and delays to meet the technical requirements of new technologies, modeling challenges, cyber security, interoperability and last but not least human capital competencies. In order to ensure these advanced technologies deployment industrialists have to afford a communication bridge with different research communities that based on experience feedback can work on scalable and generic solutions that can practically encompass these identified obstacles. It's crystal clear that in the last decade due to the emergence of a lot of new information and communication technologies and the evolvement of industry 4.0 vision, the interest towards digital twins within both research and industrial eras has significantly arise. The new shift towards cyber physical systems has resulted on the emergence of a new technological decade that makes computer sciences, communication technologies and cognitive sciences in the focus of numerous manufacturing ecosystems. The great impact that artificial intelligence has on the evolvement of a lot sectors that are related to industrial fields as well as other social aspects makes us think about the opportunities that can be offered by the combination of digital twin concepts with artificial intelligence areas like machine learning or deep learning so as to support its efficient implementation . Currently some of the technologies developed by digital services vendors like Microsoft or IBM are exploiting the potential of artificial intelligence for applications within smart factories, and digital twins' development. Our future works will focus on the exploration of this aspect with relationship to existing applications and potential opportunities offered to digital twin concept within Moroccan context that highlighted a lot of problematics related to the concept deployment as mentioned previously. Throughout surveys conducted with industrialists from different sectors and the exchange of scientific, industrial and normative experience feedback, we intend to develop an evaluation framework for digital twins. This framework through existing applications as well as gathered good practices and procedures learnt on the subject, will allow us to propose a holistic architecture for the development of digital twins that can overcome the previously identified limitations but also benefit from the technological advances of Industry 4.0 and researches conducted for its effective implementation not only as a strategic vision but also as an action plan for accelerating the performances of manufacturing plants.

Acknowledgment

This research was supported by Mohammed VI Polytechnic University de Benguerir, Morocco and OCP group. We thank our colleagues from university who provided insight and expertise that greatly assisted the research.

References

[1] M. Sony, “Industry 4.0 and lean management: a proposed integration model and research propositions,” Production & Manufacturing Research, vol. 6, no. 1, pp. 416–432, Jan. 2018, doi: 10.1080/21693277.2018.1540949.

[2] L. D. Xu, E. L. Xu, and L. Li, “Industry 4.0: state of the art and future trends,” International Journal of Production Research, vol. 56, no. 8, pp. 2941–2962, Apr. 2018, doi: 10.1080/00207543.2018.1444806.

[3] B. Rodič, “Industry 4.0 and the New Simulation Modelling Paradigm,” Organizacija, vol. 50, no. 3, pp. 193–207, Aug. 2017, doi: 10.1515/orga-2017-0017.

[4] Y. Cheng, K. Chen, H. Sun, Y. Zhang, and F. Tao, “Data and knowledge mining with big data towards smart production,” Journal of Industrial Information Integration, vol. 9, pp. 1–13, Mar. 2018, doi: 10.1016/j.jii.2017.08.001.

[5] L. Cavanini et al., “A Preliminary Study of a Cyber Physical System for Industry 4.0: Modelling and Co-Simulation of an AGV for Smart Factories,” in 2018 Workshop on Metrology for Industry 4.0 and IoT, Brescia, 2018, pp. 169–174, doi: 10.1109/METROI4.2018.8428334.

[6] A. Wortmann, B. Combemale, and O. Barais, “A Systematic Mapping Study on Modeling for Industry 4.0,” in 2017 ACM/IEEE 20th International Conference on Model Driven Engineering Languages and Systems (MODELS), Austin, TX, 2017, pp. 281–291, doi: 10.1109/MODELS.2017.14.

[7] J. Dizdarevic, F. Carpio, A. Jukan, and X. Masip-Bruin, “Survey of Communication Protocols for Internet-of-Things and Related Challenges of Fog and Cloud Computing Integration,” ACM Computing Surveys, vol. 51, no. 6, pp. 1–29, Jan. 2019, doi: 10.1145/3292674.

[8] R. Petrasch and R. Hentschke, “Process modeling for industry 4.0 applications: Towards an industry 4.0 process modeling language and method,” in 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand, 2016, pp. 1–5, doi: 10.1109/JCSSE.2016.7748885.

[9] L. S. Dalenogare, G. B. Benitez, N. F. Ayala, and A. G. Frank, “The expected contribution of Industry 4.0 technologies for industrial performance,” International Journal of Production Economics, vol. 204, pp. 383–394, Oct. 2018, doi: 10.1016/j.ijpe.2018.08.019.

[10] S. Belkhala, S. Benhadou, K. Boukhdir, and H. Medromi, “Smart Parking Architecture based on Multi Agent System,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 3, 2019, doi: 10.14569/IJACSA.2019.0100349.

[11] M. W. Grieves, “Product lifecycle management: the new paradigm for enterprises,” International Journal of Product Development, vol. 2, no. 1/2, p. 71, 2005, doi: 10.1504/IJPD.2005.006669.

[12] OCP Group, “2017 Annual report OCP Group,” 2017.

[13] “Webcast: Dr. Michael Grieves - Digital Twin: Manufacturing Excellence through Virtual Factory Replication.” [Online]. Available: http://www.apriso.com/library/video/dr_grieves_digital_twin_webcast_en.php. [Accessed: 06-Apr-2019].

[14] M. Grieves and J. Vickers, “Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems,” in Transdisciplinary Perspectives on Complex Systems, F.-J. Kahlen, S. Flumerfelt, and A. Alves, Eds. Cham: Springer International Publishing, 2017, pp. 85–113.

[15] E. Glaessgen and D. Stargel, “The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles,” in 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, American Institute of Aeronautics and Astronautics.

[16] S. Weyer, T. Meyer, M. Ohmer, D. Gorecky, and D. Zühlke, “Future Modeling and Simulation of CPS-based Factories: an Example from the Automotive Industry,” IFAC-PapersOnLine, vol. 49, no. 31, pp. 97–102, 2016, doi: 10.1016/j.ifacol.2016.12.168.

[17] F. Yao, A. Keller, M. Ahmad, B. Ahmad, R. Harrison, and A. W. Colombo, “Optimizing the Scheduling of Autonomous Guided Vehicle in a Manufacturing Process,” in 2018 IEEE 16th International Conference on Industrial Informatics (INDIN), Porto, 2018, pp. 264–269, doi: 10.1109/INDIN.2018.8471979.

[18] K. Reifsnider and P. Majumdar, “Multiphysics Stimulated Simulation Digital Twin Methods for Fleet Management,” in 54th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, Boston, Massachusetts, 2013, doi: 10.2514/6.2013-1578.

[19] “Digital Twin Technology Can Make Smart Cities Even Smarter.” [Online]. Available: https://www.govtech.com/smart-cities/Digital-Twin-Technology-Can-Make-Smart-Cities-Even-Smarter.html. [Accessed: 18-Dec-2019].

[20] N. Mohammadi and J. E. Taylor, “Smart city digital twins,” in 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, 2017, pp. 1–5, doi: 10.1109/SSCI.2017.8285439.

[21] R. Martinez-Velazquez, R. Gamez, and A. E. Saddik, “Cardio Twin: A Digital Twin of the human heart running on the edge,” in 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Istanbul, Turkey, 2019, pp. 1–6, doi: 10.1109/MeMeA.2019.8802162.

[22] R. Lutze, “Digital Twins in eHealth –: Prospects and Challenges Focussing on Information Management,” in 2019 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), Valbonne Sophia-Antipolis, France, 2019, pp. 1–9, doi: 10.1109/ICE.2019.8792622.

[23] S. Boschert and R. Rosen, “Digital Twin—The Simulation Aspect,” in Mechatronic Futures, P. Hehenberger and D. Bradley, Eds. Cham: Springer International Publishing, 2016, pp. 59–74.

[24] F. Tao, M. Zhang, and A. Y. C. Nee, “Background and Concept of Digital Twin,” Elsevier, 2019, pp. 3–28.

[25] A. Detzner and M. Eigner, “A DIGITAL TWIN FOR ROOT CAUSE ANALYSIS AND PRODUCT QUALITY MONITORING,” DS 92: Proceedings of the DESIGN 2018 15th International Design Conference, 2018. [Online]. Available: https://www.designsociety.org/publication/40558/A+DIGITAL+TWIN+FOR+ROOT+CAUSE+ANALYSIS+AND+PRODUCT+QUALITY+MONITORING. [Accessed: 14-May-2019].

[26] F. Tao, J. Cheng, Q. Qi, M. Zhang, H. Zhang, and F. Sui, “Digital twin-driven product design, manufacturing and service with big data,” Int J Adv Manuf Technol, vol. 94, no. 9–12, pp. 3563–3576, Feb. 2018, doi: 10.1007/s00170-017-0233-1.

[27] P. André, F. Azzi, and O. Cardin, “Heterogeneous Communication Middleware for Digital Twin Based Cyber Manufacturing Systems,” in Service Oriented, Holonic and Multi-agent Manufacturing Systems for Industry of the Future, vol. 853, T. Borangiu, D. Trentesaux, P. Leitão, A. Giret Boggino, and V. Botti, Eds. Cham: Springer International Publishing, 2020, pp. 146–157.

[28] K. Borodulin, G. Radchenko, A. Shestakov, L. Sokolinsky, A. Tchernykh, and R. Prodan, “Towards Digital Twins Cloud Platform: Microservices and Computational Workflows to Rulea Smart Factory,”

in Proceedings of the10th International Conference on Utility and Cloud Computing - UCC '17, Austin, Texas, USA, 2017, pp. 209–210, doi: 10.1145/3147213.3149234.

[29] A. Ebrahimi, "Challenges of developing a digital twin model of renewable energy generators," in 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE), Vancouver, BC, Canada, 2019, pp. 1059–1066, doi: 10.1109/ISIE.2019.8781529.

[30] K. Josifovska, E. Yigitbas, and G. Engels, "Reference Framework for Digital Twins within Cyber-Physical Systems," in 2019 IEEE/ACM 5th International Workshop on Software Engineering for Smart Cyber-Physical Systems (SEsCPS), Montreal, QC, Canada, 2019, pp. 25–31, doi: 10.1109/SEsCPS.2019.00012.

[31] H. Yoo and T. Shon, "Challenges and research directions for heterogeneous cyber–physical system based on IEC 61850: Vulnerabilities, security requirements, and security architecture," Future Generation Computer Systems, vol. 61, pp. 128–136, Aug. 2016, doi: 10.1016/j.future.2015.09.026.

[32] M. Schluse and J. Rossmann, "From simulation to experimentable digital twins: Simulation-based development and operation of complex technical systems," in 2016 IEEE International Symposium on Systems Engineering (ISSE), Edinburgh, United Kingdom, 2016, pp. 1–6, doi: 10.1109/SysEng.2016.7753162.

[33] T. Borangiu, D. Trentesaux, A. Thomas, P. Leitão, and J. Barata, "Digital transformation of manufacturing through cloud services and resource virtualization," Computers in Industry, vol. 108, pp. 150–162, Jun. 2019, doi: 10.1016/j.compind.2019.01.006.

[34] "Digital twin of cost," Siemens PLM Software. [Online]. Available: https://www.plm.automation.siemens.com/global/fr/topic/digital-twin-of-cost/60535. [Accessed: 22-Nov-2019].

[35] Z. Song and A. M. Canedo, "Digital twins for energy efficient asset maintenance," US20160247129A1, 25-Aug-2016.

[36] "Connected Plant – Honeywell." [Online]. Available: https://www.honeywellprocess.com/en-US/online_campaigns/connected_plant/Pages/home.html. [Accessed: 17-Apr-2019].

[37] K. Lee, "Digital Transformation using Digital Twins," p. 37, 2018.

[38] J. E. Hershey, F. W. Wheeler, M. C. Nielsen, C. D. Johnson, M. J. Dell'Anno, and J. JOYKUTTI, "Digital twin of twinned physical system," US20170286572A1, 05-Oct-2017.

[39] Arnold M. Lund, Oakland, CA (US); et al., "DIGITAL TWIN INTERFACE FOR OPERATING WIND FARMS," 17-Nov-2016.

[40] "What Is Digital Twin Technology? | PTC." [Online]. Available: https://www.ptc.com/en/product-lifecycle-report/what-is-digital-twin-technology. [Accessed: 20-May-2019].

[41] Gerald Les Vossler, Grand Rapids, MI (US), "NETWORK FOR DIGITAL EMULATION AND REPOSITORY," 30-Jun-2016.

[42] Edward A . Fowler , Houston , TX (US, "DIGITAL TWIN OF CENTRIFUGAL PUMP IN PUMPING SYSTEMS," 20-Jun-2019.

[43] "The Digital Twin: Creating digital operations today to deliver business value tomorrow," Altran United Kingdom. [Online]. Available: https://www.altran.com/uk/en/insight/the-digital-twin-how-to-create-value-now-and-prepare-for-digital-operations-for-tomorrow/. [Accessed: 22-Nov-2019].

[44] "Digital Twin towards a meaningful framework - Arup." [Online]. Available: https://www.arup.com/perspectives/publications/research/section/digital-twin-towards-a-meaningful-framework. [Accessed: 25-Dec-2019].

[45] Q. Li et al., "Smart manufacturing standardization: Architectures, reference models and standards framework," Computers in Industry, vol. 101, pp. 91–106, Oct. 2018, doi: 10.1016/j.compind.2018.06.005.

[46] A. J. C. Trappey, C. V. Trappey, U. H. Govindarajan, J. J. Sun, and A. C. Chuang, "A Review of Technology Standards and Patent Portfolios for Enabling Cyber-Physical Systems in Advanced Manufacturing," IEEE Access, vol. 4, pp. 7356–7382, 2016, doi: 10.1109/ACCESS.2016.2619360.

[47] M. Helu, A. Joseph, and T. Hedberg, "A standards-based approach for linking as-planned to as-fabricated product data," CIRP Annals, vol. 67, no. 1, pp. 487–490, 2018, doi: 10.1016/j.cirp.2018.04.039.

[48] 14:00-17:00, "ISO/CD 23247-1," ISO. [Online]. Available: http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/50/75066.html. [Accessed: 26-Oct-2019].

[49] B. Melzer, "Reference Architectural Model Industrie 4.0 (RAMI 4.0)," p. 15.

[50] B. Melzer, "Reference Architectural Model Industrie 4.0 (RAMI 4.0)," p. 15.

[51] R. Leszczyna, "A review of standards with cybersecurity requirements for smart grid," Computers & Security, vol. 77, pp. 262–276, Aug. 2018, doi: 10.1016/j.cose.2018.03.011.

[52] G. N. Schroeder, C. Steinmetz, C. E. Pereira, and D. B. Espindola, "Digital Twin Data Modeling with AutomationML and a Communication Methodology for Data Exchange," IFAC-PapersOnLine, vol. 49, no. 30, pp. 12–17, 2016, doi: 10.1016/j.ifacol.2016.11.115.

[53] L. F. C. S. Durão, S. Haag, R. Anderl, K. Schützer, and E. Zancul, "Digital Twin Requirements in the Context of Industry 4.0," in Product Lifecycle Management to Support Industry 4.0, vol. 540, P. Chiabert, A. Bouras, F. Noël, and J. Ríos, Eds. Cham: Springer International Publishing, 2018, pp. 204–214.

[54] M. Rodríguez, J. R. Oviedo, and M. Piattini, "Evaluation of Software Product Functional Suitability: A Case Study," p. 12.

[55] S. Elnagar, H. Weistroffer, and M. Thomas, "Agile Requirement Engineering Maturity Framework for Industry 4.0," in Information Systems, vol. 341, M. Themistocleous and P. Rupino da Cunha, Eds. Cham: Springer International Publishing, 2019, pp. 405–418.

[56] M. Yli-Ojanperä, S. Sierla, N. Papakonstantinou, and V. Vyatkin, "Adapting an agile manufacturing concept to the reference architecture model industry 4.0: A survey and case study," Journal of Industrial Information Integration, Dec. 2018, doi: 10.1016/j.jii.2018.12.002.

[57] "Livre Blanc : La Transformation Digitale Au Maroc. – AUSIM MAROC." [Online]. Available: http://www.ausimaroc.com/livre-blanc-la-transformation-digitale-au-maroc/. [Accessed: 29-Dec-2019].

[58] S. El Hamdi, M. Oudani, and A. Abouabdellah, "Morocco's Readiness to Industry 4.0," in Proceedings of the 8th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT'18), Vol.1, vol. 146, M. S. Bouhlel and S. Rovetta, Eds. Cham: Springer International Publishing, 2020, pp. 463–472.

[59] T. D. West and A. Pyster, "Untangling the Digital Thread: The Challenge and Promise of Model-Based Engineering in Defense Acquisition," INSIGHT, vol. 18, no. 2, pp. 45–55, Aug. 2015, doi: 10.1002/inst.12022.

[60] J. Hereu, "1. SMART CITY EXPO CASABLANCA, ED. 2018 2. VILLES À VIVRE ET INNOVATION CITOYENNE 3. TEMPS FORTS A. DISCOURS 1 LA SMART CITY EN TANT QUE PROJET," p. 50.

[61] Z. Boudanga, S. Benhadou, H. Medromi, and J.-P. Leroy, "Development perspective of a Moroccan smart city," in 2019 Third World Conference on Smart Trends in Systems Security and Sustainablity (WorldS4), London, United Kingdom, 2019, pp. 247–254, doi: 10.1109/WorldS4.2019.8903961.

[62] Weltwirtschaftsforum and A. T. Kearney Incorporated, Readiness for the future of production report 2018. 2018.

[63] "SIMIT Simulation Platform (V9.1)," p. 876.

An Ontology-Driven IoT based Healthcare Formalism

Salwa Muhammad Akhtar[1]
Makia Nazir[2], Kiran Saleem[3]
Department of Computer Science and IT
The University of Lahore
Lahore, Pakistan

Hafiz Mahfooz Ul Haque[4], Ibrar Hussain[5]
Department of Software Engineering
The University of Lahore
Lahore, Pakistan

***Abstract*—The recent developments in the Internet of Things (IoT) paradigms have significantly influenced human life, which made their lives much more comfortable, secure and relaxed. With the remarkable upsurge of the smart systems and applications, people are becoming addicted to using these devices and having many dependencies on them. With the advent of modern smart healthcare systems, and significant advancements in IoT enabled technologies have facilitated patients and physicians to be connected in real-time for providing healthcare services whenever and wherever needed. These systems often consist of tiny sensors and usually run on smart devices using mobile applications. However, these systems become even more challenging when there is a need to take intelligent decision making dynamically in a highly decentralized environment. In this paper, we propose a Belief-Desire-Intention (BDI) based multi-agent formalism for ontology-driven healthcare systems that perform BDI based reasoning to take intelligent decision making dynamically in order to achieve the desired goals. We illustrate the use of the proposed approach using a simple case study with the prototypal implementation of heart monitoring applications.**

Keywords*—*Internet of Things; BDI reasoning agents; ontology; smart healthcare

I. Introduction

Recent years have witnessed rapid advances in the smart systems' development with the incorporation of IoT paradigms in the ubiquitous computing environment. IoT is a network of interconnected devices/sensors that enable the systems to share their collected data and process information in order to achieve a common goal. IoT follows certain protocols for communication with the availability of the internet in accessing real-time device data and remote management of the system. In recent years, a significant amount of literature has revealed numerous kinds of smart systems and applications in various domains, more specifically, in healthcare systems, supply chain management systems, telemedicine, security systems, and in safety-critical systems, etc. IoT based healthcare systems and applications have a huge impact on the urban areas where there is an increasing dependency on these devices because elderly people mostly stay alone in their homes. According to research in [1], it is indicated that around 89% of elderly people likely to stay home alone and it is expected to rise by 22% (2 billion population approximately) by 2050. Chronic diseases are the diseases people suffer from, for a very long time. These diseases become more fatal to the person with age. Elderly people who usually stay alone are more affected by such chronic diseases as compared to those who stay with their caregivers. The rapid proliferation of IoT based healthcare solutions mostly consists of smarter, lightweight sensor/embedded sensors attached to comfortable wearable devices to monitor the patient's health and diagnose them accordingly. Using this mechanism, raw data is acquired by the sensors and then these systems compute the data to infer the derived results based on the provided facts from the sensors. However, the ability of data acquisition and analysis of the results could become troublesome in generating the desired results due to inadequate working on framework developments and poor applicability of the system. These systems often exhibit self-adaptive behavior in nature, mostly interact with users autonomously, run in a highly decentralized environment and can naturally be implemented as multi-agent systems. An agent is a piece of software that acquires information from the sensor/environment, perform reasoning, exchange information to/from other agents and infer the desired goals. An agent can be pro-active, reactive and has the capability to take decisions autonomously [2][3][4]. Multi-agent systems use different reasoning techniques to develop smart healthcare systems. Among others, BDI (Belief-Desire-Intention) based reasoning is considered to be most optimistic approach due to its simplistic nature and the usage of folk psychological notions which corresponds to the human behavior as human think, plan and make an intention to meet his/her desired goals [5]. In literature, numerous IoT based healthcare systems are available. There are quite few IoT based frameworks available which acquire and process ontology-driven information. However, such healthcare frameworks incorporating BDI based reasoning are still lacking and having a thirsty area of research. In this paper, we present a semantic knowledge-based healthcare formalism using BDI reasoning agents to develop an intelligent decision support system. Our proposed solution to this dilemma is the proposed framework which gathers information of the patient using different sensors and agents and if any abnormality is detected, sends the relevant data to the medical server. We construct an ontology for modelling the domain which represents the overall structure of the system and the data being collected, while BDI-based reasoning is performed in order to achieve the desired results. We develop a heart monitoring application for the prototypal implementation of the proposed system.

The rest of the paper is structured as follows. In Section II, we present the related work. Section III presents BDI reasoning based smart healthcare formalism. Section IV discusses contextual modeling using ontology. Section V provides the architecture of the proposed framework. Section VI presents the case study for the prototypal implementation of the system and finally concludes in Section VII.

II. Related Work

Literature has revealed a significant amount of work on IoT based systems and applications in different domains of computer science, more specifically, using mobile application development, Raspberry Pi and Arduino frameworks. A majority of such applications acquire information from different sensors, cameras and other devices/agents, and provide assisted support dynamically according to the given situation. In the case of safety-critical systems, these systems detect hazardous situations and take decisions intelligently. For example; authors in [6] claimed that health-monitoring devices offered a great solution for those patients who could not visit their doctors regularly. The sensors in these healthcare monitoring devices are intelligent enough and have the capabilities to detect any abnormalities and cope up with the situation accordingly. In the healthcare episode-based scenario, there are usually eight stages of interaction among patients and doctors. The framework proposed in [6] reduces these stages by allowing patients to be monitored from the comfort of their home, where agents can help the situation and interact with patients and doctors in case of an emergency. Moreover, agents can advise the patient about his/her situation and can allow the patient to directly contact the doctor by sending the query of patient's health and answering questions such as, whether the visit is important under the current health situation or not, if certain medications need to be continued or not, etc. In [7], the authors discussed that in WSN applications, the agent-based approach is important in terms of dynamic and goal-based decisions. WSN has characteristics like physical distribution, resource boundedness, information uncertainty, large scale, decentralized control and adaptiveness. Such properties can be shared and handled by multi-agents to achieve better results in scalability, decentralized control, resource-aware and situation-aware scenarios. In [8], the authors had a similar concept of using intelligent agents with healthcare technology in order to provide a better life for patients and reduce the expenses of hospital visitation when it's not compulsory. The usage of agents helped the patients to monitor their vital signs by themselves with the help of intelligent agents [9][10][11]. In [12], the authors discussed the usage of agents with WSN in healthcare. They concluded that using intelligent agents in healthcare applications greatly enhances the data processing stage. The authors also concluded that intelligent agents are very suitable in such applications due to their ability to adapt according to dynamic environments and reach a conclusion based on dynamic data.

In [13], the authors introduced an ontology-based approach, called the Medical Decision Support System (MDSS) which collects most relevant information of the user by providing a personalized questionnaire according to the patient profile and then provides the patients with customized advice. In [14], the authors produced a system based on a formal ontology that integrates patient information and alerts data retrieved from electronic health records in order to classify the importance of alerts. This can be very helpful when a huge number of alerts are generated and sent to the physicians, which makes it very difficult for the physicians to manage all these alerts simultaneously. In [15] the authors produced an Information Gathering System (IGS) that collects the most relevant information related to the patient which improves the classical approach by customizing the interview with each patient by providing a questionnaire that is driven by a domain ontology. The proposed system is complex in nature due to the continuous collection and processing of data by the sensors, the sharing of the data with the agents for reasoning, reaching a decision about the health of the patient and sharing this information with the patient's doctor and caregiver. This level of complexity makes data handling a very tedious task prone to error. Thus, the ontology was developed to make performing this task easier. Using ontology makes the process of managing the acquired data very easy, as it allows proper structuring of the entire system, including all the agents involved and their collected data. In the proposed system, the agents involved are BDI-agents. These agents use the reasoning mechanism to perform reasoning in reaching an accurate decision about the condition of the patient and whether to inform the doctor and caregiver or not.

III. BDI Reasoning based Smart Healthcare Formalism (BDI-SmartHealth)

In this section, we propose BDI reasoning based smart healthcare formalism which assists doctors and caregivers about health monitoring of the patients without any delay while the patient's health is abnormal. The system consists of different agents which perform BDI reasoning in order to infer the desired goals. In the system, the proposed framework consists of three core layers:

- Application layer
- BDI based reasoning agents
- Ontology layer

The top layer is the application layer, which consists of an android application and its different interfaces, as shown in Fig. 1. The middle layer is the reasoning layer which consists of BDI-agents who use BDI-based reasoning to reach conclusions about situations. The lower ontology layer receives input from the data collected by the sensors, responsible for structuring and handling the acquired data. Fig. 1 provides an architecture of the proposed framework.

A. Ontology Layer

In this layer, ontology is developed to model the domain for data structuring. It is assumed that agents are getting contextual data from the sensors, which is stored in the ontology and then this contextual data is sent to BDI agents for performing reasoning on it. The desired results can be monitored on the application layer.

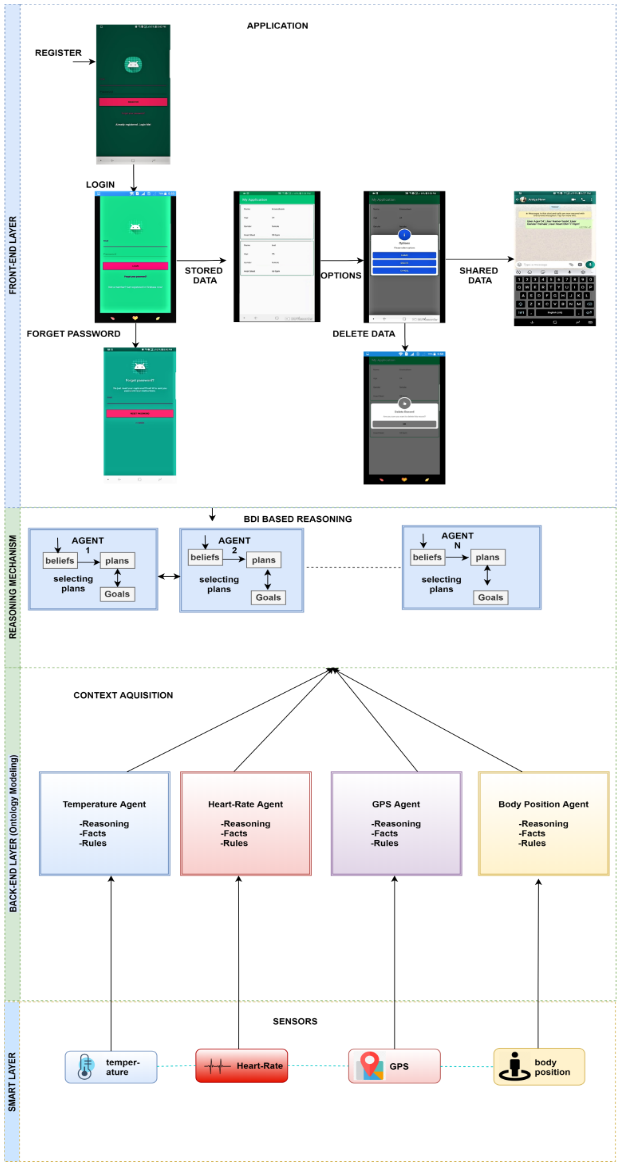


Fig. 1. BDI-Smart Health Framework.

B. BDI-Agent's Reasoning Layer

In this layer, multi-agent systems are involved with data acquisition, processing the data using BDI approach in order to reach the desired goals. These components are:

- Sensors
- Connectivity
- BDI-reasoning mechanism
- Remote server
- Storage
- Security

1) Sensors: Sensors can be WBAN (wireless body area network), wearable sensors or any other physical sensors (camera, heat sensors, etc.) to collect the data from the environment and pass it further for processing. WBAN connects devices that monitor the situation of the person, such as heart rate, body temperature, body posture, etc., processes the information, stores and shares it [16]. Sensors used in the proposed work are:

- Heart rate sensors: use to detect the pulse.
- Temperature sensor: use to detect body fever.
- GPS sensor: use to detect the location of the patient.
- Position sensor: to detect the patient's position.

2) Connectivity: Sensors can transfer the data for processing and reasoning, whether the vital signs are normal or not? Sensors send the data using Wifi or Bluetooth connectivity to the mobile application or laptops which are called the connectors, who perform reasoning on the data and send it to the remote server where the physicians/ doctors can monitor the patient's situation [17].

3) BDI- based reasoning: Sensors send the data using Wi-Fi or Bluetooth to the mobile application or other devices so that reasoning can be performed on the gathered data. When the data is sent to the connector, it performs reasoning mechanism in terms of its beliefs:

- What was the measured heart rate?
- What should be the maximum/minimum heart rate level?
- What was the physical activity of the patient at the moment?
- What was the posture position of the patient?

These questions play vital role in reaching a conclusion about the condition of the patient, as there are numerous other factors like stress levels, age and gender that can have a direct effect on the heart rate of a person. If there is some abnormality in the acquired data, then the connector selects a plan according to its intention for coping up with the situation and alerts the nearest hospital (remote server) about the medical emergency.

4) Remote server: Remote server or the medical care center is responsible for examining the patient's condition from the received data and send the required help if needed. This permits the patients to stay outside of the hospitals while allowing the doctors to examine their daily routine and vital signs continuously. All the data gathered by the sensors is sent to the remote server, where the doctors can treat the patients whenever needed.

5) Storage: Cloud and other databases are used for storage purpose. On the connector side, the data is stored on mobile application databases, while on the remote side, the data is stored either on the cloud or in the system's database.

6) Security: Security and privacy are major concerns when it comes to data transfer. In healthcare, it is the most important factor because sometimes the patient is not comfortable sharing his health status with anyone other than their doctors. The proposed application considers this concern and uses the login/ sign up option for the patient, which gives only the patient the complete access to his own records, while on the medical care side, the login/sign-up option, allows doctors the complete access of only their own patients. Multiple cryptography algorithms like AES and DES can be used to secure the data being shared.

C. Application Layer

This layer gives access to the android application prototype developed for framework validity. Android studio is used for the development of the android application to ensure the availability of a major medical facility, monitoring of the heart rate, in especially the rural areas, which comprises 60.78% of the total area of Pakistan according to a survey[1] in year 2016 [18]. We propose a BDI reasoning based approach, by employing body temperature, heart rate monitoring, blood pressure, body position, and location sensors. These sensors send the collected data to agents for performing reasoning on it. The sharing of the alert message, in case of an emergency, is done by these agents by using smartphones. We chose smartphones and android operating system as the hardware and software respectively for our solution, owing to the fact that 68% of the users in Pakistan have android systems on their smartphones, according survey conducted in [19][20], and also because, with the advent of the technological era, the availability and cost of smartphones with android system is decreasing with the launch of every new version of the Android operating system, resulting in the previous versions of the android operating systems to become more and more affordable for those belonging to low economic sector. This application allows its users to share their monitored heart rate with the medical experts via Whatsapp Messenger which is a cross-platform messaging and voice-over IP service owned by Facebook. The reason for choosing Whatsapp messenger is because until October 2018, it was ranked as one of the most popular mobile messaging application worldwide according to [21] and because of its end-to-end encryption feature, it is also secure and reliable, two very important non-functional requirements in any system, particularly when the system is responsible for saving and sharing critical data. A comprehensive elucidation of the steps involved in the working of the proposed android application is given below:

- An unregistered user will first register in the application by providing his/her basic information, and will then be given an account which will include user's e-mail and a password of the user's choosing, which can be employed by the user for login purposes, every time he/she wishes to use the services provided by the android application.

- A registered user will enter the username and password, chosen by the user during registration, and then click the 'login' button to login to the application.

- The sensors then start collecting and processing the data from their environment.

- If an anomaly occurs, this data is sent to the agent to perform further reasoning on the data.

- The agents after receiving the data and performing reasoning on it, reach a conclusion about the current condition of the patient.

- This decision is then shared with the patient's doctors and caregivers, with the help of the android application.

The doctors after receiving the data can draw a conclusion of the condition of the patient and prescribe him/her relevant measures for the improvement if necessary. The experts in this step reach a conclusion by keeping in mind the age of the patient, as with growing age, the heart rate may be affected more frequently during stress, which has become a growing issue in today's society [22] explained further in Fig. 2. The decision reached by the experts also depends on the gender of the patient as the heart rate of women was determined to be higher than men, regardless of age [23] as explained in Fig. 3, therefore the conclusion reached by the expert may vary with the patient.

Different interfaces have been developed for the Android application prototype of the proposed system to provide a better understanding of the application's working. Due to space constraints, a few screenshots of the interfaces for such an application have been presented from Fig. 4-7. Fig. 4 represents the login page, where the registered user will enter his/her e-mail and password to gain access to the application. Fig. 5 depicts the heart-beat value, Fig. 7 shows the value after heart-beat values taken and all the monitored and saved data is saved in the database which can be seen in Fig. 8.

All the monitored and saved data is saved in the database which can be seen in Fig. 7.

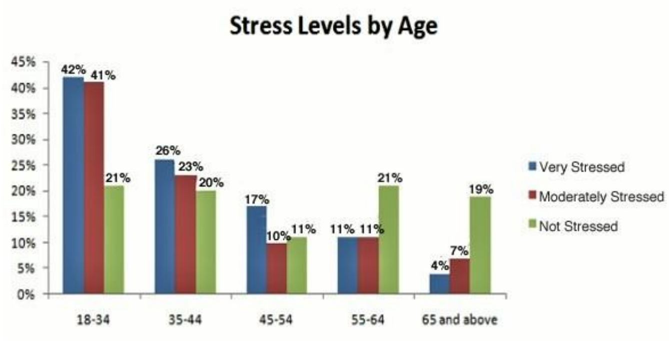
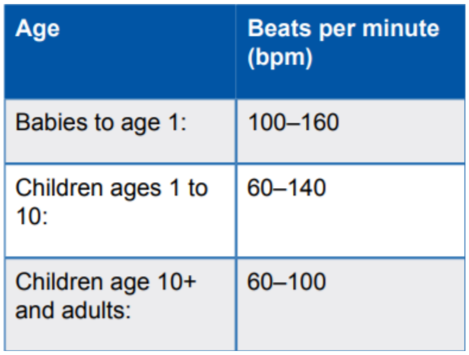


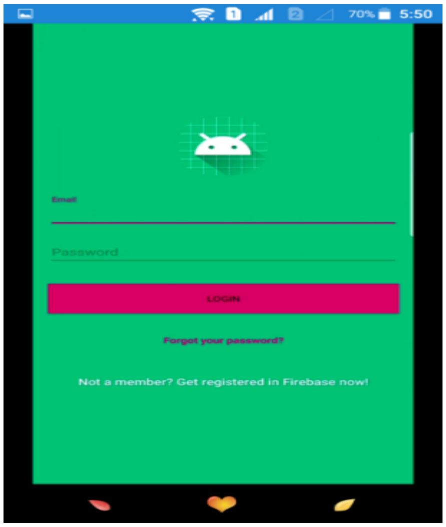
Fig. 2. Average Stress Levels by Age.



Age	Beats per minute (bpm)
Babies to age 1:	100–160
Children ages 1 to 10:	60–140
Children age 10+ and adults:	60–100

Fig. 3. Average Beats Per Minute (bpm) Depending on Age.

[1] www.tradingeconomics.com



Email

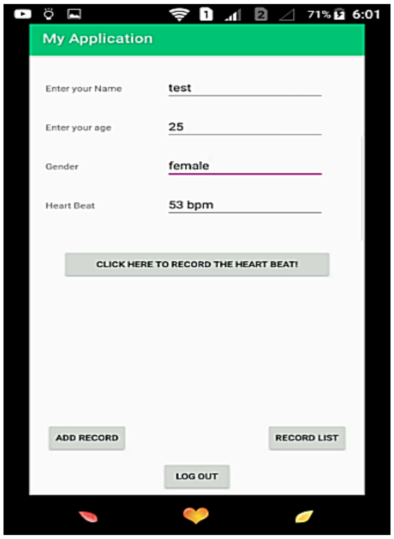
Password

LOGIN

Forgot your password?

Not a member? Get registered in Firebase now!

Fig. 4. Login Page.



My Application

Enter your Name test

Enter your age 25

Gender female

Heart Beat 53 bpm

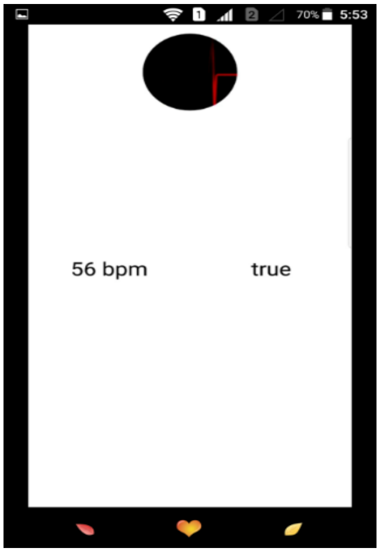
CLICK HERE TO RECORD THE HEART BEAT!

ADD RECORD

RECORD LIST

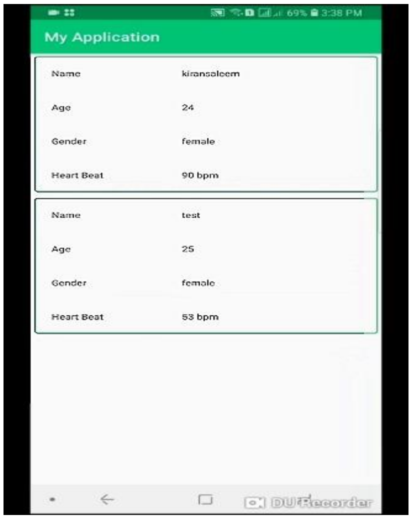
LOG OUT

Fig. 5. Main Interface.



56 bpm true

Fig. 6. Heart Rate Monitoring.



My Application

Name kiransaleem

Age 24

Gender female

Heart Beat 90 bpm

Name test

Age 25

Gender female

Heart Beat 53 bpm

Fig. 7. List of Recorded Data after Heart Rate Monitored.

IV. Context Modelling using Ontology

Ontology is a formal specification of a conceptualization of the domain which consists of different concepts/classes with their predefined relationships. There are two different versions of ontology named OWL 1 (Web Ontology Language) and OWL 2 (Web Ontology Language). Each version of the ontology has its own sub-languages. Ontology has been used on the proposed framework to fetch the data from several sensors and store them in a structured and organized manner. Ontologies are a formal way of describing the taxonomies and relationship between the concepts, entities and data. The scene description of contextual heart rate monitoring is modeled using specified description logic languages to map into logical simulation shown in Fig. 8.

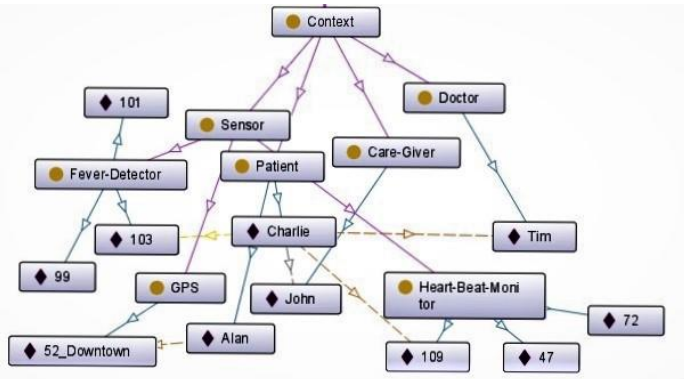


Fig. 8. Heart Rate Monitoring.

V. Architecture Design and Algo Calculation

This section shows the overall working of the reasoning mechanism of the proposed framework. In this step, the distances between two adjacent peaks are considered as the distance between the heartbeats. Knowing the average distance between two heartbeats, one can easily calculate the user's pulse in beats per minute. The normal range at rest is between 60-100 beats per minute (bpm). The basic way to calculate the rate is by taking the duration between two peaks and dividing this duration into 60. Each peak corresponds to a single heartbeat as shown in Fig. 9. The number of heartbeats and length of the measurement are all that is needed to calculate the heart rate. The resulting equation would be: Rate = 60/(peak interval).

For example, if the peak interval is 0.2 seconds then the heart rate is 60/0.2 = 300 bpm [24]. Fig. 10 shows the overall working of the reasoning mechanism of the proposed framework.

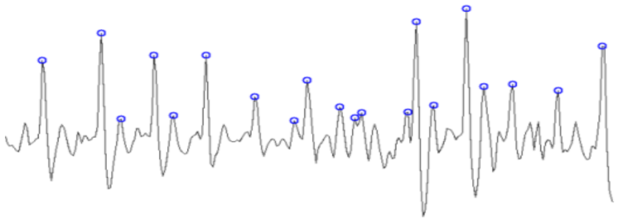


Fig. 9. Heart Rate "Peaks".

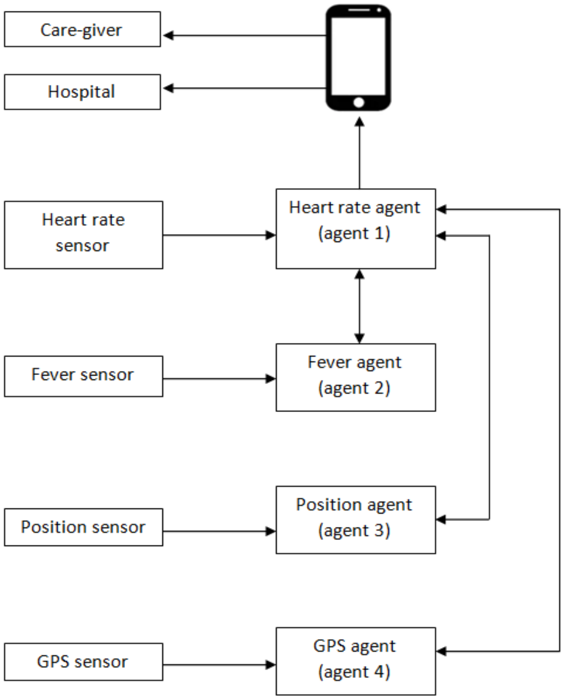


Fig. 10. Reasoning Mechanism.

VI. Case Study

A. General Scenario

To illustrate the use of the proposed formalism, we initially develop an ontology of smart healthcare systems. For this, we construct a simple example of a patient Charlie, a 60 years old male suffering from myocardial infarction. Charlie had a job in the city but has now retired. After retirement, his pension is not able to cover his living expenses in the city. Due to this reason, he has recently moved to a suburban rural part of the city. In this part of the city, the nearest medical facility (hospital) is 5 miles (8.05 km) away. Charlie has been suffering from myocardial infarction for more than 10 years. Myocardial infarction causes damage to the muscles of the heart due to interrupted blood flow because of a blood clot that develops in coronary arteries and can also occur if an artery suddenly narrows or spasms [25]. Because of this, he needs continuous monitoring of his vital signs, particularly his heart rate. Heart rate is directly related to body temperature. An increase in 10 bpm (beats per minute) occurs when the body temperature increases by even a single degree [26]. Heart rate also affects the blood pressure. When the blood pressure goes down, heart rate increases in an attempt to bring the blood pressure back to the normal level [27]. To provide Charlie with round-the-clock-care sensors are used. Some of these sensors, such as location sensors and body position sensor can be embedded in his house. Other sensors like heart rate monitor sensor, blood pressure sensor and body temperature sensor can be worn by Charlie. All these sensors transfer data to corresponding agents in a continuous manner and share information with other agents whenever needed and generate alerts, if required. Four agents are working in this system, namely, Heart-rate Agent (agent 1), Fever agent (agent 2) Position agent (agent 3) and GPS agent (agent 4). Agent 1 receives data from the heart rate sensor, Agent 2 receives data from the body temperature sensor, Agent 3 receives data from the body position sensor while agent 4 receives data from the location sensor. Each of these agents consists of Belief–Desire–Intention (BDI) reasoning, having a set of beliefs, intentions and goals. The proposed multi-agent system will be used in this scenario that is going to acquire the data from sensors.

- Heart-rate Agent (Agent 1): acquire heart rate, calculate the pulse data, and decide danger level.
- Fever agent (Agent 2): acquire body temperature from temperature sensors. Calculate temperature and decide danger level.
- Position agent (Agent 3): acquire body position from the position sensor. Calculate position, and send the position to another agent
- GPS agent (Agent 4): acquire location from the GPS sensor. Calculate location and send the location to the agent to alert the medical team.

Tables I-IV give details about the inputs, beliefs, intentions, and goals of all four agents.

TABLE. I. INPUTS, BELIEFS, INTENTIONS, AND GOALS OF AGENT 1

Heart Rate Monitor (Agent 1):

Inputs:

- Acquire the pulse data, calculate and danger level and interact with the agent.

Beliefs:

- Heart rate> 140, danger level (severe)
- Heart rate > 110 danger level (very high)
- Heart rate> 90 danger level (high)

Intention:

- If danger level severe
  - Interact with agent 3 (position agent) to get the position.
  - If the danger level is still severe
    - Interact with agent 4.
    - Get help from the medical center.
- If danger level high
  - Interact with agent 3 (position agent) to get the position.
  - If the danger level is still high according to the position
  - Contact with the caregiver and find a better treatment for further process

Goal:

- Provide better care so the heartbeat can come to its normal range.

TABLE. II. INPUTS, BELIEFS, INTENTIONS, AND GOALS OF AGENT 2

Agent 2:

Inputs:

- Acquire body temperature.

Belief:

- Temperature > 103 danger level severe
- Temperature > = 100 danger level high
- Temperature <=99 danger level normal

Intention:

- If danger level severe
  - Interact with agent 3 (position agent) to get the position.
  - If the danger level is still severe
  - Interact with agent 4.
  - Get help from the medical center
- If danger level high
  - Interact with agent 3 (position agent) to get the position.
  - If the danger level is still high according to the position
    - contact with the caregiver and find a better treatment for further process

Goal:

- Provide better care so the body temperature can come to its normal range.

TABLE. III. INPUTS, BELIEFS, INTENTIONS, AND GOALS OF AGENT 3

Agent 3:

Inputs:

- Acquire positions from motion sensors

Belief:

- If danger level severe of pulse data and Position is running, then check the normal range of pulse rate for running position.
- If its high according to it, then the danger level is severe and if the position is not running, then danger level is severe
- If danger level high and very high of pulse data, then get the position

Intention:

- Send the position and danger level to the contacted agent

Goal:

- Send position and danger level to the contacted agent so better treatment can be provided.

TABLE. IV. INPUTS, BELIEFS, INTENTIONS, AND GOALS OF AGENT 4

Agent 4:

Inputs:

- acquire location from GPS sensor

Belief:

- Patient's current location must be in latitude-longitude coordinates for accuracy

Intention:

- Provide an accurate location of the patient

Goal:

- Provide the shortest and safest route for the medical facility to arrive

B. Illustrative Example

Consider a scenario in which Charlie is sitting in his house reading a book, and he is being monitored by the healthcare system. Suddenly the heart rate monitoring sensor detects that Charlie's heart rate has dropped considerably, all the way to 45 bpm, instead of remaining between the normal range of 60-100 bpm. The heart rate sensor alerts the heart rate agent (agent 1), about this anomaly. While this is happening, the fever sensor detects that Charlie's body temperature has decreased to 34.5 °C, which is lower than the normal range of 36.5-37.0 °C. The fever sensor alerts the fever agent (agent 2), about this. Agent 1 and agent 2 communicate these pieces of information with each other and conclude that Charlie may be suffering from bradycardia, a major reason for myocardial infarction. However, to be more accurate agent 1 requests the position agent (agent 3) to provide him with Charlie's current body position. Once agent 1 has received information that Charlie's vital signs have gone beyond the specified ranges even though he is not doing any form of exertion, it becomes certain that Charlie is having a heart attack. After this, agent 1 requests a GPS agent (agent 4) to send Charlie's accurate location. Once agent 4 sends this information after receiving it from the GPS sensor, agent 1 sends all these details to the hospital as well as to Charlie's caregiver by using the android application installed on Charlie's phone. Charlie's phone after receiving the alert of heart attack sends a message to Charlie's emergency contact about Charlie's current situation as well as the hospital/doctor. The hospital/doctor after receiving information about Charlie's current situation reaches a conclusion about the treatment he should be given and notifies the ambulance service about Charlie's location to rescue the patient as soon as possible.

VII. Conclusion and Future Work

In this paper, we present an ontology driven IoT based healthcare framework that uses BDI reasoning agents to develop an intelligent decision support system. The proposed solution can be suitable to provide health-care services to the patients living in far-off locales, where such facilities are not readily available. The proposed android application prototype allows automatically sharing the patient's critical health-care data with their corresponding doctors and caregivers. This application also allows patients to have a way of knowing the acuteness of their condition at any given time by observing their previous medical records. In future work, we develop a logical framework for the development and deployment of a BDI reasoning based multi-agent system for the heterogeneous system. In addition, we shall add the feature in the application for an automatic comparison of the patient's previous and current records.

References

[1] Akram, M., Alam, H. M., & Iqbal, Z. (2018). Impact of Financing on Production with Mediating Role of Rural Population: Evidence from Agricultural Sector of Pakistan. Journal of the Research Society of Pakistan–Vol, 55(2), 85–92.

[2] Fu, J. and Fu, Y. (2015). An adaptive multi-agent system for cost collaborative management in supply chains. Engineering Applications of ArtificialIntelligence, 44:91 – 100.

[3] Chen, H. and Tolia, S. (2001). Steps towards creating a context-aware software agent system. Technical report, Hewlett Packard Labs, Palo Alto HPL-2001.

[4] Kwon, O. B., and Sadeh, N. (2004). Applying case-based reasoning and multi-agent intelligent system to context-aware comparative shopping. Decision Support Systems, 37(2):199 – 213.

[5] Georgeff, M., Pell, B., Pollack, M., Tambe, M. and Wooldridge, M., "The Belief-Desire-Intention Model of agency", 1999 Proceedings of Agents, Theories, Architectures and Languages (ATAL'99).

[6] V. Chan, P. Ray, and N. Parameswaran, "Mobile e-Health monitoring: an agent-based approach", TELEMEDICINE AND E-HEALTH COMMUNICATION SYSTEMS, IET Commun., Vol. 2, No. 2, February 2008.

[7] Giancarlo Fortino, Stefano Galzarano, Raffaele Gravina, Antonio Guerrieri, "Agent-based development of wireless sensor network applications", CEUR Workshop Proceedings, Vol. 741, No. 123-132, 2011.

[8] Giancarlo Fortino, Stefano Galzarano, Raffaele Gravina, Antonio Guerrieri, "Multi-Agent Application for Chronic Patients: Monitoring and Detection of Remote Anomalous Situations", Springer International Publishing Switzerland 2016 J. Bajo et al. (Eds.): PAAMS 2016 Workshops, CCIS 616, pp. 27–36, 2016.

[9] Wood, A., et al.: Context-aware wireless sensor networks for assisted living and residential monitoring. IEEE Network 22(4), 26–33 (2008).

[10] Zato, C., et al.: PANGEA: a new platform for developing virtual organizations of agents. Int. J. Artif. Intell. 11(13A), 93–102 (2013).

[11] Corchado, J.M., et al.: GerAmi: improving healthcare delivery in geriatric residences. IEEE Intell. Syst. 23(2), 19–25 (2008).

[12] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references).

[13] Thygesen, K., Jaffe, A. S., Chaitman, B. R., White, H. D., Zealand, N., & Canada, P. J. D. (2018). Fourth Universal Definition of Myocardial Infarction (2018). Journal of the American College of Cardiology, 72(18), 2231–2264.

[14] Lamine Benmimoune, Amir Hajjam, Parisa Ghodous, Emmanuel Andres, Samy Talha, et al.. Ontology-based Medical Decision Support System to Enhance Chronic Patients' Lifestyle within Ecare Telemonitoring Platform. International Conference on Informatics, Management and Technology in Healthcare ICIMTH, Jul 2015, Athènes, Greece. pp.279-282. hal-01263682.

[15] Rosier, Arnaud & Mabo, Philippe & Temal, Lynda & van Hille, Pascal & Dameron, Olivier & Deléger, Louise & Grouin, Cyril & Zweigenbaum, Pierre & Jacques, Julie & Chazard, Emmanuel & (DUCHEMIN) LAPORTE, Laure & Henry, Christine & Burgun, Anita. (2016). Remote Monitoring of Cardiac Implantable Devices: Ontology Driven Classification of the Alerts. Studies in health technology and informatics. 221.

[16] Mladen Milošević, Michael T. Shrove, Emil Jovanov, "Applications of smart-phones for ubiquitous health monitoring and wellbeing management", applications of smartphones for ubiquitous health monitoring and wellbeing management jita 1(2011) 1:7-15.

[17] BROENS, Tom, et al. Towards an application framework for context-aware m-health applications. International Journal of Internet Protocol Technology, 2007, vol. 2, no. 2, p. 109-116.

[18] Lamine Benmimoune, Amir Hajjam, Parisa Ghodous, Emmanuel Andres, Samy Talha, et al.. Ontology-Based Information Gathering System for Patients with Chronic Diseases: Lifestyle Questionnaire Design. Progress in Artificial Intelligence, Sep 2015, Coimbra, Portugal. pp.110-115, ff10.1007/978-3-319-23485-4_11. hal-01263333.

[19] Muhammad, N., McElwee, G., & Dana, L. (2015). Barriers to the Development and Progress of Entrepreneurship in Rural Pakistan. International Journal of Entrepreneurial Behaviour & Research, 44(August), 814–839.

[20] Tariq, K., Tariq, R., Ayesha, A., Hussain, A., & Shahid, M. (2019). Effects of smartphone usage on psychological wellbeing of school going children in Lahore, Pakistan. Journal of the Pakistan Medical Association, 69(7), 955–958.

[21] Kamel Boulos, M. N., Giustini, D. M., & Wheeler, S. (2016). Instagram and WhatsApp in health and healthcare: An overview. Future Internet, 8(3), 1–14.

[22] Marasco, V., Stier, A., Boner, W., Griffiths, K., Heidinger, B. and Monaghan, P. (2017) Environmental conditions can modulate the links among oxidative stress, age, and longevity. Mechanisms of Ageing and Development, 164, pp. 100-107.

[23] Mahmoud A. Alomari, Nihaya A. Al-Sheyab & Ali H. Mokdad (2019): GenderSpecific Blood Pressure and Heart Rate Differences in Adolescents Smoking Cigarettes, Waterpipes or Both, Substance Use & Misuse.

[24] Horton, J. F., Stergiou, P., Fung, T. S., & Katz, L. (2017). Comparison of Polar M600 Optical Heart Rate and ECG Heart Rate during Exercise. Medicine and Science in Sports and Exercise, 49(12), 2600–2607.

[25] Latvala, A., Kuja-Halkola, R., Rück, C., D'Onofrio, B. M., Jernberg, T., Almqvist, C., Lichtenstein, P. (2016). Association of resting heart rate and blood pressure in late adolescence with subsequent mental disorders: A longitudinal population study of more than 1 million men in Sweden. JAMA Psychiatry, 73(12), 1268–1275.

[26] Debashis Saha, Amitava Mukherjee. (2003). Pervasive Computing: A Paradigm for the 21stCentury.IEEE Computer Society, 36(3), 25-31.

[27] Ying, N. M. (2019). Bioengineering Principle and Technology Applications, Chapter 2 Measurement Of Body Temperature And Heart Rate For The Development Of Healthcare System Using Iot Platform., vol 2.

Ransomware Behavior Attack Construction via Graph Theory Approach

Muhammad Safwan Rosli[1], Raihana Syahirah Abdullah[2]*
Warusia Yassin[3], Faizal M.A[4], Wan Nur Fatihah Wan Mohd Zaki[5]
Centre of Advanced Computing Technology, Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia[1, 2, 3, 4, 5]

***Abstract*—Ransomware has becoming a current trend of cyberattack where its reputation among malware that cause a massive amount recovery in terms of cost and time for ransomware victims. Previous studies and solutions have showed that when it comes to malware detection, malware behavior need to be prioritized and analyzed in order to recognize malware attack pattern. Although the current state-of-art solutions and frameworks used dynamic analysis approach such as machine learning that provide more impact rather than static approach, but there is not any approachable way in representing the analysis especially a detection that relies on malware behavior. Therefore, this paper proposed a graph theory approach which is analysis of the ransomware behavior that can be visualized into graph-based pattern. An experiment has been conducted with ten ransomware samples for malware analysis and verified using VirusTotal. Then, file system among features were selected in the experiment as a medium to understand the behavior of ransomware using data capturing tools. After that, the result of the analysis was visualized in a graph pattern based on Neo4j which is graph database tool. By using graph as a base, the discussion has been made to recognize each type of ransomware that acts differently in the file system and analyze which node that have the most impact during analysis part.**

Keywords—Ransomware; behavior analysis; graph theory; file activity system; Neo4j

I. Introduction

Information security is one of the critical issues that has been addressed in order to maintain the operation of the system constantly [1]. Today, cybercriminal tend to target vulnerable users and communities such as company of business, government sectors and critical infrastructure for example healthcare. The attacks can cause high severity and impact in most cases that even small fraction of time influencing detection and prevention need to be very concern and critical. With the intrusion and attacks, the attackers can gain access of confidential data from the victims or injecting various malware inside victim’s machine [2]. With new challenges such as sophisticated malware that has been rampaging in our network, traditional conventional solution like signature-based detection that relies on malware attack pattern does not give higher impact and less efficient in preventing malware attacks [3].

Thus, a few solutions, techniques and approaches have been developed using sandboxes with features that capable to filter and distinguish between benign and malicious files. However, as the solutions which used multiple of sandboxes with virtual machines grow larger, they also consume huge amount of resources such as RAM, machine storage which are time consuming [3]. So, to mitigate the concern issues, researchers need to come with different approaches and solutions to defend against current and future threats and also to understand the behavior of the malware attacks and their interactions with victim’s machine [4].

The main problem remain persists yet and it still needs to keep on update where the researchers need to understand the malware behavior whether it is in network traffic or file activity system in the form of statistical and dynamic. This research also stressed out the problem in visualizing malware behavior since the data can be represented in an easy way to be understand such as in the form of graph instead a typical data form such as comma-separated value (CSV). Therefore, the main objective of this research is to study multiple sets of ransomware that will be selected into testing environment. With the result of the data from the experiment, the data were translated and visualized into graph form by using graph database tools in assisting the research development.

In summary, this paper makes the following contributions:

- The research shows an alternative approach and analysis behavior of ransomware by constructing several ransomware samples using file system activity as feature selection.
- The research analyses the ransomware using Process Monitor to capture data log and classifies behavior of ransomware as the log produces multiple attack vector of ransomware.

II. Ransomware at Glance

When it comes to an attack that causes colossal impact, ransomware is one of the malwares that shows high severity in cybersecurity threats. An individual as well as big corporations that heavily depend on network would be facing this risk and need to come up with mitigation strategies. Currently, the popular use of machine learning in many sectors has inspired malware researcher to use the approach as ransomware detection system that helps to increase detection rates [5]. Furthermore, the increases of attacks from new ransomware families shows that attackers improving themselves with numerous cunning and sophisticated features such as encryption mechanisms or propagation of worm [6]. Hence, the motivation of this research is to study anomaly behavior of ransomware and analyze the behavior based on ransomware distinct features. Generally, ransomware will act aggressively

*Corresponding Author.

by starting to encrypt the files of victim personal computer and then delivering a ransom note with a set of instructions for payment usually by using popular cryptocurrency which is Bitcoin [7].

Like other malwares, ransomware also has its own lifecycle that can be seen in Fig. 1. Several actions are needed to make ransomware attack successful when it infects the computer [8]. The chronological attack starts when the victim downloading suspicious link in email attachment or accidently drive-by download from suspicious website. This will lead into next steps which is the victim executes ransomware file since typical ransomware are executable file format. When the ransomware has been executed, the malware will try to establish the connection in Command & Control (C&C) so that the attacker can gain encryption key of victim to bargain with the victim.

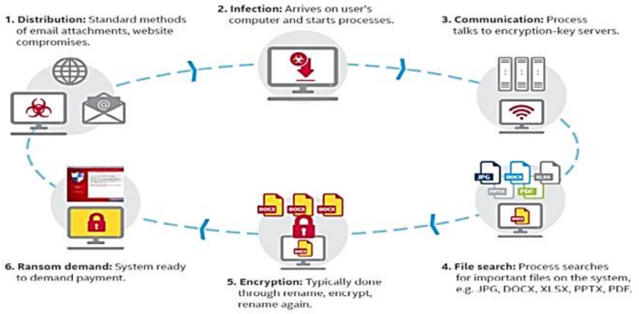

1. Distribution: Standard methods of email attachments, website compromises.
2. Infection: Arrives on user's computer and starts processes.
3. Communication: Process talks to encryption-key servers.
JPG
DOCX
XLSX
PPTX
PDF
DOCX
DOCX
DOCX
4. File search: Process searches for important files on the system, e.g. JPG, DOCX, XLSX, PPTX, PDF.
5. Encryption: Typically done through rename, encrypt, rename again.
6. Ransom demand: System ready to demand payment.

Fig. 1. Ransomware Lifecycle [4].

Next, the ransomware will search for related file with specific extension such as pdf, docx, xlsx, pptx, and jpg. Then, encryption will be done by renaming the file, encrypting the file, and then renaming it again. This steps will show most of the encrypted file with unique extension based on the type of ransomware. After the file in the directory has been encrypted, the ransomware will display a ransom note including the instruction and step by step to pay the ransom, mostly using Bitcoin transaction in The Onion Ring (TOR) protocol.

The author in [9] has stated three Indicators of Compromise or IoCs. These IoCs has been analyzed based on the result of ransomware behavior detection. The first indicator that has been identified is file changes in file system. By encrypting the files, ransomware changes the property of the files such as changing file extension and name of files. A second indicator which is file entropy can observe the randomness of file in file system. This is important for ransomware behavior detection since the encryption of ransomware causes high entropy which triggers the detection threshold and the system will detect it is as an attack. The third indicator is canary files which is a fake file that is implemented with real files. These files can set an alarm for a system if ransomware tries to encrypt the files thus, an early detection can be achieved.

The author in [10] classified ransomware as polymorphic and metamorphic also predicted threats for future ransomware that will be used by the attacker such as polymorphic blending of traffic or sandbox evasion technique. Therefore, most researchers are now focusing on Machine Learning (ML) techniques since it is capable to analyze ransomware behavior pattern thoroughly compared to other static analysis which depends on signature-based pattern in ransomware detection [11]. With constant changing of ransomware behavior, the researchers need to be alerted and further improved for current solution or framework.

III. Definition of Graph

A graph can be denoted as G = (V, E) which consists of vertices, V and edges, E. Element of vertices in graph is called nodes where entities, variable have properties, where element edges are called relation or connectivity. Vertices, V and Edges, E can be denoted as $V = \{v_i\}$ and $E = \{k_i\}$ such i = 1,2…,n and k = 1,2,…,m respectively. The value of k can be referred to the edge between v_i and v_j. There are multiple types of graphs depending on the degree of nodes and edges in graph structures such as undirected graph, directed graph, mixed graph, multigraph and weighted graph.

The basic of undirected graph is the two nodes which are connected to each other with identical edge such edge (a, b) is equal to edge (b, a). The total number of edges in this graph can be denoted as $n(n$-1)/2 without a loop in the graph [12]. A directed graph consists of nodes and the edges which are identical and can be denoted as G = (V, E). These nodes in the graph can be a set of multiple paired with edges in a form of line or arrow [12]. Three elements which can be seen in a directed graph are directed edge, in-degree and out-degree.

In addition, mixed graph consist of directed and undirected edges can be denoted as G = (V, E, A). Usually, ordered pair and unordered pair is called s arc and edge of a graph respectively [12]. Naturally, a multigraph is undirected type of graph with multiples edges. Consequently, the multiple edges also can be connected to multiple vertices or nodes that cause a loop or cycle. Lastly, a weighted graph will have weight assigned on its edges which represent quantities such as time, distance, force or monetary values. Network graph is one of the examples that can be referred to weight a graph as the graph contains measuring cost to an edge of network [12]. Unweighted graph can also be referred to a graph without weightage [12].

IV. Graph Theory

Graph theory is categorized as discrete mathematics in field of mathematics and offers visual representation using graph based on the given networks. Most graph theory combine with other analytical tools, several algorithms or frameworks in order to represent the analysis that has been done. Basically, a graph provides illustrative design to show relationship among the entities [12]. These entities are called node whereas vertices are relationship between the nodes. Multiple structures of graph are possible given information data in order to identify which graph has the most influential based on ranking of nodes. Most of graph theory analysis starts with graph key elements in order to understand the graph itself.

One of the key elements is cardinality that refers to the size of the set or number of elements in the set. In graph theory, vertex cardinality refers to the size or numbers of nodes or vertices. Thus, cardinality of the nodes denoted as n = |V|, where V refers to the number of nodes in the graph. Whereas, cardinality of the edges denoted as $m = |E|$, where E refers to the number of edges in the graph. Furthermore, centrality identifies the degree, distance and vertices of the node, then they rank its importance in the network [13].

Additionally, adjacency matrix is the edge between two nodes that are adjacent where the matrix represents adjacency relationship. Adjacency matrix present as A = $[a_{ij}]$ such i,j = 1,2,..n is define as square 0-1 matrix of size n x m which consists of binary value of 1 and 0 where [14]:

$$a_{ij} = \begin{cases} 1 \\ 0 \end{cases}, if\ nodes\ v_i\ and\ v_j\ are\ adjacent \quad (1)$$

Besides, incident matrix is a relationship between the nodes and edges in graph that can be define as $B = [b_{ij}]$ such i = 1,2,..n and j = 1,2,..m of size n x m where [15]:

$$a_{ij} = \begin{cases} 1 \\ 0 \end{cases}, if\ nodes\ v_i\ and\ v_j\ are\ adjacent \quad (2)$$

Degree of matrix also refers maximum values of each nodes in the graph is denote as $D = [d_{ij}]$ such i = 1,2,…n and j = 1,2,…m where [16]:

$$d_{ij} = \sum_{j=1}^{n} a_{ij} \quad (3)$$

Finally, clique is structure of a graph that consists of nodes with characteristics and has strong connection to each other's. Maximum clique is based on the number of nodes will denote as $\omega(G)$.

By using graph analysis, the interactive data analytics will become more flexible and easier to be used since the graph represent node and edges can be spotted [17]. Neo4j is a graph database that is highly praised as front-end and back-end media where social graphs represent real information. In addition, graph database also offers graph analytics, allowing method such as prediction and minimum spanning tree to become more popular in graph-based approach [18].

Graphs can represent network state transitions leading to attack goals, attacker exploitation steps related by preconditions and post conditions, intrusion alert sequences, logical dependencies for attack goals, or host attack reachability. Attack graphs have also been implemented with the relational model [19]. It also can become a useful tool for security and risk analysis since it can represent relationship between multiple node as shown in Fig. 2. However, the process of graph analysis can become tedious especially if the data is big and having a large node [20].

Computer network also can be represented by the graph analysis by defining entities such as IP address, hostnames to nodes and activities such as connection between the nodes to edges [21]. Since there are numbers of demand in graph analysis, three factors need to be considered which are graph data model, memory management and the algorithm of graph analytics [22].

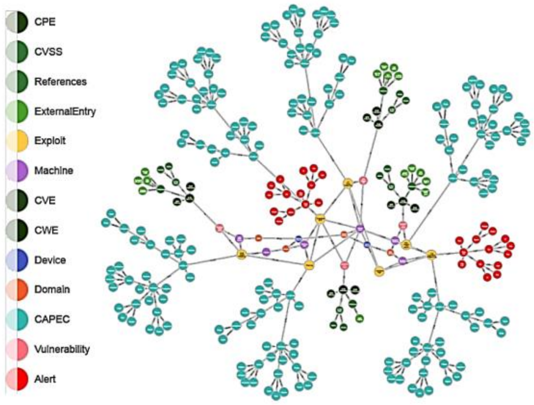


Fig. 2. Graph Model using Graph Database Tool [20].

V. Related Works

One of the disadvantages of static analysis approach is difficulty in extracting information with code obfuscation techniques. A file such as binary is also difficult to dissemble from time to time. Hence, dynamic analysis approach provides controlled environment which runtime information can be captured for analysis, since it does not affect extraction from complex technique [23]. Graph theory is a dynamic analysis approach, which helps in studying the relationship between the entities including the distance of each nodes and degree of nodes within the networks [24]. Furthermore, graph-based approach for detecting malware is an active area of research for malware targeted at PCs as well as mobiles [25]. Thus, a brief set of previous works that have been done by previous researchers, which applied graph theory approach in malware behavior, detection and graph techniques.

In malware detection, [1] proposed graph-based algorithmic technique using System-call Dependency Graphs (ScDG). The system calls that have been generated in ScDG will form a set of graphs called Group Relation Graphs (GrG). Based on the degree and vertices generated by the graphs, the author was able to investigate malware behavior in each system call in graph and enhance the accuracy of detection in their detection model.

To prevent malicious attacks in IoT devices, [26] has proposed behavior-based deep learning framework (BDLF), utilizing Stacked AutoEncoders (SAEs) and machine learning algorithm to obtain high level representation of malware behavior graphs. The framework uses Control Flow Graph (CFG) generated from the malware samples to analyze degree centrality, the size of graph, radius and shortest path of each node. This also includes the investigation of malware behavior similarities and the differences between IoT malware and android malware. Another framework called Together proposed by [27] is capable to generate massive graph provided by android malware samples and network files such as IP addresses and domains. The framework uses multiple heterogeneous graphs for network information to correlate each sub-threat network using Page Ranking algorithm to label malicious nodes in graph.

In addition, machine language instruction, OpCodes also benefit the application of graph theory in malware detection. [28] proposed a malware detection method in executable files by leverage graph using Power Iteration method embedded in graph properties such as eigenvectors and eigenvalues. Then, classification technique was used to classify each vector as malicious or benign. On the other hand, [29] proposed an android malware detection method using weighted probability graph of Dalvik Opcode. The author has presented important steps, which are Opcode sequences will be constructed into directed graph. Then, graph pruning will be executed as to reduce complexity of the graph structure while preserving critical information. Next, the author extracted and analyzed the similarity of features based on centrality and distance of graph in each malware samples using Manhattan Distance.

Furthermore, directed acyclic graph or DAG also can be used for malware classification. [30] used DAG technique as the graph inherits the properties of classification techniques in feature collection for different malware samples. In behavior-based detection method, [31] proposed graph-mining approach with malware behavior information such as system calls will be represented in Quantitative Data Flow Graph (QDFG). Then, the pattern of graph will be used for machine learning classifier to match between malicious or benign software.

The author in [32] proposed an approach for botnet detection where the method extracted malware samples and network traffic from darknet big data in order to investigate malware types and properties. The author utilized graph theorical of maximum spanning tree by implementing modified version of Kruskal’s algorithm. Cyberattacks comes from many vectors. Consequently, [4] addressed this issue using multimodal graph approach to identify possible sources or vectors such as actors, actions and means of cyber-attack. Then, the centrality of the graph will be measured in order to identify the highest value or most influential nodes in multimodal graph. Clustering technique, also one of the machine learning technique, has been widely used by researcher when it comes to malware. The author in [33] proposed malware clustering evaluation model using undirected topological graph to construct all malware samples guided by antivirus label information named Malware Relation Graph.

The research found several gaps in literatures, which are related to research problems. The researches in [1], [26], [27], [29], [30], [31], [32], and [33] have been identified with the absence of specific malware used and specific malware behaviour in literature which referred to the insufficient knowledge of malware that has been used during experiment, since each malware comes with massive amount of variant with various amount of behaviour. [28] comes with insufficient presentation of graph that is related to graph and not presented in details thus, the graph has lack of understanding. Although static analysis of graph comes with a complete analysis of an attack [4], a graph takes better advantage in dynamic analysis as the malware behaviour observation can be done in real-time with controlled environment.

VI. Methodology

To understand the behaviour of the ransomware, the research simulates the environment of the attacks by doing an experiment. This experiment consists of activities such as gathering the ransomware samples, selecting the tools to capture the behaviour of the ransomware and visualizing experiment result by using graph database tools for further discussion.

A. Design of Experiment

Fig. 3 shows a testbed environment where the experiment has been conducted along with collecting ransomware samples. The testbed environment is needed where the ransomware behavior will be captured using analysis tools in each Client and Server virtual machine. The initial of the testbed is to create the two VMs in main desktop using VMware Workstation 14 Pro then, the DHCP server of desktop will creating a set range of IPs for each ransomware samples running in each test. The operating system in each VMs are using Windows 7 and 4 GB of RAM. Table I shows the specifications of client and server virtual environment:

Two effective network analysis tools have been selected for the experiment and analysis stage where the tools are open source and have been used for multiple time for researcher to capture the data or even to analyze the data since analyzing the data are its main function. Process Monitor or known as ProcMon is a monitoring tool that shows a real-time environment of file system activity, registry and process activity. These logs are arranged into several columns that make them easier to read such as process ID, the operation of process, directory path and result of the process. ProcMon also can be used in monitor and record malware activity since it provides filtering function. While, Wireshark is used to capture the flow of network traffic and analyze the packet that has been captured through network interface card. Packet that has been captured will be presented into multiple types of information such as time, source, destination, protocol, length and the info of each packet frame. Finally, the research leverages a graph database tool to visualize the result from the experiment.

As for the datasets, the process of capturing starts with accessing a GitHub where most of ransomware samples are given for research purpose. Then, selected ransomware will be downloaded into executable format since the experiment is based on Windows operating system. Finally, the samples will be verified by the dynamic malware analysis sandbox, which called VirusTotal to authenticate the MD5 checksum as shown in Table II.

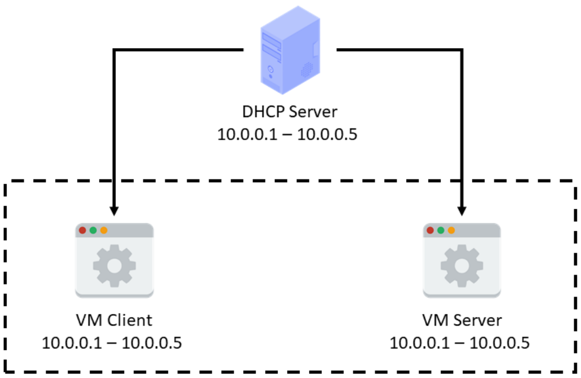


Fig. 3. Testbed Environment

TABLE. I. VIRTUAL ENVIRONMENT SPECIFICATION IN CLIENT AND SERVER

	Client	Server
Host OS	Windows 10 Pro	
Software	VMware Workstation 14 Pro	
Virtual OS	Windows 7 Ultimate SP1 64bit	
Virtual Memory	4 GB	
Virtual Processors	1 core per processor	
Virtual Hard disk	60 GB	
Virtual Network Adapter	VLAN	
Virtual Software	Wireshark v2.6.1, Procmon v3.5	

TABLE. II. RANSOMWARE DATASET PROPERTIES

Ransomware	MD5	File Size	File Type
Badrabbit	fbbdc39af1139aebba4da004475e8839	431.54 KB	Win32 EXE
Cerber	8b6bc16fd137c09a08b02bbe1bb7d670	604.5 KB	Win32 EXE
GoldenEye	e3b7d39be5e821b59636d0fe7c2944cc	254.5 KB	Win32 EXE
Jigsaw	2773e3dc59472296cb0024ba7715a64e	283.5 KB	Win32 EXE
Mamba	409d80bb94645fbc4a1fa61c07806883	2.3 MB	Win32 EXE
Mischa	8a241cfcc23dc740e1fadc7f2df3965e	878.5 KB	Win32 EXE
Rensenware	60335edf459643a87168da8ed74c2b60	96.5 KB	Win32 EXE
Satana	46bfd4f1d581d7c0121d2b19a005d3df	49.67 KB	Win32 EXE
TeslaCrypt	6e080aa085293bb9fbdcc9015337d309	257.5 KB	Win32 EXE
WannaCry	84c82835a5d21bbcf75a61706d8ab549	3.35 MB	Win32 EXE

B. Analysis Process

The flow of the experiment in Fig. 4 starts with using capturing tools that are used to capture dataset in data capturing process. Both tools are started to capture the network traffic and normal process before the execution of the malware samples. After the malware has been executed within certain period, the data analysis process starts to analyses the data from both results based on the tools, data network traffic from the Wireshark tool and resulting huge set number of PCAP (packet capture) files whereas data file activity system from Process Monitor tool resulting massive data log from the testing environment. Each of the data has been analyzed by using filter that provides by the tools to reduce the workload of analyzing both raw data information.

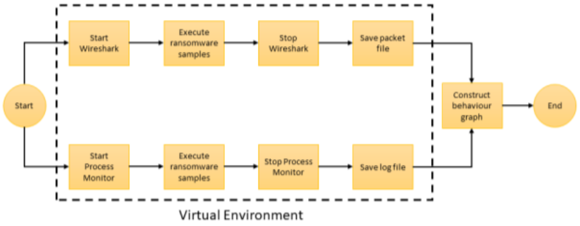


Fig. 4. Analysis Process in Flowchart

VII. ANALYSIS RESULT AND DISCUSSION

Visualizing and constructing the graph is quite challenging since node and edges need to be clarified before represented in the form of graph based on the result of experiment. Thus, a graph database tool is needed to assist in visualizing and representing the data. Neo4j is a property-graph type model, which uses node and edges concepts [34]. Multiple or single directed edge is used in Neo4j to define relationship between these nodes which means the nodes can possibly have multiple relationship with other nodes as well. A basic graph in Neo4j model consists on several elements, which are nodes, relationship, properties and labels. The nodes are described as the main element that connected to other nodes using relationship. The node and edges have properties that can be stored as key-value whereas label is described as roles to define types of node in the graph [35].

Additionally, the same concept of Relational Database Management System (RDBMS) is applied to graph database such to construct the graph, node and edges and need to be declared much like primary key and foreign key in RDBMS. Compare to other graph database tools, Neo4j provides its own database syntax called Cypher Query Language (CQL) which comparable to SQL that has been optimized for query in graph database so multiple variation of graph and complex conceptual connection can be visualized and expressed respectively [34] [36]. Therefore, by using Neo4j, the analysis result will be visualized into main graph, consists of multiple nodes that each node represents set of data analysis from the experiment.

Based on the overall graph model in Fig. 5, there are four types of nodes called Ransomware, FileOpen, FileCreate and FileExecutableUnderMalwareProcessTree. Ransomware node represent each sample has been used in the experiment. FileCreate and FileCreate nodes represent one or many types of DLL that have been accessed by each sample during the experiment and various sample files that have been infected by these ransomwares, respectively whereas FileExecutable UnderMalwareProcessTree node representing executable process that has been created during the experiment.

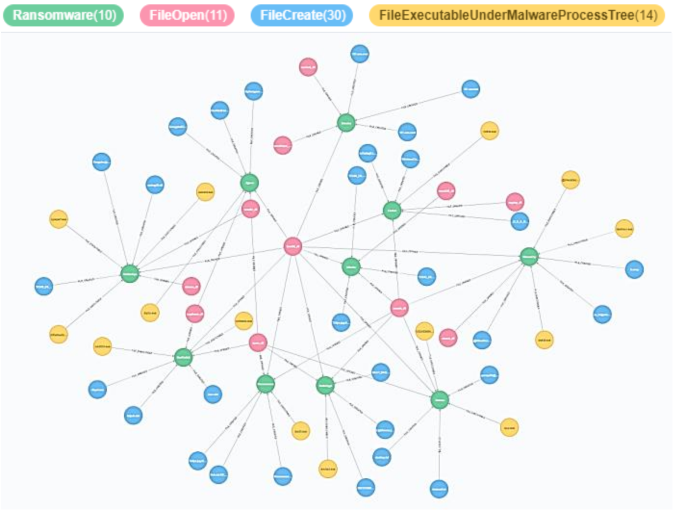


Fig. 5. Overall Graph Model.

There are several nodes shared the same behavior which are accessed the dynamic link library (DLL) during the execution of the ransomware. Other ransomware samples have their own distinguish DLL and the file that have been created in the file directory. All ransomware samples also create their own file executable process during the experiment. Below is the analysis of subgraph among ransomware behavior and its relationship between each file.

A. Badrabbit Ransomware

BadRabbit ransomware capable in deceiving victims into clicking it by creating false notification about Flash player update since it can hide using fake Adobe Flash. Then, the ransomware restarts the system after the attack entered the filesystem. During the process of execution, the ransomware will prompt UAC or User Access Control in order to obtain privilege since certain of the files need to have user permission. After that, it creates several malicious files in Windows directory such as infpub.dat which that responsible for modifying bootloader and encrypting the files.

BadRabbit ransomware subgraph contains multiple sets of nodes and edges that connected to each other as shown in Fig. 6. These set of nodes are labeled based on DLL files and has been accessed by this ransomware or the ransomware created the file or process in file system activity. From the analysis result, cryptbase.dll is among of highlighted DLL that has been accessed by this ransomware. This is because; the DLL is the Base cryptographic API DLL that was introduced in Windows NT 4.0 to provide services that enables developers to secure Windows-based applications using cryptography. Thus, the ransomware leverages the process to use its function to do malicious activity. Furthermore, it creates other file such as infpub.dat and cscc.dat, which are the main module of the malware to execute other process. Also, it creates other process under malware process tress, which are rundll32.exe and schtasks.exe. This process disguises as a normal behavior since it is created in Windows directory.

B. Cerber Ransomware

CRBR Encryptor or Cerber is among of ransomware that capable to encrypt the files even though the victims do not connect to the Internet. Like other ransomwares, file extension also has been renamed by the ransomware, namely, “.ba99”,”.98a0", ".a37b" and ".a563". However, the result of the experiment shows Cerber renamed the extension files as “.bdfa” due to variation version of ransomware.

In Fig. 7, the subgraph shows multiple nodes and edges connected to the main node of Cerber ransomware. Based on the data results, rsaenh.dll is among DLL files that have been highlighted in this subgraph because the function of this DLL is to implement 128-bit encryption of cryptographic service provider (CSP). Therefore, the ransomware leverages the process to use its function to do malicious activity. Likewise, it creates other process under malware process tress, which is mshta.exe. This process disguises as a normal behavior since it is created in Windows directory. Another DLL files that have been access are imm32.dll and cryptsp.dll. Among the files that have been captured are WindowsCodecs.dll, 1dTbfrlajT.bdfa and _R_E_A_D___T_H_I_S___SG08K_.txt. Also, it creates other executable process under malware process tree called mshta.exe.

C. GoldenEye Ransomware

GoldenEye is a type of ransomware that need to obtain administrative permission to proceed the encryption of the files. The unique behavior pattern of this ransomware is it capable to change the Master Boot Record (MBR) with custom boot loader. Then, the computer automatically reboot itself, showing a fake check disk while it performs encryption activity in the background process thus, recovering the data is practically impossible.

Based on the data result shown in Fig. 8, the ransomware creates additional malware called msimg32.dll in Windows filesystem, which is a Trojan dropper. Another typical behavior from this ransomware is it creates a file such as ransom note and ‘x4jBy3PY’ extension file, which is from the file that has been encrypted by the ransomware. Similarly, it creates other process under malware process tress, which are xwizard.exe, typeperf.exe and InfDefaultInstall.exe. These processes disguise as a normal behavior since they are created in Windows directory. Besides, other DLL files that have been access by this ransomware are wow64.dll and cryptsp.dll. Likewise, the files have been captured are Penguins.jpg.x4jBy3PY and YOUR_FILES_ARE_ENCRYPTED.TXT.

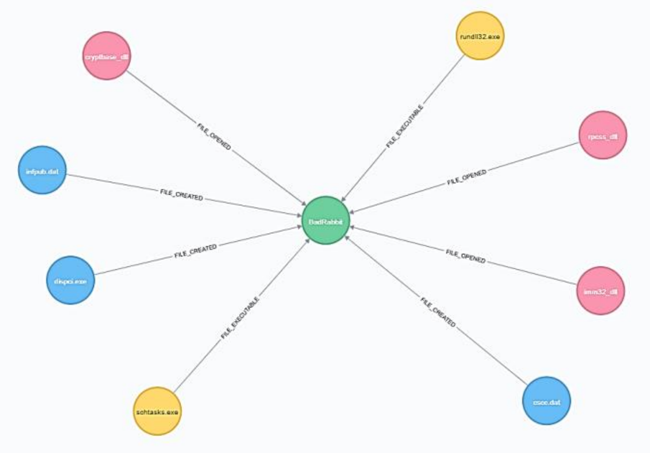


Fig. 6. Subgraph of Badrabbit Ransomware.

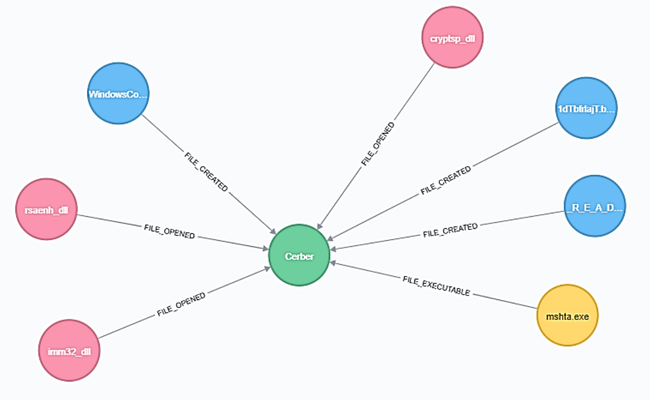


Fig. 7. Subgraph of Cerber Ransomware.

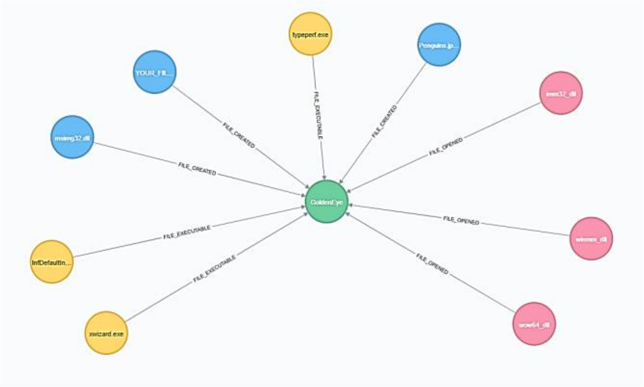


Fig. 8. Graph of GoldenEye Ransomware.

D. Jigsaw Ransomware

One of the distinct key features in identifying Jigsaw ransomware is the ransomware displays a ransom note featuring a character name Billy from movie called Saw. Based on the observation during experiment, the ransomware permanently deletes files from file directory if the ransom has not been done in specific time given.

One of the highlighted DLL files in Fig. 9 is cryptbased.dll. This DLL is the Base cryptographic API DLL that is introduced in Windows NT 4.0 to provide services that enables developers to secure Windows-based applications using cryptography. It also creates other file such as ransom note and ‘fun’ extension file, which is from the file that has been encrypted by the ransomware. Also, it creates other process under malware process tress, which are drpbx.exe. This process disguises as a normal behavior since it was created in Windows directory. Another DLL files that have been accessed are benign files such as imm32.dll and rpcss.dll. Also, there are few files are created by Jigsaw which are Hydrangeas.jpg.fun, RacWmiDatabase.sdf.fun and EncryptedFileList.txt.

E. Mamba Ransomware

HDDCryptor or known as Mamba is a type of ransomware that targets network sharing devices such as network printers, disk drives or network ports using SMB or Server Message Block. Like GoldenEye, the ransomware also requires a permission from administrator to change MBR or Master Boot Record.

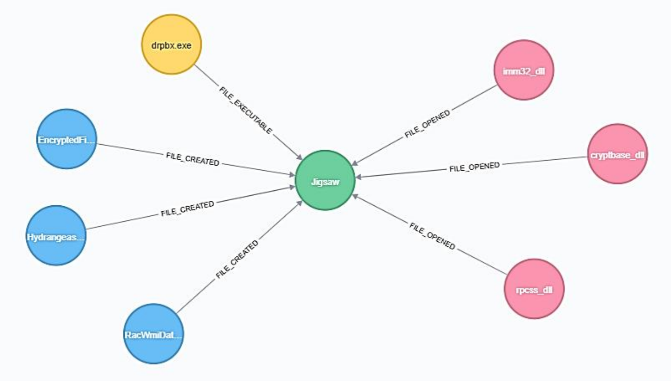


Fig. 9. Graph of Jigsaw Ransomware.

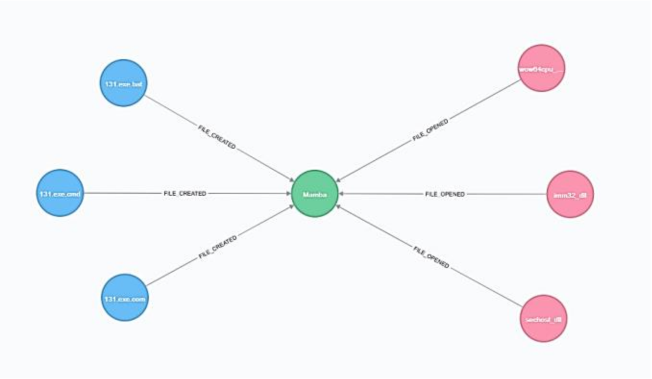


Fig. 10. Graph of Mamba Ransomware.

The subgraph of Mamba ransomware in Fig. 10 shows multiple nodes and edges connecting to main node. The highlighted nodes are DLL files that have been accessed by this ransomware called wow64cpu.dll. The DLL files are used to switch the processor from 32-bit to 64-bit mode when the application needs to be run in 64-bit format. Most of the applications that are using the Wow64 subsystem are created in SysWOW64 directory. Although it does not have an encryption in the experiment, it does have file activity processes, which are bat file, cmd file and com file in the same directory of the malware. Another DLL that has been accessed by this ransomware are imm32.dll and sechost.dll.

F. Mischa Ransomware

Mischa ransomware considered as a successor of Petya ransomware by its creators and it has become highly dangerous when it comes to sophisticated behavior. Different from Petya, the ransomware starts its behavior by scanning the system that has anti-virus software. Then the ransomware starts the encryption process while creating two ransom notes called YOUR_FILES_ARE_ENCRYPTED.HTML and YOUR_ FILES_ARE_ENCRYPTED.TXT to every folder in file directory.

The subgraph in Fig. 11 shows Mischa ransomware with multiple nodes and edges connecting to the main node. The highlighted nodes are DLL files that have been accessed by this ransomware called rsaenh.dll, which implements 128-bit encryption of cryptographic service provider (CSP). It also creates another file such as ransom note in text file and html file and ‘6NRS’ extension file image, which is from the file that has been encrypted by the ransomware. Based on our experiment, there is no process that has been created under this malware process tree.

G. Rensenware Ransomware

Rensenware is created with non-malicious intent and it is accidentally distributed in network that targets Windows OS users. However, if the ransomware finds several files that cannot be encrypted, it will crash itself and cannot be executed. The unique behavior of this ransomware is the victims are required to play certain game called “Touhou 12: Undefined Object”. The victim needs to achieve 200 million in “Lunatic” difficulty. Another interesting behavior is the ransomware does not delete the encryption key because it does not have one.

The highlighted nodes as show in Fig. 12 are files that have been created by this ransomware which each files extension that has been encrypted by this ransomware are renamed as “RENSENWARE”. Likewise, it creates other process under malware process tress, which is dw20.exe. This process disguises as a normal behavior since it was created in Windows directory.

H. Satana Ransomware

Satana or Satan ransomware is among of ransomware that operates as RaaS or Ransomware-as-a Service platform. Since it serves as a service, an attacker can implement various constraints or multiple behavior patterns based on functionality implementation. It relies on AES encryption module to encrypt victim’s data and demand for ransom using ransom note.

The highlighted nodes are files that has been created by this ransomware which it creates ransom note called “!satana!.txt” after encrypting files in each folder as shown in Fig. 13. The unique behavior of this ransomware is the encrypted file is always have this pattern which is “<email_address>__<original_name of file>”. Also, it creates other process under malware process trees, which are qxyi.exe and VSSADMIN.exe. This process disguises as a normal behavior since it was created in Windows directory.

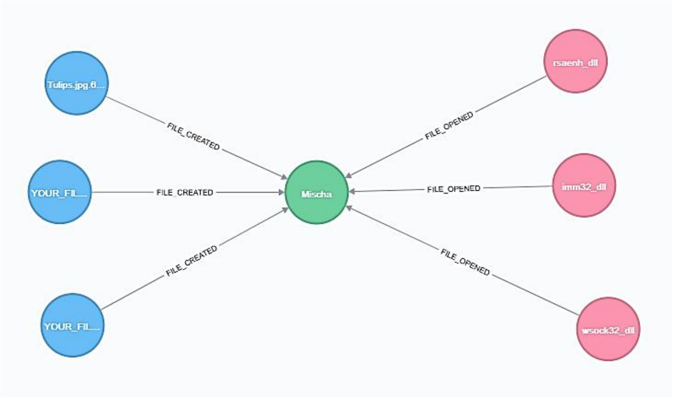


Fig. 11. Graph of Mischa Ransomware.

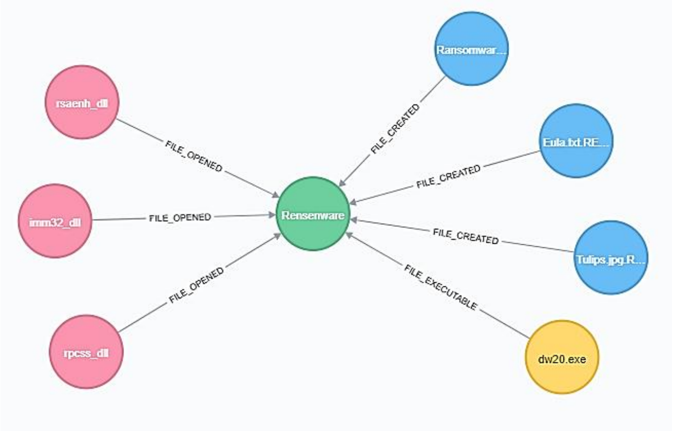


Fig. 12. Graph of Rensenware Ransowmare.

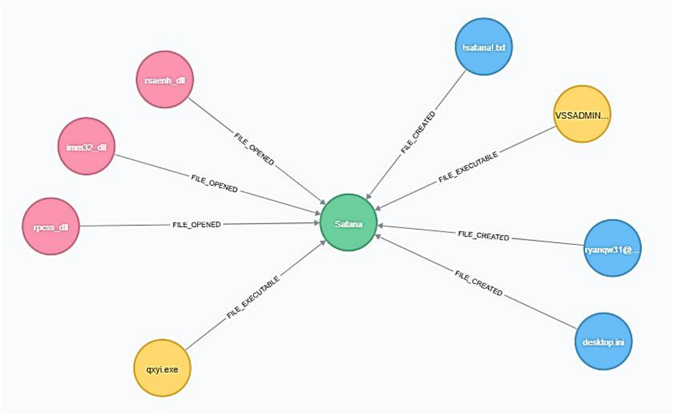


Fig. 13. Graph of Satana Ransomware.

I. TeslaCrypt Ransomware

TeslaCrypt ransomware behaves like other typical ransomware which it leaves ransom note called "HELP_RESTORE_FILES.txt" in each directory after encryption activity has been done. The unique behavior of this ransomware is it specifically target video game data such as data save or game settings in game file directory though, other variant targets different types of game files. Another similar pattern is it uses AES encryption for encrypting data files.

The nodes in Fig. 14 shows DLL that have been accessed by this ransomware are imm32.dll, rpcss.dll and rsaenh.dll whereas the nodes that have been created are Lighthouse.jpg.ecc, RECOVERY_KEY.TXT and HELP_ RESTORE_FILES.txt. Also, it creates other process under malware process trees, which is envtact.exe. This process disguises as a normal behavior since it was created in Windows directory.

J. WannaCry Ransomware

WannaCry or Wana Crypt0r is not something new when the ransomware spread the attack in May 2017 causing chaos around the world which giving awareness about how dangerous of ransomware. The ransomware uses RSA-2048 that is impossible to decrypt thus victims are required to pay ransom in Bitcoin based on the ransom note.

The highlighted nodes shown in Fig. 15 are FileCreate which is “@WanaDecryptor@.exe”. The function of this process is to show timers in ransom note and display the instruction of payment based on the language of operating system. It also creates another file such as ‘wnry’ extension, which consists of language, and normal file that is from the file that has been encrypted by the ransomware. Also, it creates other process under malware process trees, which is taskdl.exe and taskhsvc.exe. This process disguises as a normal behavior since it is created in Windows directory. Other files that have been captured during the experiment are Ransomware. WannaCry\b.wnry and m_bulgarian.wnry.

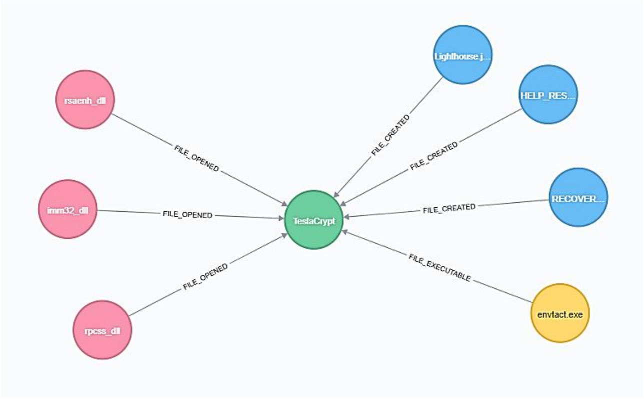


Fig. 14. Graph of TeslaCrypt Ransomware.

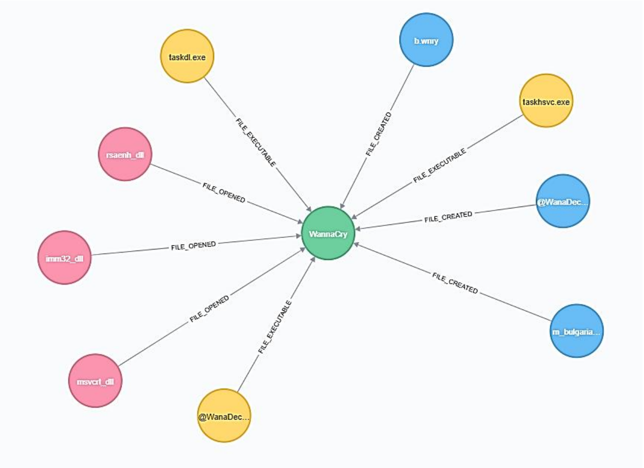


Fig. 15. Graph of WannaCry Ransomware.

VIII. Conclusion and Future Works

In conclusion, this research paper proposes an alternative method in representing the data of ransomware behavior using Neo4j graph database tools. Based on the research experiment, the data are analyzed the by classifying the ransomware behavior using analysis tools based on the log and network provided. Therefore, this research contributes in developing the level of awareness and knowledge of correlation between ransomware and graph theory approach besides providing different method in malware detection field community.

Graph theory analysis provides a new approach in malware detection. With the graph analysis, researchers can find significant relation of malware behavior, as the data from the experiment are represented as vertices and edges. The graph analysis also provides good representative based on the result from the experiment and gives better understanding of ransomware behavior in file system.

Currently, the limitation of this research is the experiment developed in offline environment, which does not project similar behavior of ransomware in online environment. Few ransomware samples also outdated or obsolete which means the old behavior does not reflect the latest ransomware with more sophisticated behavior. Thus, future research is to provide bigger scope by using multiple types of data from multiple sources such as memory and registry file. To improve the quality of the research, the next approach must have online environment features in order to capture live ransomware behavior from the experiment. The ransomware samples also have to increase as different behavior can be analyzed in comparative analysis. Moreover, with the current result of the experiment, it is a need to expand the research towards detection scheme as the analysis approach can take advantage in this scope.

Acknowledgment

This work has been supported under Universiti Teknikal Malysia Melaka research grant GLUAR/APNIC/2018/FTMK-CACT/A00018. Thank you to Research Group of Information Security Forensics and Computer Networking, Center for Advanced Computing Technology (C-ACT), Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka.

References

[1] A. Mpanti, S.D. Nikolopoulos, and I. Polenakis, “A Graph-based Model for Malicious Software Detection Exploiting Domination Relations between System-call Groups,” Proceedings of the 19th International Conference on Computer Systems and Technologies, pp. 20-26, September 2018.

[2] C. Birkinshaw, E. Rouka, and V.G Vassilakis, “Implementing an intrusion detection and prevention system using software-defined networking: Defending against port-scanning and denial-of-service attacks,” Journal of Network and Computer Applications 2019, vol. 136, pp. 71–85.

[3] C.H. Lin, H.K. Pao, and J.W. Liao, “Efficient dynamic malware analysis using virtual time control mechanics,” Computers and Security 2018, vol. 73, pp. 359–373.

[4] N. Ghose, L. Lazos, J. Rozenblit, and R. Breiger, “Multimodal graph analysis of cyber attacks,” Spring Simulation Conference (SpringSim), pp. 1-12, April 2019.

[5] R. Agrawal, J.W. Stokes, K. Selvaraj, and M. Marinescu, “Attention in Recurrent Neural Networks for Ransomware Detection,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3222-3226, May, 2019.

[6] M. Akbanov, V.G. Vassilakis, and M.D. Logothetis, “Ransomware detection and mitigation using software-defined networking: The case of WannaCry,” Computers & Electrical Engineering 2019, vol. 76, pp. 111-121.

[7] G. Hull, H. John, and B, Arief, “Ransomware deployment methods and analysis: views from a predictive model and human responses,” Crime Science 2019, 8(1), p. 2.

[8] B.A.S. Al-rimy, M.A. Maarof, and S.Z.M Shaid, “Ransomware threat success factors, taxonomy, and countermeasures: A survey and research directions,” Computers and Security 2018, vol. 74, pp.144–166.

[9] C. Chew, and V. Kumar, “Behaviour Based Ransomware Detection,” Proceedings of 34th International Conference, vol. 58, pp. 127-136, March 2019.

[10] N.K Popli, and A. Girdhar, “Behavioural Analysis of Recent Ransomwares and Prediction of Future Attacks by Polymorphic and Metamorphic Ransomware,” Advances in Intelligent Systems and Computing 2019, vol. 799, pp. 65–80.

[11] S.H. Kok, A. Abdullah, N.Z Jhanjhi, and M. Supramaniam, “Ransomware,Threat and Detection Techniques : A Review,” International Journal of Computer Science and Network Security 2019, 19(2), pp.136–146.

[12] T. Sangkaran, A. Abdullah, N. JhanJhi, and M. Supramaniam, “Survey on Isomorphic Graph Algorithms for Graph Analytics,” International Journal of Computer Science and Network Security 2019, 19(1), pp. 85–92.

[13] J. Xu, and H. Chen, “Criminal network analysis and visualization,” Communications of the ACM, 2005, 48(6) pp. 100-107.

[14] R. Mehatari, M.R. Kannan, M.R. and A. Samanta, “On the adjacency matrix of complex unit gain graph”, 2018. arXiv:1812.03747.

[15] O. De la Cruz Cabrera, M. Matar, and L. Reichel, “Edge importance in a network via line graphs and the matrix exponential,” Numerical Algorithms 2019, pp.1-26.

[16] K. Sato, “Optimal graph Laplacian”, Automatica 2019, vol. 103, pp. 374–378.

[17] R. Rossi, and N. Ahmed, “The network data repository with interactive graph analytics and visualization,” 29th AAAI Conference on Artificial Intelligence, pp.4292–4293, March 2015.

[18] G. Drakopoulos, A. Baroutiadi, and V. Megalooikonomou, “Higher order graph centrality measures for Neo4j,” 6th International Conference on Information, Intelligence, Systems and Applications (IISA), pp. 1-6, July 2015.

[19] S. Noel, E. Harley, K.H Tam, and G. Gyor, “Big-Data Architecture for Cyber Attack Graphs Representing Security Relationships in NoSQL Graph Databases,” 2015.

[20] S. Abraham, and S. Nair, “A predictive framework for cyber security analytics using attack graphs,” 2015. arXiv:1502.01240.

[21] E. Dull, “Cyberthreat analytics using graph analysis,” Cray User Group(CUG ’15), 2015.

[22] R.N Gottumukkala, S.R Venna, and V. Raghavan, “Visual Analytics of Time Evolving Large-scale Graphs,” IEEE Intelligent Informatics Bulletin, 16(1), pp.10–16, 2015.

[23] Y. Ding, X. Xia, S. Chen, and Y. Li, “A malware detection method based on family behavior graph,” Computers and Security, 2018, vol. 73, pp. 73–86.

[24] I. Martin, J.A. Hernandez, and S. de los Santos, “Machine-Learning based analysis and classification of Android malware signatures,” Future Generation Computer Systems 2019, vol. 97, pp. 295–305.

[25] A. Sharma, and B.A Prakash, “Graphs for Malware Detection : The Next Frontier,” Proceedings of the 13th International Workshop on Mining and Learning with Graphs (MLG), pp.8–10, 2017.

[26] F. Xiao, Z. Lin, Y. Sun, and Y. Ma, “Malware detection based on deep learning of behavior graphs,” Mathematical Problems in Engineering, February 2019.

[27] E.B. Karbab, and M. Debbabi, “Togather: Automatic investigation of android malware cyber-infrastructures,” Proceedings of the 13th International Conference on Availability, Reliability and Security, p. 20, August, 2018.

[28] H. Hashemi, A. Azmoodeh, A. Hamzeh, and S. Hashemi, “Graph embedding as a new approach for unknown malware detection,” Journal of Computer Virology and Hacking Techniques 2017, 13 (3), pp. 153–166.

[29] J. Zhang, Z. Qin, K. Zhang, H. Yin, and J. Zou, “Dalvik Opcode Graph Based Android Malware Variants Detection Using Global Topology Features,” IEEE Access 2018, vol. 6, pp. 51964–51974.

[30] M.K. Sahu, M. Ahirwar, and P.K. Shukla, “Improved malware detection technique using ensemble based classifier and graph theory,” 2015 IEEE International Conference on Computational Intelligence & Communication Technology, pp. 150-154, February 2015.

[31] T. Wuchner, A. Cisłak, M. Ochoa, and A. Pretschner, “Leveraging compression-based graph mining for behavior-based malware detection,” IEEE Transactions on Dependable and Secure Computing 2017, 16(1), pp. 99-112.

[32] E. Bou-Harb, M. Debbabi, and C. Assi, “Big data behavioral analytics meet graph theory: on effective botnet takedowns,” IEEE Network 106, 31(1), pp. 18-26.

[33] Y. Chen, F. Liu, Z. Shan, and G. Liang, “MalCommunity: A graph-based evaluation model for malware family clustering,” Communications in Computer and Information Science, 2018, vol. 901, pp. 279–297.

[34] R. Arora, and S. Goel, “JavaRelationshipGraphs (JRG): Transforming Java Projects into Graphs using Neo4j Graph Databases,” Proceedings of the 2nd International Conference on Software Engineering and Information Management, pp. 80-84, January 2019.

[35] Z. Zhu, X. Zhou, and K. Shao, “A novel approach based on Neo4j for multi-constrained flexible job shop scheduling problem,” Computers and Industrial Engineering 2019, vol. 130, pp. 671–686.

[36] D. Allen, A. Hodler, M. Hunger, M. Knobloch, W. Lyon, M. Needham, and H. Voigt, “Understanding Trolls with Efficient Analytics of Large Graphs in Neo4j,” Business, Technologies and Web (BTW) 2019, 2019.

A Framework for Producing Effective and Efficient Secure Code through Malware Analysis

Abhishek Kumar Pandey[1], Ashutosh Tripathi[2]
Alka Agrawal[4], Rajeev Kumar[5*], Raees Ahmad Khan[6]
Department of Information Technology
BBA University, Lucknow-UP, India

Mamdouh Alenezi[3]
College of Computer and Information Sciences
Prince Sultan University
KSA

***Abstract*—Malware attacks are creating huge inconveniences for organizations and security experts. Due to insecure web applications, small businesses and personal systems are the most vulnerable targets of malware attacks. In the wake of this burgeoning cyber security breach, this article propositions a framework for a complete malware analysis process including dynamic analysis, static analysis, and reverse engineering process. Further, the article provides an approach of malicious code identification, mitigation, and management through a hybrid process of malware analysis, priority-based vulnerability mitigation process and various source code management approaches. The framework delivers a combined package of identification, mitigation and management that simplifies the process of malicious code handling. The proposed framework also gives a solution for reused codes in software industry. Successful implementation of the framework will make the code more robust in the face of unexpected behavior and deliver a revolutionary stage wise process for malicious code handling in software industry.**

Keywords—Malware analysis; reuse code; framework; static analysis; dynamic analysis; reverse engineering; manual analysis

I. Introduction

The present cyberspace is imploding with attacks and breaches. Easy access to internet and quality less security mechanism has created much unusual and dangerous vulnerability in the current digital world. Malware is the biggest threat to the cyber world in a current situation [1]. Malware is the software that has some malicious or harmful set of operation or instructions in their source code for performing a malicious activity in a system or network [2]. Malicious software's have hidden malicious features. With name or structure, they are like normal useful software but after execution they perform harmful activities on the system. Millions of computer users are targeted by more than thousands of different malware daily. According to a study [3], in every 39 seconds, a malware attack is executed in the world.

The personal and professional tasks of today's digital generation are now software based and any software is made with source codes. Instructions and operations written by a coder into a particular language for execution on the computer are called source code [4]. Enormously growing speed of software industry is daily producing more than a hundred of new software for the users. Unfortunately, the malware creators take advantage of this huge population of software. Malware creators facilitate their malicious software with genuine software for more user accessibility. Every malicious code has some harmful features but they also have some good codes and the purpose of this framework is to provide good codes from malicious code for reuse in the industry with malware analysis. The growing and expansive rate of software industry creates the need to reuse codes for coders with some improvements instead of writing a new one.

The culture of reuse code is growing very fast in the software industry because reusing the codes reduces the efforts and, more essentially, saves on the time invested in the project. The time that a coder spends on a project is very valuable and if reusing of code reduces that valuable time, it is a great option for programmers. While working on malicious codes, it is very important to understand the harmful malicious activity of code for effective mitigation and that is the reason behind using malware analysis in the identification process.

The first segment of this article discusses the significance of reuse code in the business; the second segment characterizes the need of malware investigation. In third segment, the authors characterize the system for extricating secure great codes with malware examination and besides this portray the need and criticalness of the structure. After this clarification, in the last segment, the authors posit the conclusion and enunciate efforts directed towards future work.

II. Previous Research Initiatives

Authors of the proposed study find that many researchers provides the research article on malware analysis and portray various different type of frameworks for enhancing the malware analysis approach. In order to deliver the proposed framework authors find the following previous research initiatives.

Belal Amro provides a malware analysis technique for mobile devices that gives an analysis study of various malware analysis approaches on operating systems like android and IoS [13]. The paper focuses on frequently used phone set vulnerabilities and tries to assess their possible solution through malware analysis.

S. Chuprat et al. provides a framework and its implementation in big data environment. The proposed framework in this paper delivers an approach that analyzes and predicts the future threat of malware attack in big data

*Corresponding Author.

platform [14]. Authors of this study also provide some data in order to validate their results and framework workflow.

G. Hamsa et al. provides an analysis of various malware identification techniques and assess their pros and cons on different standards [15]. The exhausting review of paper on malware techniques provides a path for future researchers of malware and malware analysis.

The authors of this proposed study finds that there is lack of literature which is discussing the whole malicious handling approach under one roof. Many researchers provide various effective and novel approaches in order to enhance the identification and detection of malwares through malware analysis. But it is also evident that there is very less amount of literature is available that is discussing about the malicious code vulnerability identification, mitigation and management. Proposed framework will help the industry and future researchers in order to produce some useful codes through a hybrid approach associating malware analysis.

III. Importance of Reused Codes

Software industry is growing voluminously. Coders code new logics and functions every day but a new code takes too much time and efforts to be coded. Every coder faces some challenges like understanding user requirements, time of completing a project and so on [5]. Reuse of existing code eases the coders' tasks in multiple ways. Reuse of code gives a key to the coder for easily understanding the needs of client. Thus, the time of completing the project is much less when compared to the time invested in writing new codes. Embedded system development has secured an important place in the software industry in the last decades and average time duration of completing an embedded system project is a minimum of 12-14 months [6]. The phrase 'time is money', is indeed most apt for the software industry. Any product line is worthwhile only if satiates the end user's needs in a given timeframe. The immensely competitive pace at which the companies churn out products in the software arena must meet the time targets. This necessitates the reuse of code in software development. Automated Program Repair (APR) approaches also open a door and create the demand of reusable codes for creating patches and findings bugs. The basic work process of APR's are totally depends on reusable codes [16]. This type of scenario also refers to the need of effective framework that produces some useful codes from malicious vulnerable codes.

IV. Why Malware Analysis?

Malware are increasing at an alarming rate for several reasons. These reasons create many challenges and issues for the cyber expert. According to the study of Forbs Magazine, 25% of Malware target the financial information of users [9]. The study also shows that the number of hacked account a hacker has, this makes it easy for hackers to exploit. Malware analysis helps in objective identification of malware or malicious code. There are three main techniques of malware analysis (i) Static Analysis (ii) Dynamic Analysis (iii) Reverse Engineering.

Many researchers have proposed their ideas on malware analysis methodologies. Yuhei Kawakoya et al. shows the methodology of malware analysis with the help of sandboxing and API calls analysis. The paper tells the process of taint assisted malware analysis and enhances the malware identification steps [7]. Kamla Kant Sethi et al. gives a framework of malware analysis for classifying the malware with identification of the malware by using Sandboxing Tools and Machine level learning tools for extracting exact information about malicious software [8]. Li Li et al. portray a systematic review on the need for static analysis in malware detection but the methodology that is described in the paper uses the automated tools for static code analysis of malware. The methodology uses the call graph analysis technique to examine the calls, variables, and classes of code and other significant attributes of a source code [11]. Christian Camilo et al. shows the significance of machine level approach in their paper and focuses the whole analysis process of malware on machine learning for better results [12].

These research studies are based on enhancing the malware analysis process for better results. However, there is a need for mitigation of malicious codes also and further research initiatives must pivot on this. Every researcher needs to focus on the useful codes written with malicious codes for helping the software industry by providing codes for reuse as well as identifying malware and mitigating them.

V. Framework

Malware analysis deals with the study of how malware functions and about the possible outcomes of infection given by a specific malware. When an attacker writes a malicious application code, he also uses or writes some good code for hiding the malicious activity of that application and also for increasing the user acceptability of application. This is akin to steganography. The objective of this framework is to extract or retrieve the good code from malicious code for reuse. The Framework is a full package of identification, mitigation and managing the code by combining malware analysis for extracting useful codes. The authors have classified this framework into three phases (1) Monitoring (2) Mitigating and (3) Managing. A brief explanation of these three phases is enumerated below:

A. Phase 1: Monitoring Phase

Objective of this phase is to understand the purpose, functionality and structure as well as the vulnerabilities of the malware for extracting good codes and easy mitigation and management. Monitoring phase is a combination of all three methodologies of malware analysis (static analysis, dynamic analysis and reverse engineering). The developer uses automated analyzers in the monitoring phase for detecting and examining the malware easily and this is done in considerably less time. The authors categorize the monitoring phase into three sub-phases that are shown in Fig. 1 and described as follows:

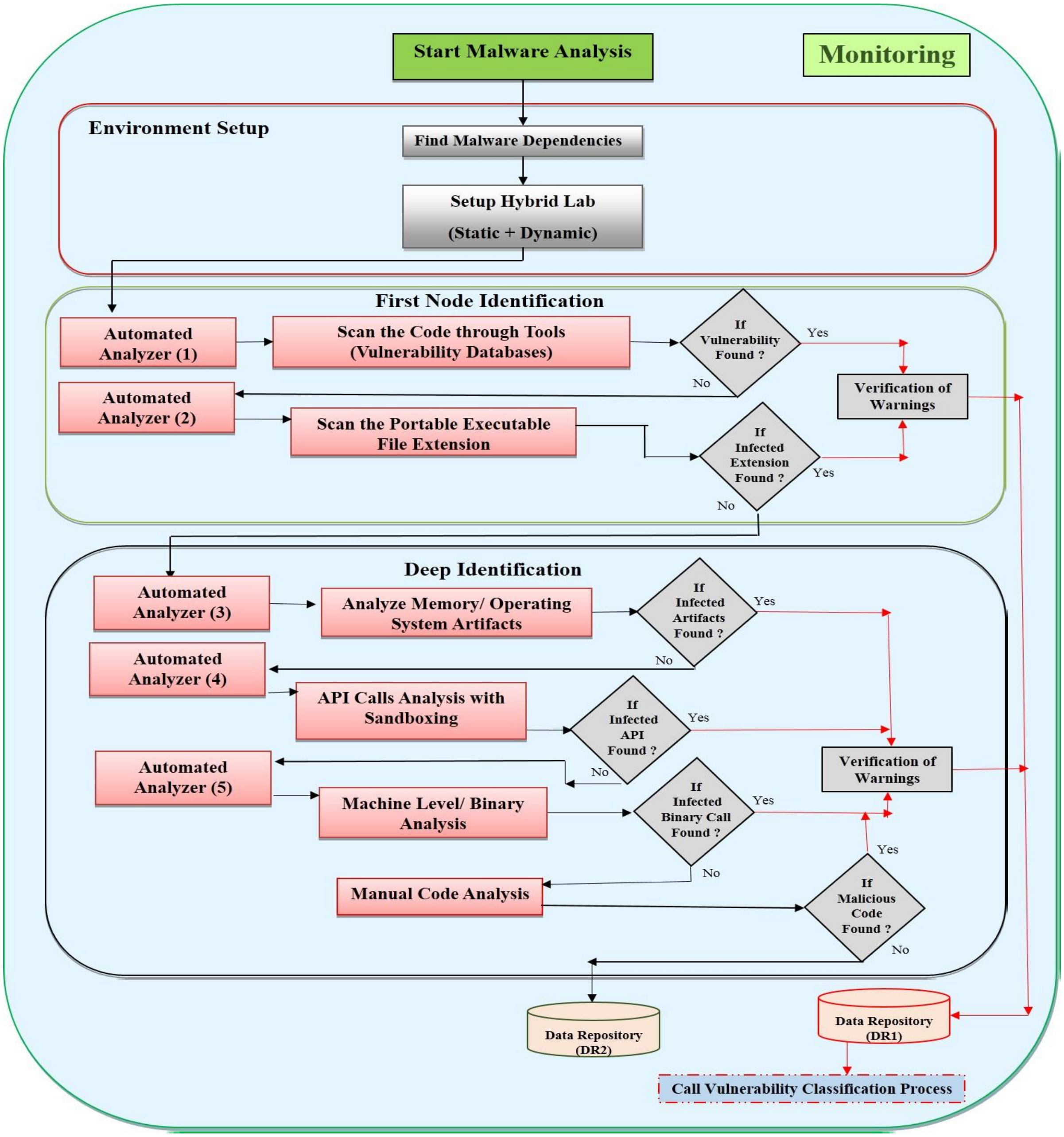

Start Malware Analysis
Monitoring
Environment Setup
Find Malware Dependencies
Setup Hybrid Lab
(Static + Dynamic)
First Node Identification
Automated Analyzer (1)
Scan the Code through Tools (Vulnerability Databases)
If Vulnerability Found ?
Yes
No
Verification of Warnings
Automated Analyzer (2)
Scan the Portable Executable File Extension
If Infected Extension Found ?
Yes
No
Deep Identification
Automated Analyzer (3)
Analyze Memory/ Operating System Artifacts
If Infected Artifacts Found ?
Yes
No
Automated Analyzer (4)
API Calls Analysis with Sandboxing
If Infected API Found ?
Yes
No
Verification of Warnings
Automated Analyzer (5)
Machine Level/ Binary Analysis
If Infected Binary Call Found ?
Yes
No
Manual Code Analysis
If Malicious Code Found ?
Yes
No
Data Repository (DR2)
Data Repository (DR1)
Call Vulnerability Classification Process

Fig. 1. Monitoring Phase.

1) Environment setup: First sub-phase of the monitoring phase is environment setup. In this sub- phase, the authors set up an environment for executing the analysis process. Dynamic analysis of malware is always done under some restricted environment for a better and secure outcome. The dynamic analysis deals with malware at motion. The following are the processes that an examiner takes while setting the analysis environment.

- Find Malware Dependencies: It is very important in dynamic analysis to run all features and services of malware for understanding the objective and finding the vulnerabilities clearly. So it is important to find malware dependencies and install them in the lab to perform dynamic analysis process.
- Setup Hybrid Lab (Static + Dynamic): After finding dependencies, set up a hybrid lab which is a mixture of the static and dynamic lab for further analysis. The static analysis deals with malware at rest, it means in this process malware is not executed on the system. Static analysis is fully secure and harmless examination process of malware, but dynamic analysis deals with malware at motion. In dynamic analysis, malware is executed on the system under a controlled environment. Reverse engineering is complementary for dynamic analysis.

2) First node identification: Second sub-phase of the monitoring phase is first node identification. This phase deals with some common methods to find out malware vulnerabilities. This phase uses signature-based identification methods for recognizing old malware classes. Steps that are taken in this sub-phase are enlisted below:

- Scan the Code through Tools (Vulnerability Databases): In this step we scan the code through various old vulnerability database (by tools) for finding the match. If the vulnerability is found then we verify the warning manually and if manual verification is also found to be yes, then we save that vulnerability into Data Repository (DR1) and, if not, then we go for the next step which is scanning the portable executable file extension.
- Scan Portable Executable File Extension: In this step, we scan portable executable file extension by various tools for finding the infected extension and if the infected extension is found by a tool, we verify the warning manually. If a warning is yes, then we save the vulnerability into DR1 otherwise we go to the next sub-phase which is Deep Identification.

3) Deep identification: Third sub-phase of monitoring phase is deep identification. In this phase, the examiner analyzes the malware by various industry level professional methods and finds the vulnerabilities. Steps that are taken in this sub-phase are:

- Analyze Memory/Operating System Artifacts: In this step, experts analyze memory/operating system artifacts both manually and by the tools. If something is detected, the examiner verifies the warning first. If the warning is yes, he saves that vulnerability into DR1 and if it is no, then the expert proceeds to the next step which is API Calls Analysis with Sandboxing.
- API Calls Analysis with Sandboxing: In this step, the expert uses sandboxing tools and with the help of that tool the examiner analyzes API calls for malicious API's. If the tool finds any malicious API then it blinks the warning and after that the expert verifies the warning. If the warning is true then the expert saves it to DR1. If false, then the next step begins which is Machine Level/Binary Analysis.
- Machine Level/Binary Analysis: After using all static and dynamic method in the last automated analysis, examiner uses reverse engineering method for finding vulnerabilities in the code. In this step, the expert uses reverse engineering malware analysis tools for finding the malicious binary calls. If the tool shows the warning, the developer verifies that warning with an expert. If a warning is yes then the expert saves that vulnerability into DR1 repository. If not, then he goes for next step which is the Manual Code analysis.
- Manual Code Analysis: In this step, the examiner analyzes the malware code and finds vulnerabilities and malicious piece of code manually. If the analyst finds malicious code or vulnerability, he calls for superior checking (verify warning) and if the senior coding expert will verify the warning to be true, then the analyst saves that vulnerability in DR1. Should it be false, then he saves this code into a new repository called the Data Repository DR2 as a vulnerability-free code.

B. Phase 2: Mitigation Phase

This phase is to mitigate the detected/identified vulnerabilities in the codes. The step-by-step process of mitigation of vulnerabilities is depicted and elucidated in Fig. 2:

1) Vulnerability classification: In this step, the analyst classifies the vulnerabilities that are discovered in previous phase. Afterwards, the analyzer goes for the next step which is measuring the Priority of vulnerabilities (Quantitatively).

2) Measuring the priority of vulnerabilities: In this step, the examiner evaluates the priority of vulnerability, quantitatively and mitigates these vulnerabilities according to their severity level. If severity level of the vulnerability is high then the analyst removes it. If the severity level of the vulnerability is medium, then the analyst repairs the codes. If the severity level of the vulnerability is low, then the analyst tries to fix the issue. After mitigating the vulnerability issues in the codes, the examiner saves the code in Data Repository (DR3) and calls for the managing phase.

C. Phase 3: Managing Phase

Objective of this phase is to manage mitigated code (DR3) and vulnerability-free code (DR2) for future reuse. Management of extracted code is very necessary because while a coder uses an old code in reuse, it is always a challenge for the programmer to manage that code5. This phase will help the coders in the industry by reducing their work (Managing Code) slightly. The authors categorize the monitoring phase into three sub-phases that are shown in Fig. 3 and are described as follows:

1) Maintaining the codes (As per requirement specification): In this step, examiner follows recent trending process of software development industry which is also called managing code. With the help of a good coder, an examiner manages the codes from (DR2) & (DR3) and after successful management of code, the expert saves the managed code into Data Repository (DR4).

2) Verification of the functionality: After successful management of code, the expert verifies the functionality of the code. If the analyst finds any functionality issue, then the analyst directly calls for maintenance of the coding process, and if no issue is found then the examiner goes for next step which is Measuring the Complexity of Design.

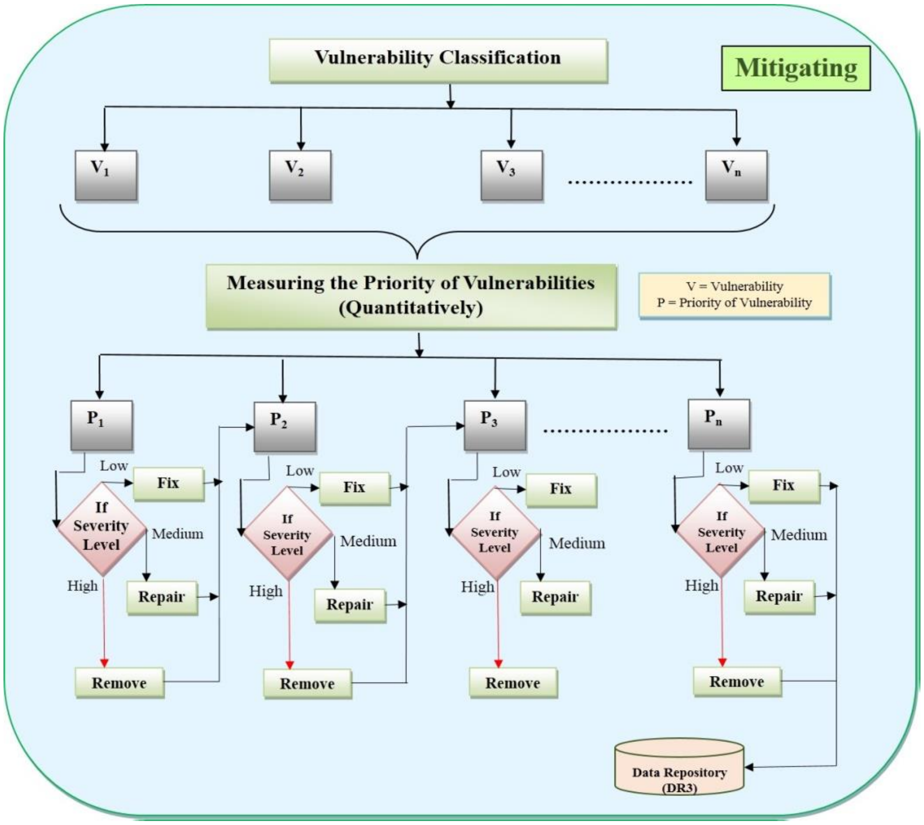

Vulnerability Classification
Mitigating
V1
V2
V3
Vn
Measuring the Priority of Vulnerabilities (Quantitatively)
V = Vulnerability
P = Priority of Vulnerability
P1
P2
P3
Pn
Low
Fix
If Severity Level
Medium
Repair
High
Remove
Data Repository (DR3)

Fig. 2. Mitigating Phase.

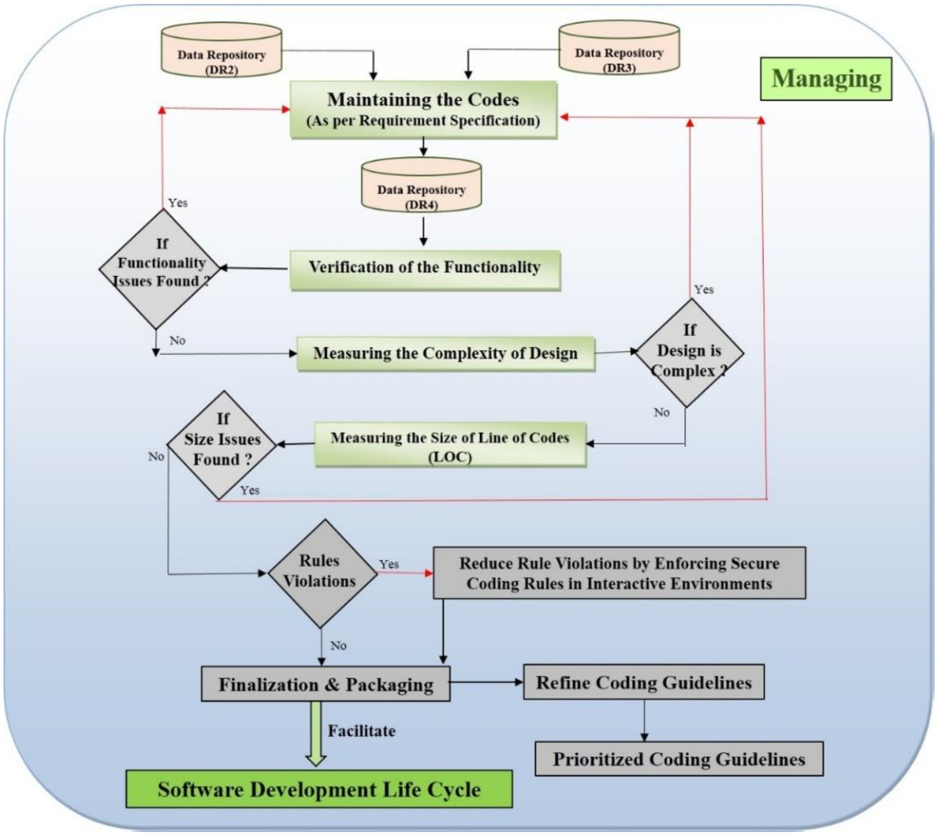

Data Repository (DR2)
Data Repository (DR3)
Managing
Maintaining the Codes (As per Requirement Specification)
Data Repository (DR4)
Yes
If Functionality Issues Found ?
Verification of the Functionality
No
Measuring the Complexity of Design
If Design is Complex ?
Yes
No
If Size Issues Found ?
Measuring the Size of Line of Codes (LOC)
No
Yes
Rules Violations
Yes
Reduce Rule Violations by Enforcing Secure Coding Rules in Interactive Environments
No
Finalization & Packaging
Refine Coding Guidelines
Facilitate
Prioritized Coding Guidelines
Software Development Life Cycle

Fig. 3. Managing Phase.

3) Measuring the complexity of design: In this step, the analyst assesses the complexity of the design of code and if the examiner finds any issue in the complexity of design, then the examiner directly calls for maintaining the coding process. Otherwise the next phase is followed which is measuring the Line of Codes (LOC).

4) Measuring the size of Line of Codes (LOC): In this step, the analyst measures the line of codes for assessing the size of code. If the examiner finds any size issue in the code, the expert directly calls for maintaining the coding process and if no issue is found then the expert goes for next step.

5) Rule violation: In this step, the expert checks the rule violation of code, if the result is yes, and rules are violated more than acceptance, the expert reduces rule violation by enforcing the Secure Coding Rules in interactive environment. After this process, the examiner goes for Finalization & Packaging step and if minimum rules are violated, they are acceptable. So, the examiner goes for the next step called-Finalization & Packaging.

6) Finalization and packaging: In this step, the expert finalizes the code and prepares it for use by facilitating it with software development life cycle. This process helps the industry developers in their projects by providing ready to use managed codes.

7) Refine coding guidelines: This step is for coders who are interested in writing secure codes. In this step, the examiner provides the guidelines for a coder for writing secure code after analyzing the full malicious code.

8) Prioritize the guidelines: This step will help the coders to understand the provided guidelines easily by arranging the guidelines according to their priority or need in programming.

VI. Significance of the Framework

Signature-based identification of malware was very useful and effective for last 10 years but in the current scenario, Corrado Aaron Visaggio and his team from Italy developed an engine that alters and modifies the malware code automatically and misinforms the signature-based analyzers [10]. The engine works on the shape of the malicious code, not on the behavior of the code. This sort of improvement creates the need for a full bundle with the blend of each of the three investigation forms and furthermore needs to take a shot at the code for malware analysis. The framework is providing all the necessary requirements that are needed in the current situation of malware analysis and software industries.

The framework is focusing on the clear and perfect vulnerability identification mechanism with the help of malware analysis techniques. For mitigating these vulnerabilities, the framework uses prioritization and severity assessment methods. After mitigation comes managing and for this the secure code framework is produced which manages the code. Thereafter, an expert programmer then assesses the complexity, reliability and size of code for easy reusability. After all these steps, the framework provides the guidelines for future developers and facilitates the produced code into the software development life cycle for further uses. If we look at this framework deeply, it is a full bundle supply of ready to use codes. The framework provides the following features for developers and researchers.

- The framework provides ready to use, a maintained code for developers for their existing projects, if the code is compatible with their project.
- The framework gives well-structured and accurate malware analysis procedure for finding code vulnerabilities.
- The framework is able to identify the malicious codes and mitigate these vulnerabilities. Furthermore, it produces secure code for industry reuse.
- The framework provides the procedure for providing secure reused codes with the help of three-phase framework and creates an easy approach for the coder to reuse code.
- The framework also provides the time feasible method for identification, mitigation and managing the malicious code.

VII. Conclusion and Future Work

Malware coders are attempting to increase their area of infection and impact of harm very massively. Evidently, security mechanism of web is penetrated on a daily basis with huge number of malware attacks occurring every day. Advancement of malicious codes on a daily basis is creating big gap in old identification and examination methodologies for malware. Besides this, a large number of software is also creating the challenge for coder in development of new logics and functions every day. This kind of challenge has increased the significance of reuse codes in the industry. The framework is shown in Fig. 4. It maps the phase-wise steps to produce secure codes from malicious code with the help of malware analysis. The framework will help in identification of malware and then mitigating the malicious vulnerabilities, moreover managing, mitigating, securing, and producing no vulnerable code for industry reuse. Successful implementation gives the direction for future analysis and suggests the guidelines for coders. The implementation of the intended framework will help the researchers to develop a useful and reliable strategy for producing or writing secure codes for future work on this proposition.

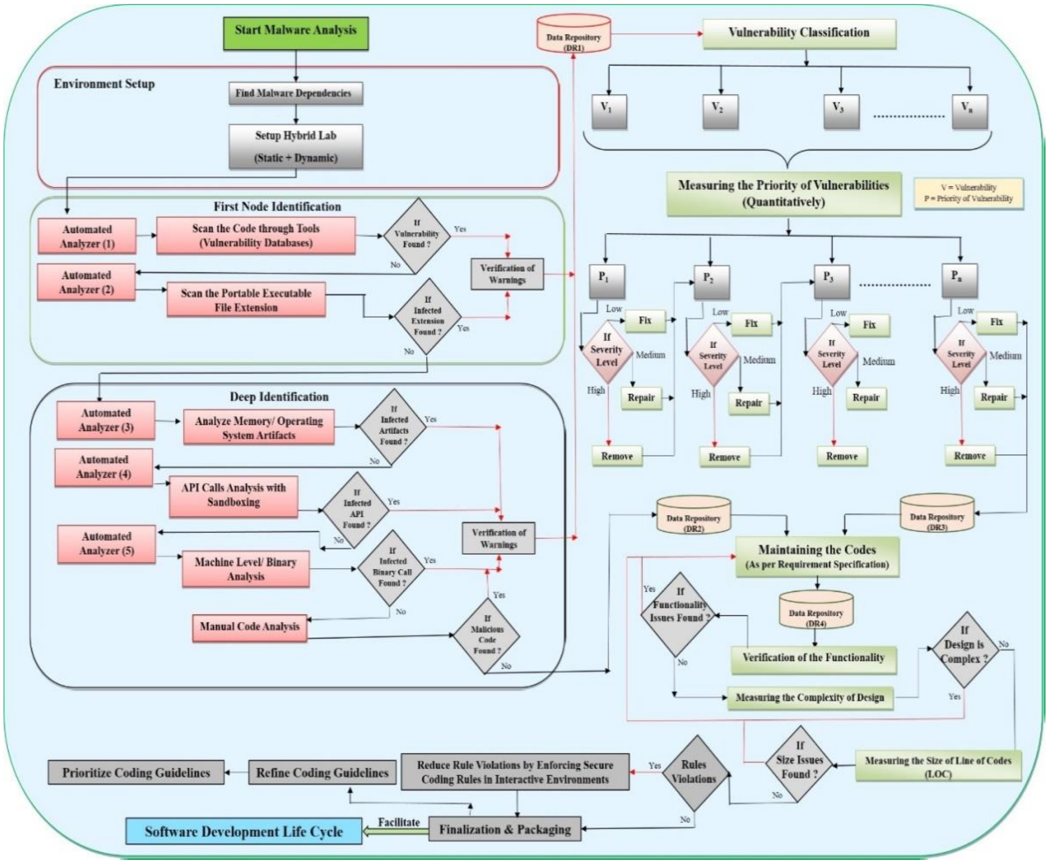

Start Malware Analysis
Data Repository (DR1)
Vulnerability Classification
Environment Setup
Find Malware Dependencies
Setup Hybrid Lab (Static + Dynamic)
V₁
V₂
V₃
Vₙ
Measuring the Priority of Vulnerabilities (Quantitatively)
V = Vulnerability
P = Priority of Vulnerability
First Node Identification
Automated Analyzer (1)
Scan the Code through Tools (Vulnerability Databases)
If Vulnerability Found ?
Yes
No
Verification of Warnings
Automated Analyzer (2)
Scan the Portable Executable File Extension
If Infected Extension Found ?
Yes
No
P₁
P₂
P₃
Pₙ
Low
Fix
If Severity Level
Medium
High
Repair
Remove
Deep Identification
Automated Analyzer (3)
Analyze Memory/ Operating System Artifacts
If Infected Artifacts Found ?
Yes
No
Automated Analyzer (4)
API Calls Analysis with Sandboxing
If Infected API Found ?
Yes
No
Verification of Warnings
Automated Analyzer (5)
Machine Level/ Binary Analysis
If Infected Binary Call Found ?
Yes
No
Manual Code Analysis
If Malicious Code Found ?
Yes
No
Data Repository (DR2)
Data Repository (DR3)
Maintaining the Codes (As per Requirement Specification)
Data Repository (DR4)
If Functionality Issues Found ?
Yes
No
Verification of the Functionality
Measuring the Complexity of Design
If Design is Complex ?
Yes
No
Measuring the Size of Line of Codes (LOC)
If Size Issues Found ?
No
Rules Violations
Yes
No
Reduce Rule Violations by Enforcing Secure Coding Rules in Interactive Environments
Refine Coding Guidelines
Prioritize Coding Guidelines
Finalization & Packaging
Facilitate
Software Development Life Cycle

Fig. 4. Framework for Producing Secure Code through Malware Analysis.

ACKNOWLEDGMENT

Authors are thankful to the College of Computer and Information Sciences, Prince Sultan University for funding this research study.

REFERENCES

[1] Pandey, A. K., Tripathi, A. K., Kapil, G., Singh, V., Khan, M. W., Agrawal, A., Kumar, R., & Khan, R. A. Trends in Malware Attacks: Identification and Mitigation Strategies. In M. Husain, & M. Khan (Eds.), Critical Concepts, Standards, and Techniques in Cyber Forensics (pp. 47-60). Hershey, PA: IGI Global, 2020.

[2] All about malware, Available at: https://www.malwarebytes.com/malware/.

[3] CyberSecurity Statics and Facts For 2017- 2018, Available at: https://privacy.net/cybersecurity-statistics/.

[4] Source Code, Available at: https://www.techopedia.com/definition/547/source-code.

[5] The Challenges of Code Reuse (How to Reuse Code Effectively), Available at: https://www.perforce.com/blog/qac/challenge-code-reuse-and-how-reuse-code-effectively.

[6] Why Code Reuse Matters, Available at: https://www.apress.com/de/blog/all-blog-posts/why-code-reuse-matters/15477476.

[7] YuheiKawakoya, EitaroShioji, Makoto Iwamura, Jun Miyoshi. Taint-Assisted Sandboxing for Evasive Malware Analysis. Journal of Information Processing; Vol.27 297-314, 2019

[8] Kamlakant Sethi, Shankar Kumar Chaudhary, Bata Krishna Tripathy, Padmalochan Bera. A Novel Malware Analysis Framework for Malware Detection and Classification using Machine Learning Approach. 19th International Conference on Distributed Computing and Networking, Varanasi. 2018.

[9] Cybercrime: 25% Of All Malware Targets Financial Services, Credit Card Fraud Up 200%, Available at: https://www.forbes.com/sites/zakdoffman/2019/04/29/new-cyber-report-25-of-all-malware-hits-financial-services-card-fraud-up-200/#7f4932eb7a47.

[10] A Group of the researchers from the Iswatlab team at the University of Sannio demonstrated how is easy to create new malware that eludes antimalware, Available at: https://securityaffairs.co/wordpress/51714/malware/evading-antimalware.html

[11] Li Li, Tegawende F. Bissyande, Mike Papadakis, Siegfried Rasthofer, Alexandre Bartel, Damien Octeau, Jacques Klein, Yves Le Traon. Static Analysis of Android Apps: A Systematic Literature Review. Journal of Information and Software Technology Elsevier; Vol. 88 pp 67-95, 2017.

[12] Christian CamiloUrcuqui Lopez, Andres Navarro. Framework for Malware Analysis in Android, Sistemas&Telemática, 14(37), 45-56, 2016.

[13] Belal Amro. Malware Detection Techniques For Mobile Devices, International Journal of Mobile Network Communications & Telematics, Vol.7, No.4/5/6, 2017.

[14] Suriayati Chuprat, Aswami Ariffin, Shamsul Sahibuddin, Mohd Naz'ri Mahrin, Firham M. Senan, Noor Azurati Ahmad, Ganthan Narayana, Pritheega Magalingam, Syahid Anuar, Mohd Zabri Talib. Malware Forensic Analytics Framework Using Big Data Platform, Springer Nature Switzerland, 881, pp. 261–274, 2019.

[15] G. Hamsa, Deepti Vidyarthi. Study And Analysis Of Various Approaches For Malware Detection And Identification, Vol 1, Issue-10, 2013

[16] Qi Xin and Steven P Reiss. "Better Code Search and Reuse for Better Program Repair". In: Proceedings of the 6th IEEE/ACM International Conference on Genetic Improvement. 2019.

Person Re-Identification System at Semantic Level based on Pedestrian Attributes Ontology

Ngoc Q. Ly[1], Hieu N. M. Cao[2]
Department of Computer Vision and Cognitive Cybernetics
VNUHCM-University of Science
Ho Chi Minh City
Vietnam

Thi T. Nguyen[3]
Computer Vision and Cognitive Cybernetics
VNUHCM-University of Science
Ho Chi Minh City
Vietnam

***Abstract*—Person Re-Identification (Re-ID) is a very important task in video surveillance systems such as tracking people, finding people in public places, or analysing customer behavior in supermarkets. Although there have been many works to solve this problem, there are still remaining challenges such as large-scale datasets, imbalanced data, viewpoint, fine-grained data (attributes), the Local Features are not employed at semantic level in online stage of Re-ID task, furthermore, the imbalanced data problem of attributes are not taken into consideration. This paper has proposed a Unified Re-ID system consisted of three main modules such as Pedestrian Attribute Ontology (PAO), Local Multi-task DCNN (Local MDCNN), Imbalance Data Solver (IDS). The new main point of our Re-ID system is the power of mutual support of PAO, Local MDCNN and IDS to exploit the inner-group correlations of attributes and pre-filter the mismatch candidates from Gallery set based on semantic information as Fashion Attributes and Facial Attributes, to solve the imbalanced data of attributes without adjusting network architecture and data augmentation. We experimented on the well-known Market1501 dataset. The experimental results have shown the effectiveness of our Re-ID system and it could achieve the higher performance on Market1501 dataset in comparison to some state-of-the-art Re-ID methods.**

Keywords*—*Person Re-Identification (Re-ID); Pedestrian Attributes Ontology (PAO); Deep Convolution Neuron Network (DCNN); Multi-task Deep Convolution Neuron Network (MDCNN); Local Multi-task Deep Convolution Neuron Network (Local MDCNN); Imbalanced Data Solver (IDS)

I. Introduction

Re-ID is the problem of recognising and associating a person at different physical locations over time after the person had been previously observed visually elsewhere. Solving the Re-ID problem has gained a rapid increase in attention in both academic research communities and industrial laboratories in recent years. It has many applications, such as tracking people across cameras, images retrieval, or customer behavior analysis [1]. Due to using appearance features from input images, this problem suffers from the common challenges in visual recognition: illumination, pose variation, occlusion, intra-class and inter-class variations. Early studies aim to make full use of hand-crafted visual features [2-8] or metric learning [2,4,5, 9-11] to build a best descriptor for each person. These methods can solve one or some of the above challenges, but are very computational expensive and do not reach high results of accuracy. In recent years, with the growth of convolutional neural networks (CNNs) and careful-annotated benchmarks, CNN-based models which learn deep features from data outperform hand-crafted methods by a large margin and achieve remarkable accuracy [12]. To obtain more discriminative features to deal with the inter-class challenge, some works try to extract local features from local regions in different ways, such as pose normalization [13-15], part-based learning [16-19], or attention mechanism [20-23]. Although these great works gain very high performance in accuracy and mAP, they are still only employed deep features, which do not contain semantic information and cannot be explained by human.

Pedestrian Attribute Recognition is a task that predicts a number of predefined attributes describing a pedestrian. Similar to Re-ID, this task takes bounding boxes of pedestrian captured by cameras as inputs. Attributes are semantic information. They are extracted based on attribute learning model. They could be robust to challenges such as pose, lighting, camera characteristics. According to Re-ID problem, facial attributes and cloth attributes are considered. Combining a set of large enough attributes can help improve the discrimination of Re-ID features. Furthermore, unlike low-level visual features or high-level deep features, attributes are easy to understand for human [24]. Attributes can also be expanded to a range of other applications, such as clothes retrieval, face retrieval. Most existing Re-ID studies use global features to predict all attributes [24-27]. However, most attributes appear in local positions, so global features are insufficient to recognize them. Some works notice this drawback and improve by divide global features into local parts [28,29], but they still consider attributes as an auxiliary branch to enrich deep features.

In this work, we proposed two simple CNN-based models, one for extracting deep global features, and the other for predicting pedestrian attributes. In the learning stage of the attribute recognition model (ARM), we split feature maps at a specific mid-level layer into multiple branches, with respect to human's body parts. Each branch use a local feature map, which is horizontally split from global feature map, to predict a group of corresponding relevant attributes. The attributes groups are applied from a predefined PAO, which can help leverage the intra-class correlation of attributes into the learning process. Besides, we take into consideration the imbalance of attributes and handle it by employing the Matthews correlation coefficient (MCC). In the inference stage, different from previous methods, for each query image, we firstly use attributes prediction to filter out mismatch candidates, and then the remaining ones will be used to find out best matching by deep global features.

The main contributions of this paper are as follows: 1) We propose the Pedestrian Attribute Ontology to conduct Pedestrian Attribute Learning Process and Re-ID process; 2) We propose the Pedestrian Attribute Learning Model based on Local Multi-task learning; 3) We propose integrating Imbalanced Data Solver based on MCC to Re-ID system; 4) We propose new Re-ID method based on Deep Global Features and Pedestrian Semantic Information.

II. Related Works

A. *Hand-Crafted-Features-based Re-ID*

Traditional approaches mainly focus on designing discriminative visual hand-crafted features. Colors and texture are usually employed. In [2], RGB and HSV color vectors are extracted from input images and then fed into a Maximum Likelihood model to learn the image similarities. In another approach [3], Gabor filters are used to extract texture features. Covariances of these features are also employed by the Region Covariance Descriptor in [6,7]. In [8], a robust features named Local Maximal Occurrence (LOMO) are proposed. LOMO is obtained by sliding a window on input images and taking the maximum values of different features from all patches under the window. Apart from proposing discriminative features, other works [5, 9-11] try to design an effective metric to learn the similarity and difference between images.

B. *CNN-based Re-ID*

CNNs have been widely employed in person re-identification due to their great performance in many different computer vision tasks [30–32]. Earlier works use global features extracted from a CNN to train a siamese network [33–35]. For example, [33] proposed a Deep Ranking model aiming to maximize the rank of Euclide distance of same identity's feature vectors. Author in [34] employ Recurrent Neural Network to make use of motion information for more discriminative person description. Author in [35] proposed a Pyramid Person Matching Network to learn the correspondence of misalignment components in image pairs. Attention mechanisms are applied in many models [20-23] to focus on salient parts to extract more useful and discriminative information. Recent studies start to consider part-level features as complementary features for their models due to its fine-grained information. Early part-based approaches apply predefined rigid grids on input images as local parts [36,37]. This way of partition is insufficient because person detection boxes are not always correct. In a very detail partition way, [38] train a semantic parsing model to localize pixel-level body parts. A weighted sum layer is then used to fuse global and local features for identity classification. Extra pose estimators [13,14] or spatial constraints [15] are also utilized to normalize deformable pedestrian parts to obtain more robust features. In another way to learn local features [16-18], global features are horizontal pooled and then separated into vectors, each of them is considered as containing information of a correspond body part. Author in [17] do the same methods but splitted global features into equal stripes in multiple granularities. Each stripe is then adopted separately to an identity classification. Many of the above methods achieve remarkable performance in Re-ID task. However, none of them consider semantic information, such as attributes, but only try to make robust deep global/local features.

C. *Attribute for Re-ID*

Attributes are signatures of concepts. It is difficult to recognize concepts, but recognizing signatures is much easier. Attributes help re-identify pedestrians from coarse to fine, and help to understand images in a detail level. Therefore, attributes can help increase the discrimination between pedestrians. There are many works investigating attributes as auxiliary information to Re-ID. In [24,25], a DCNN classifier is first trained on an independent attributes annotated dataset, then the attributes predicted by that model is used to train and fine-tune another person re-id model. It can be noticed that this kind of methods would push the error of former attribute model to the Re-ID model, especially the attribute model still do not consider the imbalance problem. In [26,27], an end-to-end Multi-task DCNN model is proposed to do both attribute recognition and Re-ID tasks simultaneously. In these works, each attribute probability is predicted by forwarding a same global feature vector to a corresponding linear layer. This vector is also used to retrieve nearest neighbors in inference stage of Re-ID task. In fact, many attributes just appear in small regions on human body, so using a unique global feature vector to learn all attributes is inefficient. Recognizing this drawback, [28,29] proposed part-based CNN models, in which a global feature map from a middle layer is horizontally split into 4 disjoint equal local feature maps, each one is then forwarded to some other convolutional layers followed by a last linear layer to predict probabilities of a group of attributes. The improvement in this method is that it use multiple local features to predict groups of suitable attributes. However, some attributes are distributed over more than one part, so it is confused to choose output of which part to evaluate attributes recognition. Therefore, in inference stage of Re-ID task, the authors do not use attributes predictions anymore, but only enrich features by concatenating all deep local and global features. Furthermore, none of the above methods handle the imbalance data problem of attributes, which is an inherent problem in many classification tasks.

Therefore, in this paper, followed the methods in [28], which is to build a DCNN model that split middle global feature map into multiple local parts. But instead of distributing each attribute over multiple parts and concatenating local and global deep features, we proposed novel methods for improvement: 1) We build a Pedestrian Attribute Ontology (PAO) for better attributes learning, and also for easily expanding in the future; 2) Based on PAO, we build a Local Multi-task DCNN model (Local MDCNN) to exploit inner group and inter group correlations between attributes; 3) We incorporate an Imbalanced Data Solver (IDS) to our Pedestrian Attribute Recognition module; and 4) we build a novel Person Re-identification system flexibly combining global deep features and pedestrian semantic information (facial and cloth attributes).

III. Method

Briefly, our contribution is a novel Person Re-identification system based on Deep Global Features and Pedestrian Attributes. In this section, we focus on the main points of our system. In the off-line stage, we build a PAO and then a Local MDCNN to learn pedestrian attributes. Besides, we use transfer learning to train a siamese network to learn person global deep features. In the on-line stage, for each query image, we firstly use pedestrian semantic information, so-called attribute, to pre-filter candidate images, and then use global deep features to find nearest neighbor in the remaining ones, or vice versa.

Our Re-ID plays a very important role in a multi-camera tracking person system. In each viewing range of camera, the tracking task can be performed by the conventional methods, but when the person moves from view range of camera (i) to camera (i+1), Re-ID is very useful to identify the monitored person is being lost track.

Our system is organized into two phases: *1) Offline Phase:* This phase is designed to build the PAO to support the Deep Global Features Learning model (DGFL model) and the Pedestrian Attributes Learning model (PAL model, aka the Local MDCNN model). In order to improve the performance of learning models, we take into account the imbalanced data problem, and to prepare features for gallery set for matching process in the online phase. The PAO is manually built based on domain knowledge in the field of fashion attributes and facial attributes. It is a hierarchical semantic tree. Its purpose is to exploit the correlations between attributes for not only better learning but also easier expanding in the future. The DGFL model is designed and trained to extract the deep global features of the input image of each person. The PAL model is designed and trained to extract the predefined attributes of each person. The deep features and the semantic information such as attributes are then mutually combined in the online phase. The imbalanced data problem also is solved in this phase by the IDS. Its purpose is to find the best thresholds for each attribute prediction in the training set. In the online phase, the best chosen thresholds are used to convert continuous predicted outputs of PAL model to corresponding binary values to use for the query process. *2) Online Phrase:* This phase is organized to run the query process including deep features extraction, attributes information extraction and retrieval. After getting deep features from DGFL model and attributes information from PAL model, for each query image, we firstly use attributes to filter out candidate images with different attributes of the query image, and then use global deep features to find nearest neighbor in the remain ones, or do the steps vice versa.

A. *Pedestrian Attribute Ontology (PAO)*

Our PAO is inspired by our Face Attribute Ontology (FaAO) and our Fashion Attribute Ontology (FasAO) [39]. PAO helps us to exploit inner group and inter group correlations between attributes, it is then very useful for training the Local MDCNN. PAO also helps the developer to easily update new attributes in the future. To do this, we firstly manually classify attributes into groups based on their corresponding correlations. These groups are combined into a semantic hierarchical tree, also known as attributes ontology. To bring the PAO into deep model, we base on an observation that, most of attributes can be seen at a local region instead of a whole human body. For example, "is wearing hat" can be predicted by learning from "head" region. Therefore, our Local MDCNN learns attributes from local features instead of global ones. Follow [39], we employ a clothing attribute dataset named DeepFashion [40] to build a general FasAO, and our prior knowledge to build a general FaAO. In experiment, we re-build specific PAO on another dataset that has both Re-ID and attribute label.

In [41], Gruber stated that ontology is a formal, explicit specification of shared concepts. An ontology is formed by four principle components: individuals, classes, concepts and relations between them. The class components can has multiple layers. In our case, we formulate person attribute ontology by five components:

- Person (individuals): a layer representing people objects.
- Regions (classes): a layer representing human’s body regions, consisted of five parts: head, upper body, lower body, whole body (upper and lower) and foot.
- Categories (classes): a layer representing types of corresponding clothes in each region.
- Attributes (concepts): clothing attributes with respect to each category and facial attributes.
- Relations: consisted of 3 relations: part of (between regions and individuals), has a (between regions - categories and categories - attributes), is a (between attributes and their values)

The semantic hierarchical tree consisted of three main levels: Regions, Categories and Attributes. Fig. 1 shows our PAO. In Fig. 1, human body firstly is split into five regions. In each region, there are multiple categories of clothing items. And for each item, there are relevant attributes depending on its kind. Basically, it is not too difficult to know which items should be put into which body regions. Here we take some examples from DeepFashion dataset and visualize in Fig. 2 to show some popular kinds of clothing items. The PAO shows two properties of attributes which are inner group correlation and inter group correlation. These two properties help us in the step of designing deep model that are:

- Firstly, when training a deep attribute recognition model, global features are usually be used to predict all attributes. But, in real life, people just need to see a local region to find out attributes related to this region. For example, we can know if a man is wearing hat or not by looking at his head, and do not need to look at the other regions. This is the inter group correlation between attributes. Ontology help us to see which attributes should not go together and therefore should not be predicted in same local features.

- Secondly, there are many attributes having co-appearance relation to each other. For example, a person having beard is usually a man, so the two attributes Is Male and Having beard usually being zeros or ones together. This is the inner group correlation. Ontology help us to group these attributes into same classes, and therefore deep model should predict them in same local features

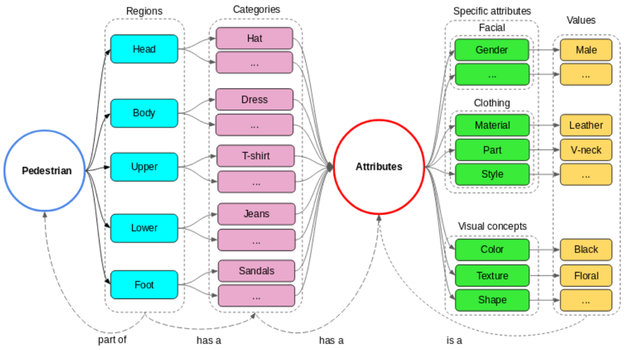


Fig. 1. Our Pedestrian Attribute Ontology.

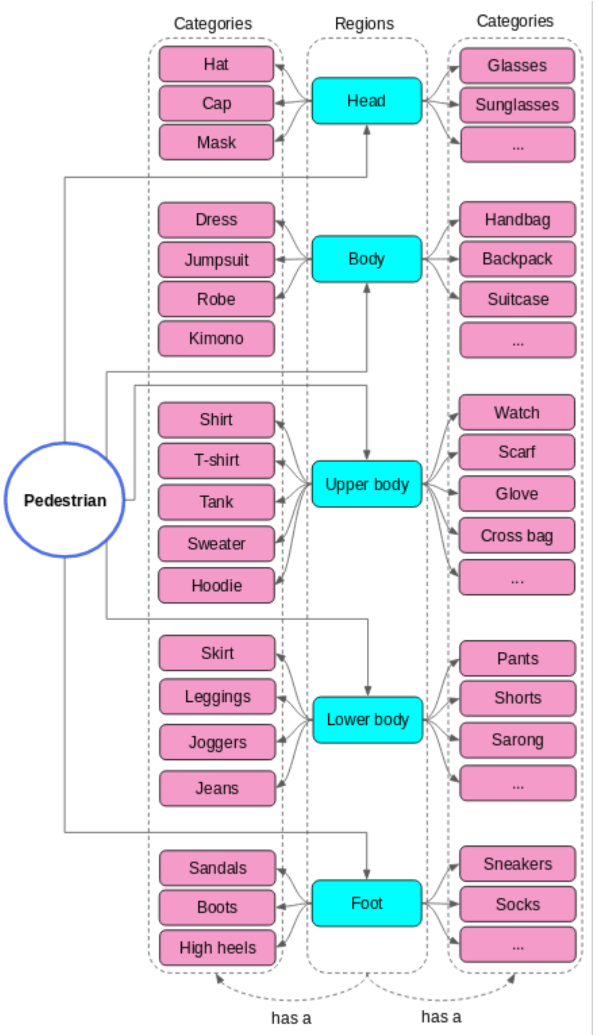


Fig. 2. Some kinds of Clothing Items Extracted from the PAO.

Clothing attributes are very numerous and variety. Follow Ly et al. we choose six types of clothing attributes to demonstrate our ontology, and divide them into two groups: i) general attributes which are attributes that most of items would have, include: color, texture, shape; and ii) specific attributes which are attributes that only exist on some items. Besides, we also show some facial attributes, which can be recognized from the head position. Some examples for clothing attributes and facial attributes are showed in Fig. 3 and Fig. 4, respectively.

In summary, Pedestrian Attribute Ontology is a hierarchical semantic tree, in which attributes are classified into groups. It not only can improve learning process of deep model (in comparison to the models without it), but also easily update more attributes if it is necessary in the future. In the next sub-section, we base on this ontology to design an effective deep model for attribute recognition task.

1) Models: Our system bases on two models: a Person Deep Global Features Learning model and a Pedestrian Attribute Learning model. The former one is trained to extract deep global features vectors and the latter one is trained to extract attributes vectors. Since deep features achieved high performances on many tasks in computer vision, it should not be ignored in our system. However, deep features do not contain semantic information. With only one input image, we cannot understand what do deep features mean, but with attributes features, we can know which attributes exist on the persons in the input images. Therefore, both of the above features can mutually support to get high performance in our system.

a) Person Deep Global Features Learning model: Since we are trying to proof by experiments that attributes prediction can help improve Re-ID results, so we just build a simple person deep global features learning model instead of using complex architecture like other great works. Concretely, we transfer 50-layers Residual Network [32] which was pretrained on the famous image classification dataset ImageNet. In our architecture, we remove the last 1000-units linear layer and append a 1x1 convolutional layer to reduce dimension from 2048 down to 256. This last 256-D vector is the feature vector of the input bounding box, which is then used for matching in the inference stage. Fig. 5 show our Re-ID model architecture in the offline phase.

A couple of images are of the same person when their corresponding deep features have high similarity. To train the model to achieve this goal, in training stage, we use Triplet Loss [42] as our loss function. This function was used by many Re-ID models in particular and image retrieval models in general. Its goal is to learn the similarity between same ID inputs and the divergence between different ID inputs. Equation (1) shows the formula of this function. Whenever the training process is finished, Euclidean dissimilarity distance between feature vectors of images of same person (f^a *and* f^p) should less than those of different person (f^a *and* f^n) by a margin m.

$$\text{TripletLoss} = \max(0, \| f^a - f^p\|_2^2 - \| f^a - f^n\|_2^2 + m \qquad (1)$$

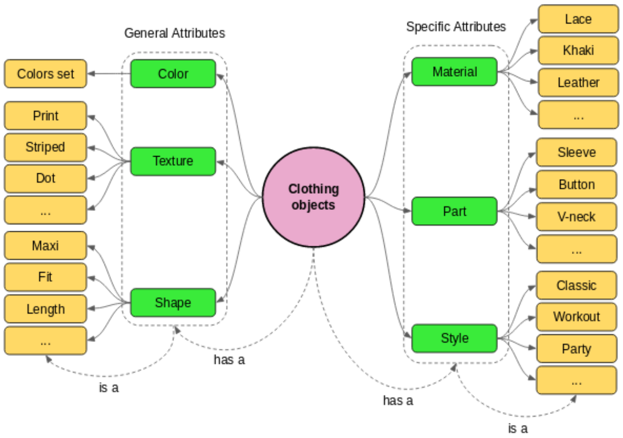


Fig. 3. Some Clothing Attributes with their Values, Extracted from the PAO.

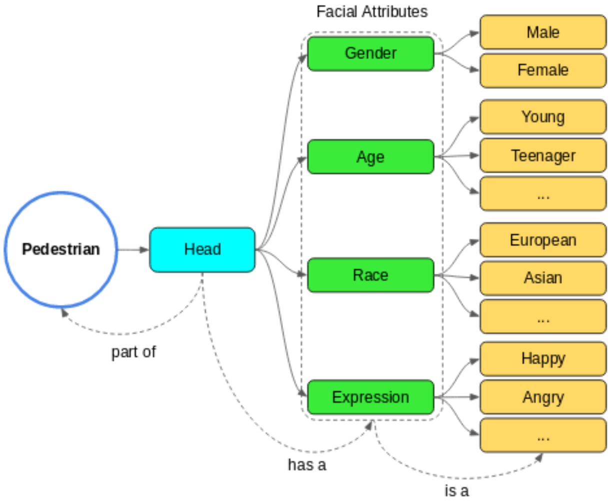


Fig. 4. Some Facial Attributes with their Values, Extracted from the PAO.

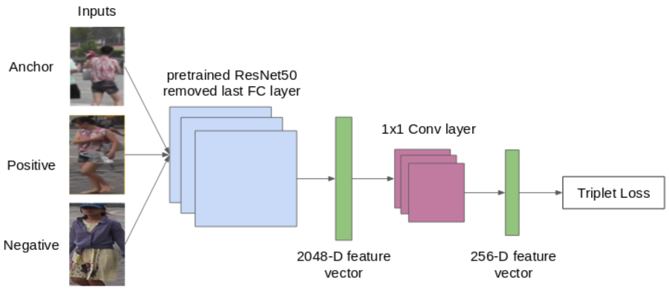


Fig. 5. Person Deep Global Features Learning Model. The Inputs of each Timestep Are a Triple of Images, Including Anchor, Positive and Negative one. The Model Parameters are Shared between them.

In the inference stage, with an input image, its corresponding deep features vectors are extracted and then compared to other pre-extracted vectors of gallery images. The nearest gallery one, i.e. the one has smallest Euclidean dissimilarity distance, would be chosen to match with input image as the same person. With the above strategy, the more similar the two individuals are, the smaller Euclidean dissimilarity distance between the corresponding deep features vectors has. However, this leads to another drawback of deep features: mismatched results would rise in cases of different persons having same appearance (such as same cloths, same pose). In these cases, we need more detail information to distinguish them instead of global deep features only. And pedestrian attributes are semantic information at fine-grained level. Therefore, the attributes are used in our system to filter out false positive candidate images in those cases to get better performance for Re-ID system.

b) Person Attribute Learning model: Pedestrian Attribute Learning model is designed based on Pedestrian Attribute Ontology. From the ontology, to make sure that the inner group and inter group correlation can be leveraged into deep model, we build PAL model with three levels: regions, categories and attributes. Firstly, a middle layer global feature map is split into four equal parts, which means each one occupies 25% ratio height of person body. From top to bottom, the parts respectively are head, upper body, lower body, foot region. The body region is the merger of upper body and lower body. Secondly, each region split features are learnt in a corresponding local sub-network to extract information relevant to that local part. And finally, local features of each part are fed into multiple smaller branches, which then predict the attributes in the PAO.

In briefly, our PAL model has three parts: i) a global part that learns common features; multiple attribute learning sub-networks, consist of: ii) local parts that learn local features, and iii) attribute parts that learn specific features and predict a group of suitable attributes. Fig. 6 shows our PAL model and Fig. 7 show a sub-network from our PAL model, which is taken from the head region. We also transfer 18-layer Residual Network [32] into our architecture. ResNet18 has 5 layer groups. We apply the conv_0, conv_1 and conv_2 to the first part, conv_3 to the second one and conv_4 for the last one, which is also shown in Fig. 6 and Fig. 7.

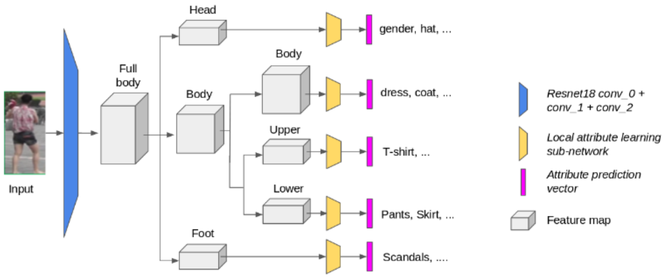


Fig. 6. Our Pedestrian Attribute Learning Model.

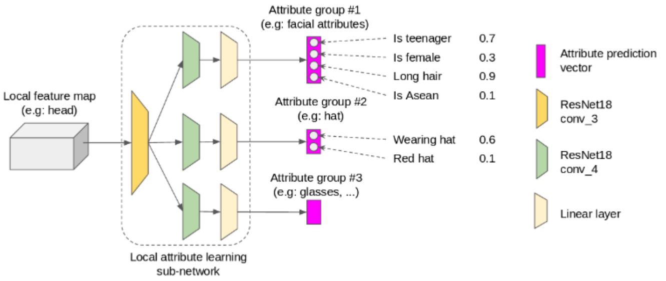

Local feature map (e.g: head)
Local attribute learning sub-network
Attribute group #1 (e.g: facial attributes)
Is teenager 0.7
Is female 0.3
Long hair 0.9
Is Asean 0.1
Attribute group #2 (e.g: hat)
Wearing hat 0.6
Red hat 0.1
Attribute group #3 (e.g: glasses, ...)
Attribute prediction vector
ResNet18 conv_3
ResNet18 conv_4
Linear layer

Fig. 7. A Sub-Network from our Pedestrian Attribute Multi-Task Learning Model, which is Taken from the Head Region.

Outputs of the model are a vector whose number of dimensions is equal to number of binary attributes. In the learning stage, we use Binary Cross Entropy function for each attribute, and get average of those of all attributes as our final loss function. With each M-dimension output vector $\hat{a}_i$ and a M-dimension ground truth vector $\hat{a}_i$, where M is the number of binary attributes, then the average Binary Cross Entropy loss function formula is shown in Equation (2):

$$\text{AvgBCELoss} = -\, 1/M \sum_{j=1}^{M}(a_{ij}\log\hat{a}_{ij} + (1 - a_{ij})\log(1 - \hat{a}_{ij} \quad (2)$$

2) Handling imbalanced data: One of the important improvements of our method is that we incorporate an Imbalance Data Solver into our Person Re-identification system. Imbalanced data is a common problem in classification tasks. There are many ways to handle it, such as: oversampling/ undersampling, use weighted loss function, etc. In the task of multi-label classification like pedestrian attributes recognition, we cannot increase or decrease the amount of samples because it will affect all attributes, and weighting loss will lead to a bunch of hyper-parameters must to be tuned. Therefore, we choose the way of adjusting the thresholds of binary attributes, instead of using common values of 0.5. Concretely, for each attribute, we perform a grid search to choose a best threshold from a list of predefined candidate thresholds. The best one is which we get the highest value of Matthews correlation coefficient [43] when using it to convert from probability to binary prediction. Matthews correlation coefficient (MCC) [43] is a famous metric used to measure the quality of a binary classifier. Its formula takes into consideration the 4 popular values of classification problem: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), which is shown in Equation (3).

$$\text{MCC} = \text{Matthews}_{\text{corrcoef}} = \frac{TPxTN - FPxFN}{\sqrt{(TP+FP)x(TP+FN)x(TN+FP)x(TN+FN)}} \quad (3)$$

Matthews correlation coefficient has value range from –1 to 1:

- It achieves maximum value of 1 when both FP and FN are zeros, which means no sample are false predicted and the classifier result matches exactly the ground truth.
- Contrary, it achieves minimum value of -1 when both of TP and TN are zeros, which means classifier result is completely opposite to the ground truth.
- And it is 0 if the prediction is random.

Therefore, we want to choose which threshold makes the coefficient gain as highest score as possible.

IV. Results and Evaluation

A. Data Set

We demonstrate our proposed method on a large Re-ID data set named Market1501 [44]. This data set contains 32668 bounding-box images of 1501 persons, which are captured from six different cameras in front of a supermarket near Tsinghua University, China. The authors divided it into three parts: training set has 12936 images of 751 persons and test set has 19281 images of the other 750 persons. From the test set, from 1 to 6 random images of each individual are chosen to form a query set, and remaining are the gallery set. Re-ID task is performed by matching images in query set with images in gallery set, and return top-k most similar ones.

For the pedestrian attribute recognition task, we use this Market1501-attribute data set, which is basically the Market1501 data set, but the attribute annotations are proposed by other authors Lin et al. There are total 27 attributes, but we only use 25 binary attributes which have proportions of positive samples rate more than 0.5%. Table I show these attributes which are clustered into local positions by us.

B. Using only Person Deep Global Features Learning Model

First, we evaluate our Person Deep Global Features Learning (PDGFL) model when do not use complementary attribute information. We train our PDGFL model in 60 epochs, by Adam optimizer algorithm, with default hyper-parameters, except the learning rate is set to 3.10─4. Input images are rescaled to 192x96 before fed into the network. Feature vector dimensions are 256. We evaluate 3 version of Residual Network [32]: ResNet18, ResNet50 and ResNet101. Results are reported in mAP, top-1, top-5 and top-10 accuracy, which are shown in Table II. In our experiments, ResNet101 achieves highest performance. ResNet50's result is less than ResNet101's by a very small gap, but it has only a half of number of layers compare to ResNet101. This means even more complex and deeper model still cannot distinguish similar appearance individuals.

C. Attribute Recognition Model

Attribute recognition model is the principle component of our proposed methods. We demonstrate our model in 2 scenarios: without/with Ontology, without/with Matthews correlation coefficient (both are using Ontology). In all attribute recognition experiments, we use the same train-validation-test split and optimizer algorithm as Re-ID experiment, except number of epochs is now set to 10.

Firstly, we re-build the attribute ontology on Market1501-attribute data set. The ontology is shown in Fig. 8.

TABLE. I. 25 Attributes from Market1501-Attribute Data Set

Position	Attribute
head	gender, hair length, wearing hat
body	carrying backpack, carrying handbag, carrying bag
upper	sleeve length, 8 colors of upper clothing
lower	length of lower clothing, type of lower clothing, 8 colors of lower clothing
foot	none

TABLE. II. Query Result of different Models without Complementary Attribute Information

Model	Top-1	Top-5	Top-10	mAP
ResNet18	77.7%	90.6%	93.5%	57.9%
ResNet50	**81.4%**	91.8%	94.7%	**65.1%**
ResNet101	**82.0%**	93.1%	95.6%	**66.0%**

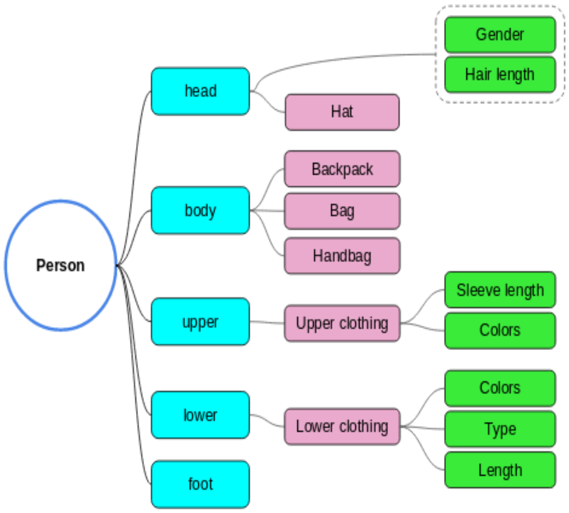


Fig. 8. The PAO Implemented on Market1501-Attribute Data Set.

Secondly, we evaluate our proposed attribute recognition model in four versions:

- Baseline: This is simply a ResNet18 network replaced last 1000-units linear layer by 25-units linear layer. In other words, this model predicts all attributes from a unique global features.
- ONTO: This is the above proposed network, which has multiple branches to predict attributes from suitable local features. But in this version, we still do not handle the imbalance problem.
- ONTO + MCC: This is simply the same as ONTO version, but we handle the imbalance data problem by apply Matthews correlation coefficient in adjusting thresholds. Threshold candidates are predefined by an arithmetic progression from 0.01 to 0.99 with step of 0.01. The best thresholds are then used in test phase to convert continuous outputs to binary values.
- ONTO + MCC + LM: This is the version which has 4 models corresponding to 4 parts: head, body, upper and lower. We observed that, instead of training a unique multi-task model containing all of local branches, training separate models for each local branch make it easy to update new attributes and re-train in the future.

Table III shows the results of 4 versions. We can see some observations:

- With Ontology: As depicted in column Baseline and ONTO, average F1-score increases 12% when using complementary ontology, compare to a plain network. There are 24/25 attribute's which have F1-scores increasing, too. This proofs that using local features, or concretely using ontology, is more powerful than using global features in attribute prediction.
- With Matthews correlation coefficient: As depicted in column ONTO and ONTO + MCC, average F1-score now increases 13% when using MCC. All of attributes have F1-scores increasing, too. Some attributes are improved by a large gap. For example, attribute "wearing hat" has largest growth with 31%, although this attribute positives rate is only 2.6%. This shows that handling imbalance problem by adjusting thresholds help improve significantly the quality of a multi-label classifier.
- With Local multi-task training: As depicted in column ONTO + MCC and ONTO + MCC + LM, average F1-score again increases 13% when training multiple models corresponding to each local part. And all of attributes have F1-scores increasing, too. Moreover, there are many attributes having result greater than 90%, includes: up white, up red, up yellow, lower type, and down black. This shows that, training separate models for each position can also improve the prediction result.

D. Attribute Filtering for Person Re-Identification

1) Compare to the case of using only global deep features: From the results of attribute recognition models in Table III, because of not all of attributes give remarkable scores, instead of use all of them, we only choose five highest F1-score attributes to improve re-id performance. Concretely, for each query image in query set of Market1501 data set:

a) Firstly: We use each of these attributes to filter out candidates in gallery set that mismatch the attributes of query image.

TABLE. III. RESULTS OF DIFFERENT VERSIONS OF PROPOSED MODEL (F1-SCORES)

Position	**Attribute**	**Baseline**	**ONTO**	**ONTO + MCC**	**ONTO + MCC + LM**
Head	gender hair length wearing hat	40.26 51.32 06.74	71.45 65.62 22.82	78.03 73.43 53.42	88.91 86.38 68.26
Body	backpack bag handbag	45.89 43.57 11.40	53.06 46.94 32.47	70.12 53.54 54.57	84.76 77.10 64.77
Upper	sleeve length up black up white up red up purple up yellow up gray up blue up green	11.96 47.63 73.58 71.47 20.02 72.86 21.63 45.84 41.53	44.86 69.53 74.05 73.88 47.74 76.80 52.53 58.23 44.68	66.57 83.06 79.58 88.93 67.91 81.37 68.89 68.91 64.22	74.94 88.75 91.92 90.05 84.99 94.16 82.32 84.92 84.99
Lower	lower length lower type down black down white down pink down yellow down gray down blue down green down brown	59.56 78.78 74.19 48.63 62.77 0.21 51.20 58.77 29.29 43.85	74.41 82.20 83.67 49.35 61.80 11.08 57.92 61.43 44.52 53.61	80.85 89.20 87.85 69.58 66.79 31.34 64.61 64.04 54.96 68.46	83.26 92.16 91.19 78.10 89.07 43.08 78.84 81.34 76.30 82.39
	Average	**43.32**	**56.59**	**69.20**	**81.72**

b) Secondly: the remaining candidates are used to find K nearest neighbors by comparing Euclidean dissimilarity distance of deep features extracted by the above ResNet50 Re-ID model.

The 5 chosen attributes are: up red, up white, up yellow, lower length and down black. Table IV shows attributes pre-filtering results in mAP and top-K. As depicted in Table IV, pre-filtering by attribute down black gives best results in top-K accuracy, and by attribute up red gives best result in mAP. Consider all of these 5 attributes in pre-filtering, although the top-k accuracy values are improved by a small gap, the mAP values increase remarkably, at least 9.3% comparing to the case that does not use attributes in pre-filtering. This indicates that, if in the future we can have better attribute recognition models, so that more attributes have high prediction results, then Re-ID results would be derived to a better performance too.

2) Compare to the case of using global and local deep features: Most of previous works use attributes as auxiliary task for enhancing the global/local deep features, and do not employ attributes prediction in the test stage. Therefore, we perform some experiments to compare the two cases: using global and local deep features with using global deep features and attributes information. Combination of global and local deep features in our experiments are extracted as follow:

a) For each position: we get the output of the local part in our above network (Fig. 6), which is a feature map.

b) Then: we apply a max-pooling operation to convert feature map to a feature vector. This is the local deep feature vector of the corresponding position.

c) Finally: we concatenate deep global features and local feature of one of the 4 regions (head, body, upper, lower) to form a unique deep feature vector and then use it to perform matching process in test stage.

Combination of deep global features and attributes information is exactly the pre-filtering strategy in the previous section. For each position, we compare two cases: i) use all attributes of that position; and ii) use only one attribute with best prediction of that position. Results of the comparisons of all of 4 positions are showed in Table V, the arrows indicates the result of using attribute information is higher or lower than using local features, and the bold values is the highest values between 3 cases in that positions.

As depicted in Table V, when using all attributes in filtering step, top-k accuracies and mAP in all positions are lower than combinations of global and local features. However, it is the opposite for the case of using only one best attribute. In the position head, attribute gender has higher the top-1, top-5 and mAP at about 3-5%, and the top-10 accuracy is smaller only 0.1% compare to using complementary local features. In the position body, attribute backpack is not as good as local features, because it has a not-too-good prediction F1-score, about 84%. In the positions upper and lower, using corresponding best attributes gives performance totally higher than deep local features. Fig. 9 shows some samples that query results are rearranged and improved by attribute filtering.

TABLE. IV. QUERY RESULTS BY ATTRIBUTES PRE-FILTERING USING 5 ATTRIBUTES HAVING BEST PREDICTION

Attribute Pre-filtering	**Top-1**	**Top-5**	**Top-10**	**mAP**
None	81.4%	91.8%	94.7%	65.1%
up red	↑83.3%	↑93.9%	↑92.6%	**↑75.2%**
up white	↑83.6%	↑93.7%	↑95.8%	↑74.4%
up yellow	↑83.8%	↑93.6%	↑95.9%	↑74.9%
lower type	↑84.2%	↑94.1%	↑96.0%	↑74.7%
down black	**↑85.2%**	**↑95.3%**	**↑96.9%**	↑74.8%

TABLE. V. COMPARISON BETWEEN COMBINATION OF GLOBAL AND LOCAL DEEP FEATURES AND COMBINATION OF GLOBAL DEEP FEATURES AND ATTRIBUTE INFORMATION

Case	**Top-1**	**Top-5**	**Top-10**	**mAP**
Position: Head				
Deep Global + Local Feature	78.4%	89.7%	93.9%	56.8%
Deep Global feature + All local attributes	↓75.1%	↓87.2%	↓92.6%	↓53.3%
Deep Global feature + Only attribute *gender*	**↑81.6%**	**↑91.1%**	↓93.8%	**↑61.2%**
Position: Body				
Deep Global + Local Feature	**80.4%**	**89.7%**	94.9%	62.0%
Deep Global feature + All local attributes	↓72.7%	↓79.6%	↓85.1%	↓50.3%
Deep Global feature + Only attribute *backpack*	↓78.3%	↓91.4%	**↑95.5%**	**↑63.1%**
Position: Upper				
Deep Global + Local Feature	80.3%	91.5%	94.8%	61.7%
Deep Global feature + All local attributes	↓77.7%	↓89.4%	↓92.9%	↓59.9%
Deep Global feature + Only *up yellow*	**↑83.8%**	**↑93.6%**	**↑95.6%**	**↑74.9%**
Position: Lower				
Deep Global + Local Feature	78.3%	90.9%	94.5%	61.1%
Deep Global feature + All local attributes	↓69.1%	↓77.4%	↓82.5%	↓47.8%
Deep Global feature + Only attribute *lower type*	↑84.2%	↑94.1%	↑96.0%	↑74.7%

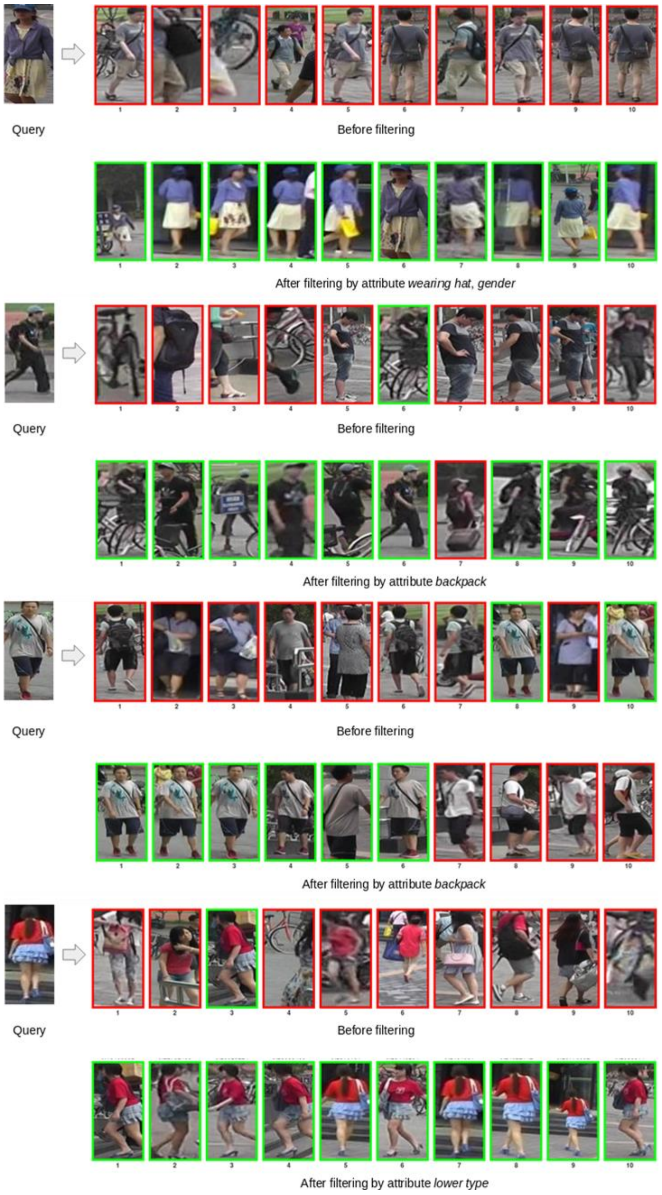

Query
Before filtering
After filtering by attribute *wearing hat, gender*
Query
Before filtering
After filtering by attribute *backpack*
Query
Before filtering
After filtering by attribute *backpack*
Query
Before filtering
After filtering by attribute *lower type*

Fig. 9. Some Samples that Query Results are Improved by Attribute Filtering.

3) Compare to other methods: We use attribute "down black" which has the best performance in improving Re-ID results in comparison with related works. As depicted in Table VI, our method achieves the higher performance than the other works in mAP, top-5 and top-10 accuracy. Note that these works only use attribute as auxiliary information in learning stage. The results show that using attribute as a pre-filter in inference stage can achieve equivalent or even better performance. The methods presented in [24], [26], [28] were selected to compare performance with our method for the following reasons: all these methods have used pedestrian's attributes in Person Re-Identification in different ways but have not yet used attribute pre-filters and considered data imbalance. We would like to show that our method can overcome their drawbacks listed in Section II (part C).

TABLE. VI. COMPARISON WITH OTHER METHODS ON MARKET1501 DATA SET

Methods	**Top-1**	**Top-5**	**Top-10**	**mAP**
Schumann and Stiefelhagen [24]	83.61%	92.61%	95.34%	62.6%
Lin et al. [26]	84.29%	93.2%	95.19%	64.67%
Zhang and Xu [28]	**86.58%**	94.48%	96.73%	68.08%
Ours, pre-filtering by attribute *down black*	85.2%	**95.3%**	**96.9%**	**74.8%**

V. CONCLUSIONS

In this paper, we present a new method using semantic information like as pedestrian attributes to improve person re-identification performance. Our methods is a unified Re-ID system consisted of two main modules: 1) Pedestrian Attributes Learning model (PAO + Local MDCNN + IDS); 2) Person Re-ID model (Deep Global Features based Person Re-ID + Pedestrian Attribute based Person Re-ID). We show that the performance of our Re-ID system is better than some state-of-the-art Re-ID methods. In the future, if more powerful attribute recognition model were proposed, Re-ID task would be driven to a better performance and Re-ID system at semantic level will be integrated to Visual Question Answering (VQA) to improve the intelligence of video surveillance system.

ACKNOWLEDGMENT

This research is funded by Viet Nam National University Ho Chi Minh City (VNUHCM) under grant no. B2018-18-01. Thank to AIOZ Pte Ltd. company for the valuable support on internship cooperation.

REFERENCES

[1] S. Gong, M. Cristani, S. Yan, C. C. Loy, "Person Re-Identification; Springer Publishing Company," Incorporated, 2014.

[2] L. Bazzani, M. Cristani, V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," Computer Vision and Image Understanding 2013, 117, pp. 130–144. doi:10.1016/j.cviu.2012.10.008.

[3] Y. Zhang, S. Li, "Gabor-LBP Based Region Covariance Descriptor for Person Re-identification," Image and Graphics (ICIG), 2011 Sixth International Conference on 2011. doi:10.1109/ICIG.2011.40.

[4] B. Prosser, W. S. Zheng, S. Gong, T. Xiang, "Person Re-Identification by Support Vector Ranking," 2010, Vol. 2, pp. 1–11. doi:10.5244/C.24.21.

[5] M. Gou, F. Xiong, O. Camps, M. Sznaier, "Person Re-Identification Using Kernel-Based Metric Learning Methods," 2014. doi:10.1007/978-3-319-10584-0_1.

[6] W. Ayedi, H. Snoussi, M. Abid, "A fast multi-scale covariance descriptor for object re-identification," Pattern Recognition Letters - PRL 2011, 33. doi:10.1016/j.patrec.2011.09.006.

[7] S. Ba̹k, E. Corvee, F. Bremond, M. Thonnat, "Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid," 2011, pp. 179 – 184. doi:10.1109/AVSS.2011.6027316.

[8] S. Liao, Y. Hu, X. Zhu, S. Li, "Person re-identification by Local Maximal Occurrence representation and metric learning," 2015, pp. 2197–2206. doi:10.1109/CVPR.2015.7298832.

[9] K. Weinberger, J. Blitzer, K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," 2006; Vol. 10.

[10] M. Guillaumin, J. J. Verbeek, C. Schmid, "Is that you? Metric learning approaches for face identification," 2009 IEEE 12th International Conference on Computer Vision 2009, pp. 498–505.

[11] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof, “Large Scale Metric Learning from Equivalence Constraints,” 2012. doi:10.1109/CVPR.2012.6247939.

[12] L. Zheng, Y. Yang, A. G. Hauptmann, “Person Re-identification: Past, Present and Future,” ArXiv 2016, abs/1610.02984.

[13] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, “Pose-Driven Deep Convolutional Model for Person Re-identification,” 2017 IEEE International Conference on Computer Vision (ICCV) 2017, pp. 3980–3989.

[14] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, W. Gao, “Attention driven person re-identification,” Pattern Recognition 2019, 86, 143–155.

[15] D. Li, X. Chen, Z. Zhang, K. Huang, “Learning Deep Context-aware Features over Body and Latent Parts for Person Re-identification,” 2017.

[16] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, J. Sun, “AlignedReID: Surpassing Human-Level Performance in Person Re-Identification,” ArXiv 2017, abs/1711.08184.

[17] Y. Sun, L. Zheng, Y. Yang, Q. Tian, “Wang, S. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline),” ECCV, 2017.

[18] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, Y. Xu, “Deep-Person: Learning Discriminative Deep Features for Person Re-Identification,” 2017.

[19] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, “Learning Discriminative Features with Multiple Granularities for Person Re-Identification,” ACM Multimedia, 2018.

[20] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, “End-to-End Comparative Attention Networks for Person Re-Identification,” IEEE Transactions on Image Processing 2017, 26, pp 3492–3506.

[21] A. Rahimpour, L. Liu, A. Taalimi, Y. Song, H. Qi, “Person re-identification using visual attention,” 2017 IEEE International Conference on Image Processing (ICIP) 2017, pp. 4242–4246.

[22] S. Zhou, J. Wang, D. Meng, Y. Liang, Y. Gong, N. Zheng, “Discriminative Feature Learning with Foreground Attention for Person Re-Identification,” IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 2018.

[23] D. Ouyang, Y. Zhang, J. Shao, “Video-based person re-identification via spatio-temporal attentional and two-stream fusion convolutional networks,” Pattern Recognition Letters 2019, 117, 153–160.

[24] A. Schumann, R. Stiefelhagen, “Person Re-identification by Deep Learning Attribute-Complementary Information,” 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2017, pp.1435–1443.

[25] C. Su, S. Zhang, J. Xing, W. Gao, Q. Tian, “Deep Attributes Driven Multi-Camera Person Re-identification,” ArXiv 2016, abs/1605.03259.

[26] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Y. Yang, “Improving Person Re-identification by Attribute and Identity Learning,” ArXiv 2017, abs/1703.07220.

[27] N. McLaughlin, J. M. del Rincón, P. C. Miller, “Person Reidentification Using Deep Convnets With Multitask Learning,” IEEE Transactions on Circuits and Systems for Video Technology 2017, 27, pp. 525–539.

[28] G. Zhang, J. Xu, “Person Re-identification by Mid-level Attribute and Part-based Identity Learning,” ACML, 2018.

[29] Y. Chen, S. Duffner, A. Stoian, J. Y. Dufour, A. Baskurt, “Pedestrian Attribute Recognition with Part-based CNN and Combined Feature Representations,” VISIGRAPP, 2018.

[30] A. Krizhevsky, I. Sutskever, G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Neural Information Processing Systems 2012, 25. doi:10.1145/3065386.

[31] K. Simonyan, A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv 1409.1556 2014.

[32] K. He, X. Zhang, S. Ren, J. Sun, “Deep Residual Learning for Image Recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, pp. 770–778.

[33] S. Z. Chen, C. C. Guo, J. Lai, “Deep Ranking for Person Re-Identification via Joint Representation Learning,” IEEE Transactions on Image Processing 2016, 25, pp. 2353–2367.

[34] N. McLaughlin, J. M. del Rincón, P. C. Miller, “Recurrent Convolutional Network for Video-Based Person Re-identification,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, pp 1325–1334.

[35] C. Mao, Y. Li, Z. Zhang, Y. Zhang, X. Li, “ Pyramid Person Matching Network for Person Re-identification,” ACML, 2017.

[36] H. Shi, X. Zhu, S. Liao, Z. Lei, Y. Yang, S. Z. Li, “Constrained Deep Metric Learning for Person Re-identification,” ArXiv 2015, abs/1511.07545.

[37] S. Wu, Y. C. Chen, X. Li, A. Wu, J. You, W. S. Zheng, “An enhanced deep feature representation for person re-identification,” 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) 2016, pp. 1–8.

[38] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, M. Shah, “Human Semantic Parsing for Person Re-identification,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018, pp. 1062–1071.

[39] N. Ly, T. Do, B. Nguyen, “Large-Scale Coarse-to-Fine Object Retrieval Ontology and Deep Local Multitask Learning,” Computational Intelligence and Neuroscience 2019, 2019, pp. 1–40. doi:10.1155/2019/1483294.

[40] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, “DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[41] T. R. Gruber, “Toward principles for the design of ontologies used for knowledge sharing,” Int. J. Hum.-Comput. Stud. 1993, 43, pp. 907–928.

[42] F. Schroff, D. Kalenichenko, J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015, pp. 815–823.

[43] S. Boughorbel, F. Jarray, M. El-Anbari, “Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric,” PloS one, 2017.

[44] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, “Scalable Person Re-identification: A Benchmark,” Computer Vision, IEEE International Conference on, 2015.

Geo Security using GPT Cryptosystem

Eraj Khan[1], Abbas Khalid[2], Arshad Ali[3,*], Muhammad Atif[4], Ahmad Salman Khan[5]
Department of Computer Science & Information Technology, The University Of Lahore, Lahore, 55150, Pakistan[1, 2, 3, 4]
Department of Software Engineering, The University of Lahore, Lahore, 55150, Pakistan[5]

***Abstract*—This paper describes an implementation of location-based encryption using a public key cryptosystem based on the rank error correcting codes. In any code based cryptosystem, public and private keys are in the form of matrices based over the finite field. This work proposes an algorithm for calculating public and private key matrices based on the geographic location of the intended receiver. The main idea is to calculate a location specific parity check matrix and then corresponding public key. Data is encrypted using public key. Some information about the parity check matrix along with other private keys are sent to receiver as cipher-text, encrypted with another instance of the public or GPT cryptosystem using public key of the receiver. The proposed scheme also introduces a method of calculating different parity check matrix for each user.**

Keywords—Location based security; code based cryptosystem; cipher-text; GPT

I. Introduction

Companies all over the world are extending their business models and reaching out to the consumers across the globe. Digital content distribution has overtaken physical format to become the dominant stream for generating revenue. Meanwhile, proliferation of global network interconnections along with ultrahigh density storage devices have made millions of documents online. Although it has made knowledge sharing easy and efficient but this large storage of digital contents has presented unique challenges. Now, the chances of information theft have increased than ever before. Therefore protection of confidentiality, privacy and integrity of information from unauthorised access has become significant challenge for researchers. Encryption provides a way to protect integrity and confidentiality of data which ensures that data is protected from unauthorised access. Traditional cryptographic algorithms provide assurance that only the intended users can access the encrypted data. It is still useful to have an extra layer of security on top of the existing encryption that guarantees that the authorized user can only access the contents at the specific location. It provides information protection against an authorised user who is not at authorised location. If an authorised user tries to decrypt the cipher text at an unauthorised location such as airports, train stations and other public places, the decryption should fail. It can be achieved, by combining decryption key with the location of intended recipient. The idea of combining location of intended recipient with encryption and decryption process was first introduced in [1]. In this paper authors have proposed a geolocking mechanism to be used with traditional cryptographic algorithms. There is a wide range of cryptographic algorithms available which are based on different mathematical problems. Most popular public key cryptosystems are either based on hardness of factorization of large integers (RSA) or on finding discrete logarithms over various groups (ElGamal). Although these algorithms are still considered secure if used with recommended key size and other parameters but after the seminal paper of Peter Shor [2], algorithms based on these problems are known to be broken. In [2], author provided efficient randomized algorithms for solving these problems on hypothetical quantum computer with small probability of errors. Code-based cryptography is a strong candidate for post quantum security algorithms along with hash-based and lattice-based cryptographic algorithms [3]. It is based on that mathematical which can withstand an attack by the adversary equipped with quantum computer [4].

First code-based public key cryptosystem was proposed by Robert McEliece in 1978 [5]. The cryptosystem proposed by McEliece was based on the hardness of decoding a general linear code. In a linear binary code, the problem of finding a codeword is NP-complete. Although it was a very strong algorithm but due to its large and impractical key size which was 219 bits, it didn't gain much of attention. In 1986, Herald Niederreiter [6] proposed another code-based public key cryptosystem. The proposed cryptosystem used the scrambled version of the parity check matrix H as the public key. Due to use of parity check matrix as public key the key size is reduced from 219 to 218. Both of these cryptosystems were based on Hamming metric for calculating code lengths. In 1991, Gabidulin, Paramanov and Tretjakov (GPT) [7] proposed that if rank metric is used instead of Hamming metric, then key size of the code based public key cryptosystem can be reduced further. Based on this idea they proposed another cryptosystem based on rank codes called GPT cryptosystem. Use of rank metric instead of Hamming metric provided two advantages to the GPT cryptosystem. First it has reduced the key size to 214. Secondly as compared to the cryptosystems proposed in [5] and [6] the GPT cryptosystem is much stronger against decoding attacks. As rank codes are well structured and due to this property, over the years several attacks have been launched against GPT cryptosystem. Initially there were series of attacks on the GPT cryptosystem are published in [8-11]. To defend against these attacks several variants of GPT cryptosystems are proposed as well [12-15]. There were some recent attacks on the GPT cryptosystem published in [16-18] but to withstand these attack recently another construction of GPT cryptosystem is proposed by Loidreau P. [19]. Although GPT cryptosystem is continuously under threats over the years. However, it gained so much popularity that it is still considered as a credible post-quantum alternative to traditional cryptography [20]. Various encryption approaches are discussed by research community [21-24].

*Corresponding Author.

This work proposed a technique for implementing geo encryption using a GPT public key cryptosystem based on rank error correcting codes. Variant of GPT cryptosystem proposed in [19] is considered in this work because it withstands all the attacks published against the system so far. As GPT cryptosystem is a code based cryptosystem therefore both public key and private key are in the matrix form. In this paper, a technique for calculating the public key and private key for GPT cryptosystem based on the receiver location is presented.

The rest of the paper is organized as follow: Section II provides related work. It consists of two parts. In first part geo encryption is discussed whereas in second part background information about rank codes is provided. The proposed scheme is described in Section III and results are discussed in Section IV. Finally paper is concluded in Section V.

II. Related Work

Related work section is divided in two sections. First one is about geo encryption and second is about GPT cryptosystem.

A. GEO Encryption

The term geo encryption or location based security refers to the encryption technique that restricts the access to the encrypted data to a specified location at specified time even for a legal user. This restriction can be based on location and time dependent parameters. The main idea is to ensure that data cannot be used other than the authorized location and time. In [1], Logan and Denning proposed a framework for the implementation of geo encryption for digital movie distribution as shown in Fig. 1. They proposed a hybrid approach to implement geo encryption for digital movie distribution which means both public key and private key algorithms are used. The actual data is encrypted using private key encryption algorithm and then the key used for encryption is XORed with a geo lock which is computed using location and decryption time of the intended receiver. This XORed data is then encrypted again using public key encryption algorithm. At the other end, the receiver will first decrypt the encrypted key using private key and then to get the session key the output will XORed with geo lock which is computed using the same function as used at the sender. The session key will be then used to decrypt the data.

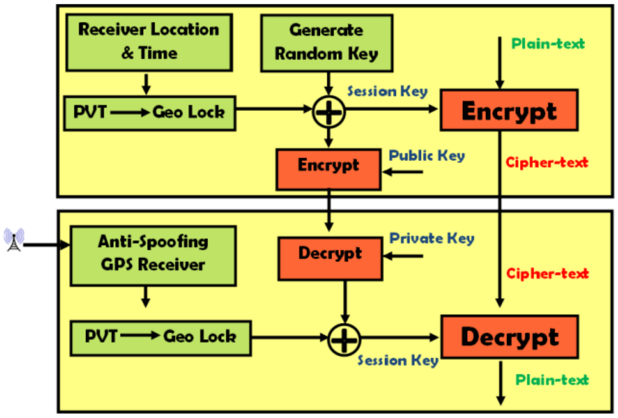


Fig. 1. Geo Codex.

B. Rank Codes

The rank distance codes is first provided in [25]. Let F_q and $F_q{}^N$ represent a finite base field of q elements and an extension field of degree N respectively. If $a = (a_1, a_2, a_3, \ldots, a_n)$ is a vector having coordinates from extension field then the Rank of a is defined as the maximal number of a_i, which are linearly independent over the base field and it can be denoted as $rk(A|F_q)$. The Rank distance between any two vectors a and b is the rank of the difference between a and b i.e. $d(a, b) = rk(a - b|F_q)$. In case of any matrix having all its elements from extension field, its column rank will be all those columns, which are linearly independent over base field. The column rank of any matrix A can be denoted as $rk(A|F_q)$.

In [26], the detailed description about the theory optimal MRD codes is given. The k x n generator matrix G of any MRD code is defined as

$$G = \begin{pmatrix} g_1 & g_2 & g_3 & \cdots & g_n \\ g_1^{[1]} & g_2^{[1]} & g_3^{[1]} & \cdots & g_n^{[1]} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ g_1^{[k-1]} & g_2^{[k-1]} & g_3^{[k-1]} & \cdots & g_n^{[k-1]} \end{pmatrix} \tag{1}$$

where $g_1, g_2, g_3, \ldots, g_n$ are randomly chosen elements from extension field F_q^N. All elements of the first row of generator matrix must be linearly independent over the base field F_q, here $g^{[i]} := g^{q^{i \bmod N}}$ represents the i^{th} Frobenius power of *g*. If *q=2*, then each element of current row is square of the elements present in the same column in previous row. If $m = m_1, m_2, m_3, \ldots, m_k$ is a *k*-dimensional information vector then the corresponding code vector of dimension *n* will be:

$$g(m) = mG_k \tag{2}$$

If $y = g(m) + e$ is a received code-word and if rank of the error vector e is, $rk(e \mid F_q) = s \le t = \left\lfloor \frac{d-1}{2} \right\rfloor$, then the information vector m can be easily gotten back by applying decoding algorithms on y. For decoding any MRD code, another ((n-k) × n) matrix called parity check matrix is need and it is denoted as H. The generator matrix G and parity check matrix H are orthogonal to each other, i.e. $G.H^T = 0$. A parity check matrix can be represented as

$$H = \begin{pmatrix} h_1 & h_2 & h_3 & \cdots & h_n \\ h_1^{[1]} & h_2^{[1]} & h_3^{[1]} & \cdots & h_n^{[1]} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_1^{[d-1]} & h_2^{[d-1]} & h_3^{[d-1]} & \cdots & h_n^{[d-1]} \end{pmatrix} \tag{3}$$

where elements $h_1, h_2, h_3, \dots, h_k$ are from F_q^N and like generator matrix all elements of the first row must be linearly independent over F_q. The notation $h^{[i]} := h^{q^{[i \bmod N]}}$ means the i^{th} Fresenius power of h. The detail of rank codes is described in [21].

C. Description of Stndard GPT Cryptosystem

Several variants of GPT cryptosystem are proposed due to several attacks against the original system. The main structure of almost all the variant remains the same. The difference lies in the construction of various matrices comprising the public key. To be more precise, it's how the elements of various matrices must be chosen so that an attack could be ineffective against the system. For proposed scheme, parameters suggested in [19] are considered which withstands all the known attacks to date. The public key of GPT cryptosystem is given below:

$$G_{pub} = S\, G\, P^{-1} \tag{4}$$

The S is a k × k non-singular, row scrambler matrix over$F_q{}^N$. G is the generator matrix as given in eq (1). The matrix P is an invertible having entries from $F_q{}^N$ as described in [19].

III. GEO Encryption using GPT

The main advantage of using public key cryptosystems over private key cryptosystem is that in former, one does not need to transfer the private keys to the receiver to decrypt the cipher text instead one encrypts the message using public key of the receiver provided through any certificate authority or public directory. The challenge of implementing geo encryption using a public key cryptosystem is to restrict the receiver from decrypting the cipher text without being on the permitted location or at inappropriate time. It means that receiver must verify its location and time to accurately decrypt the cipher text. On the other hand, the receiver is unable to calculate the private keys based on its location and time parameters alone without knowing the structure of the public key. Therefore, to implement the location and time restrictions partial information about the private keys will be sent to receiver and to accurately calculate the private key and to verify its location and allowed decryption time, the receiver has to calculate the rest of the key based on its location and time parameters. Fig. 2 shows the overview of the proposed scheme. In GPT cryptosystem, all keys are in the form of matrices based over F_q^N.

At sender, first a parity check matrix will calculated based on the geographical coordinates of the receiver then a corresponding generator matrix and public key matrix will be calculated. A data will be encrypted using this public key and transmitted over any channel. Some information about the calculation of parity check matrix along with two other matrices which serve as private key will be encrypted together using another instance of GPT cryptosystem using the public key of the intended receiver which means that to decrypt this data we do not need to provide the private key to receiver as it already has it. At the receiver, first of all the encrypted keys will be decrypted to get the two private key matrices and some information about calculating parity check matrix, in parallel to this a key generation which will geographical location parameter to calculate remaining information needed to calculate the parity check matrix. The output of this function along with the information receiver from the sender will be used to calculate the parity check matrix.

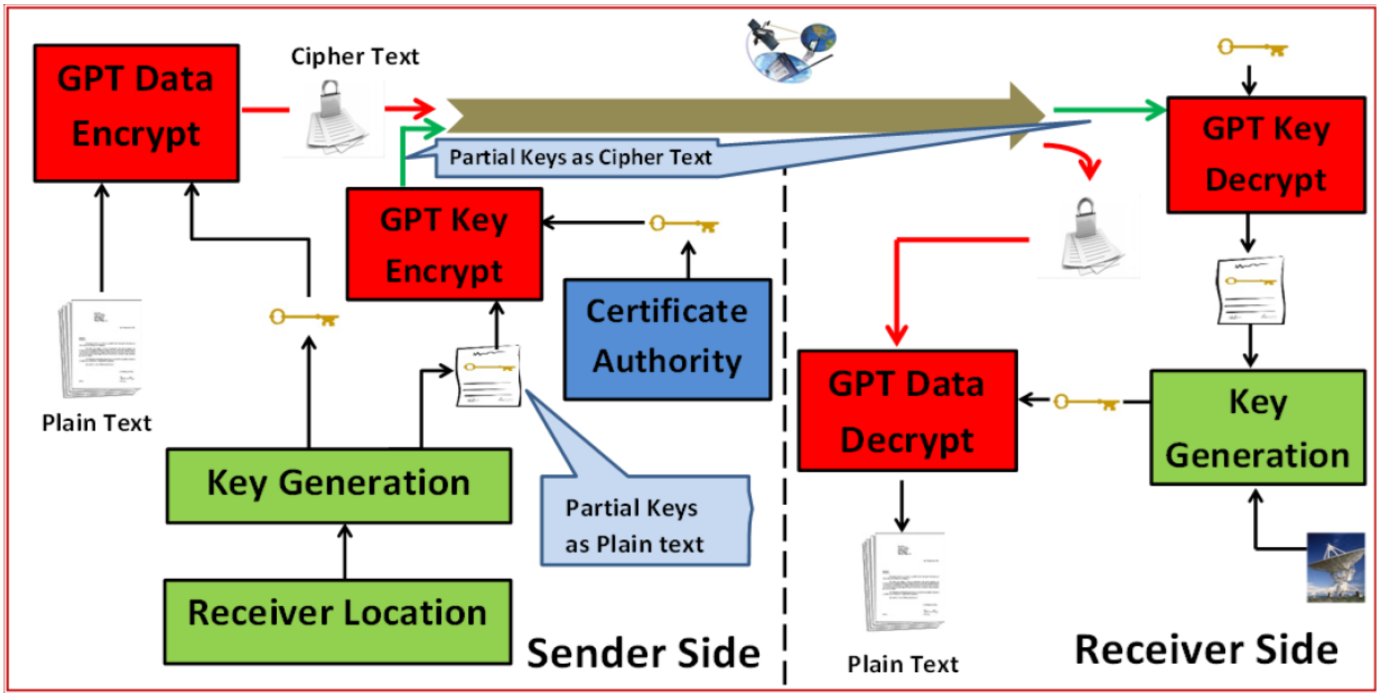


Fig. 2. Geo Encryption with GPT.

Sender side: The key generation algorithm at the sender side is presented in algorithm given below. First of all, the sender will check whether the message is already encrypted. If it is already encrypted, then the sender is not needed to calculate the parity check matrix, generator matrix and public key instead it will calculate an appropriate initial vector and encrypt it along S and P matrices and send it to the receiver along with already encrypted message. So there will be two cases:

Case 1: First time Encryption. The sender will calculate parity check matrix H by calculating an integer constant Φ using the location and time of the week parameters of the intended receiver using a pseudo random permutation. Any pseudo random permutation can be used which could take location and time as inputs and return a big integer as output. It should be noted that the size of the integer constant Φ must of $\leq 2^N - 1$, where N is the degree of the extension field. e.g. if N=8, then the largest value Φ can have is 255.

$$\varphi \leftarrow f(lat, lon, TOW) \tag{5}$$

After calculating Φ, write it in the multiplicative factors of powers of two.

$$\varphi = (x_{n-1}2^{n-1} + x_{n-2}2^{n-2} + \ldots + x_0 2^0) \tag{6}$$

```
Input: Location coordinates and time such as latitude, longitude and time of the week
If Message NOT Encrypted Before Then
        Compute $\varphi \leftarrow f(latitude, logitude, TOW)$
        Write $\varphi$ in the multiplicative factors of powers of 2
        Compute $h_{LV}$ by replacing all the powers of 2 for which coefficient is one
        With corresponding elements from $F_q^N$ and rest with zeros
        Repeat
            Compute $h_{LV}$ by choosing elements from $F_q^N$
            Compute final generating vector $h_f$ for parity check matrix as
                $h_f = h_{LV} \oplus h_{IV}$
        Until All elements of $h_f$ are not linearly independent
            Compute parity check matrix as given in eq 4
            Compute corresponding Generator matrix using matrix reduction algorithm
Else
        If Same location but different receiver
            Compute β by randomly choosing elements from $F_q^N$
            Compute $h_{f\,new} = \beta \times h_f$
            Compute $h_{IV} = h_{f\,new} \oplus h_{LV}$
        Else
            Compute $\phi \leftarrow f(latitude, logitude, TOW)$
            Write $\varphi$ in the multiplicative factors of powers of 2
            Compute $h_{LV}$ by replacing all the powers of 2 for which coefficient is one
            With corresponding elements from $F_q^N$ and rest with zeros
            Compute $h_{IV} = h_f \oplus h_{LV}$
        End if
End if
```

Here x_i are the coefficient having values either 0 or 1 and $i = 0, 1, \ldots, n-1$. In above equation where coefficient is 1, replace powers of 2 with the corresponding elements from the $F_q{}^N$ and insert 0 for rest all. Call it location generating vector h_{LV}. i.e. $h_{LV} = h_{LV_1}\ h_{LV_2}\ h_{LV_2} \ldots h_{LV_n}$

```
Input: $n \times n$ matrix
For $i = 1 \rightarrow n$ then
    If first element of first column $\neq 0$ then
        Do Nothing
    Else
        If left most of the rest of the columns having first element $\neq 0$ & diagonal element $= 0$ then
            Swap the first column with this column
        Else If left most of the rest of the columns having first element $\neq 0$ & diagonal element $\neq 0$
            Swap the first column with this column
        End If
    End If
    If first element of the first column $\neq 0$ then
        Divide the column by its first element
    Else
        Do Nothing
    End If
    Zero the first element of each column except the first column by subtracting an appropriate multiple of the first column
    Rotate rows upwards and column leftwards
End for
```

In next step choose an initial generating vector h_{IV}. There are two reason of choosing this generating vector. First, it will make sure that all the elements of final generating vector which is the first row of parity check matrix are linearly independent. Second, it will completely distort the elements of h_{LV}. To correctly calculate the parity check matrix at receiver, the sender will transmit h_{IV} to the receiver.

Calculate the final generating vector h_f for parity check matrix by taking XOR of h_{IV} and h_{LV}.

$$h_f = h_{IV} \otimes h_{LV} \tag{7}$$

The h_f is the first row of the parity check matrix, rest all rows are frobenious power of each element of the previous row. After calculating parity check matrix H, the sender will calculate a corresponding generator matrix G orthogonal to parity check using matrix reduction algorithm provided below which was Originally proposed in [13].

The message will be encrypted using equation 7 and sent to the receiver using any communication channel. As sender has encrypted the message without giving any prior information about private keys to the receiver, so it will also transmit S^{-1}, P^{-1} and h_{IV} to the receiver in the form of another cipher text encrypted using the public key of the receiver which can be obtained from certificate authority or public directory.

Case 2: Message is already encrypted. If the message is already encrypted then the sender will check whether the intended receiver sharing the same location parameters with the previous receiver for which message was encrypted because in that case the sender will compute new parity check matrix. The parity check matrix of rank codes is quite structured. The generator matrix which is orthogonal to one parity check matrix is also orthogonal to any other parity check matrix which is calculated using the same generating vector multiplied with any randomly chosen element from extension field. It means, if h is the generating vector for a parity check matrix H which is orthogonal to a generator matrix G, then another parity check matrix $\tilde{H}$ which is generated using another generating vector $\tilde{h} = \beta \times h$ is also orthogonal to generator matrix G, where β is a randomly chosen element from extension field. The parity check matrix with $\tilde{h}$ will also be orthogonal to the generator matrix in eq. 3.

$$h_{f_{new}} = \beta \times h_f \tag{8}$$

New h_{IV} will be calculated as

$$h_{IV} = h_{f_{new}} \otimes h_{LV} \tag{9}$$

and sent to the receiver. If the receiver location is different, then the sender will calculate the h_{LV} and will XOR this with the h_f to get h_{IV}.

Receiver side: The algorithm for key generation at receiver is given below. All the steps of key generation algorithm at receiver are similar to steps at the sender except the elements of h_{IV} are not randomly chosen instead the receiver will use the h_{IV} provided by the sender.

```
Input: Location coordinates and time such as latitude, longitude and time of the week
Compute Matrix generating constants
Compute $\varphi \leftarrow f(latitude, logitude, TOW)$
Write $\varphi$ in the multiplicative factors of powers of 2
Compute $h_{LV}$ by replacing all the powers of 2 for which coefficient is one with corresponding elements from $F_q^N$ and rest with zeros
Decrypt the generating vector $h_{IV}$ received from the sender
Compute the final generating vector $h_f$ for parity check matrix as

$h_f = h_{LV} \oplus h_{IV}$
Compute the parity check matrix as given in eq. 4
Compute corresponding Generator matrix using matrix reduction algorithm
```

IV. Analysis and Discussion

In this section, different aspects of the proposed scheme will be analysed and discussed.

A. Security

In the proposed scheme, there are two types of messages which are transmitted from sender to receiver. First is the data itself and second are the private keys to decrypt this data. Both of these messages are encrypted using the GPT cryptosystem first and then transmitted over the channel. Therefore it can be said that the overall security of the proposed scheme is equal to that of security of the cryptosystem itself. Although the keys are transmitted from sender to receiver but these are not enough to decrypt the encrypted data. Only S^{-1}, P^{-1} and h_{IV} are provided. Using the PRP, the receiver has to calculate h_{LV} and combine it with h_{IV} to get parity check matrix. The only way in which an adversary can attack the system is to calculate h_{IV} by correctly guessing the location parameters and decryption time and then using the pseudo random permutation (PRP) to calculate h_{LV}. Therefore it is suggested that the user must use a secure PRP to get the Φ and it must be secret. Even though adversary correctly calculates the h_{LV}, it is not enough because he/she still needs the encrypted S^{-1}, P^{-1} and h_{IV} to decrypt the cipher text. One of the potential attack against any code based cryptosystem is the decoding attack. In decoding attack, an adversary tries to recover the plain text by correcting the errors using a general decoding algorithm without any knowledge of the structure of the code. The aim of the adversary is to try to decode the encoded/encrypted message to the nearest possible codeword. If the adversary is successfully to decode the encrypted message then he/she can recover the original plain text correctly. The general decoding algorithms do not consider the inherent structure of the code. They treat the published code as random. In [10], the authors published two general decoding algorithms to decode an arbitrary linear rank codes. These algorithms can correct errors of rank $t = \left\lfloor \frac{d-1}{2} \right\rfloor$ in $O^{(k+t)^3 t^3 q^{(t-1)(N-t)}}$ and $O^{(Nt)^3 q^{(t-1)(k+1)}}$ operations in F_q. Fig. 3 and Fig. 4 show the operation complexities of these algorithms with respect to key size.

The operation complexity is calculated for three different values of *n* and *k*. It can be seen in both Fig. 3 and Fig. 4 that when n=24 and k=20 the cryptosystem provides good information rate $k/n = 0.833$ but at the same time it is not secure at all and can be easily broken in about 2^{38} operations. For *n=28* and *k=14* the information rate will be $k/n = 0.5$

and the first algorithm requires about 2^{148} operations and second algorithm requires about 2^{113} operations which is quite secure with the current computing power.

B. Key Size and Information Rate

Although algebraic code based cryptosystems are considered as cryptosystems for post quantum computing but they are still not widely accepted for application development due to their huge key size and data expansion. As compared to McEliece [5] and Niederreiter [6], GPT cryptosystem has reduced key size with almost same level of security. Results in Fig. 5 shows the key size versus information rate for different values of *t* where $t \leq \left\lfloor \frac{n-k}{2} \right\rfloor$ is the error correcting capability of the code.

The proposed work randomly chooses h_f to completely mix all elements of generating vector of parity check matrix, so security of proposed system is same as the security of original GPT cryptosystem.

C. Decoding Speed

In [10], two fast decoding algorithms are proposed to decode any rank distance code. First one is Matrix Decoding Algorithm (MDA) and second is Decoding based on Right Euclidean Decoding Algorithm (DREDA). Space limitations discourage from going through each and every step of mentioned algorithms and arithmetic operations required in these steps, instead the arithmetic operations required in these algorithms are summarized in Table I.

Here *t* is error correcting capability of the code and is defined as $t = \left\lfloor \frac{d-1}{2} \right\rfloor$ and *n* is the code length. Table I shows that the number arithmetic operations depend on the size of *t*. Table II and Table III show the exact number of operations required for different values of *t*. Here *n=30* and *k* is changed for different information rates.

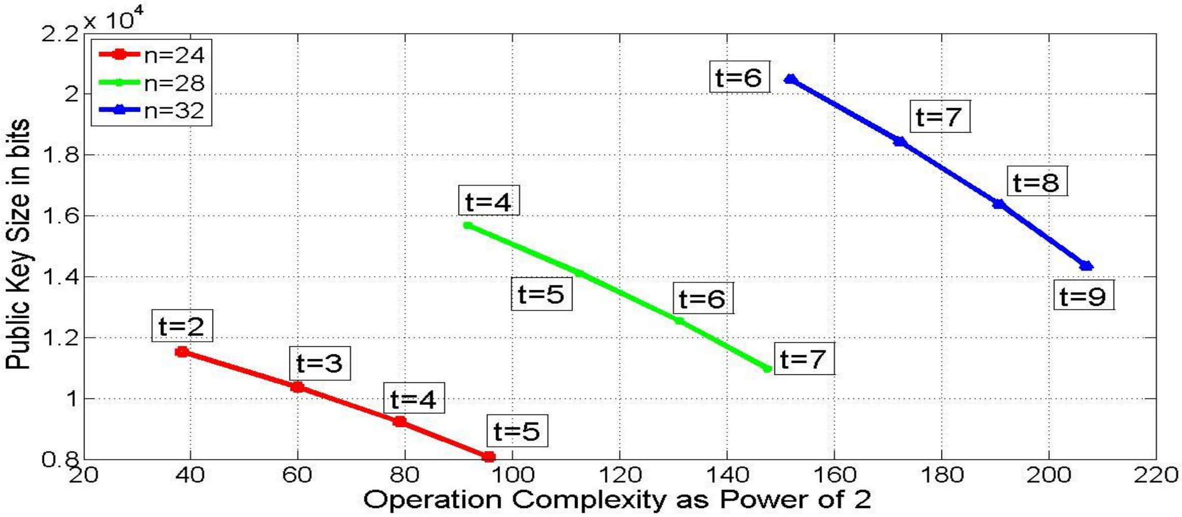


Fig. 3. Key Size vs Complexity of First Decoding Algo.

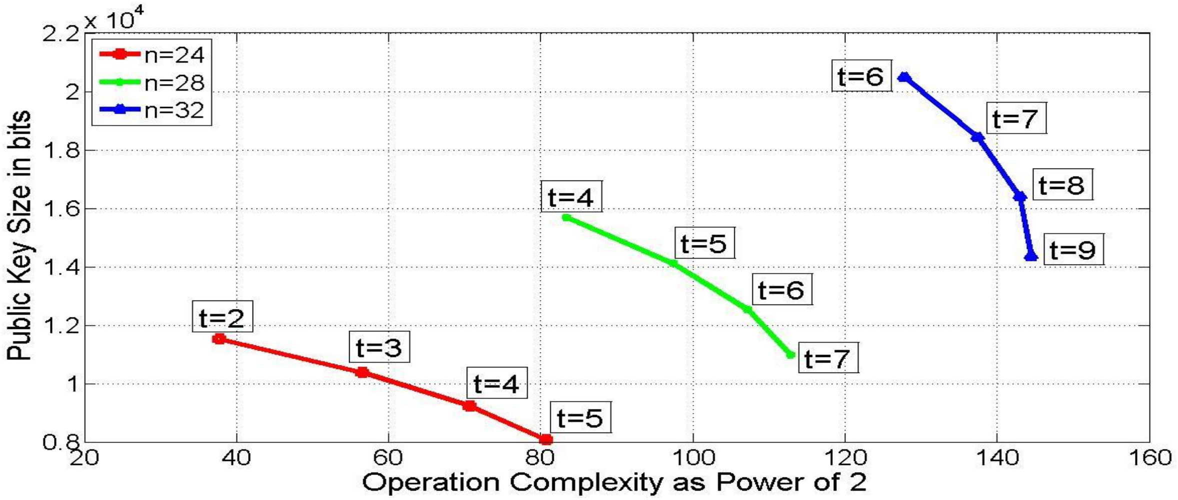


Fig. 4. Key Size vs Complexity of Second Decoding Algo.

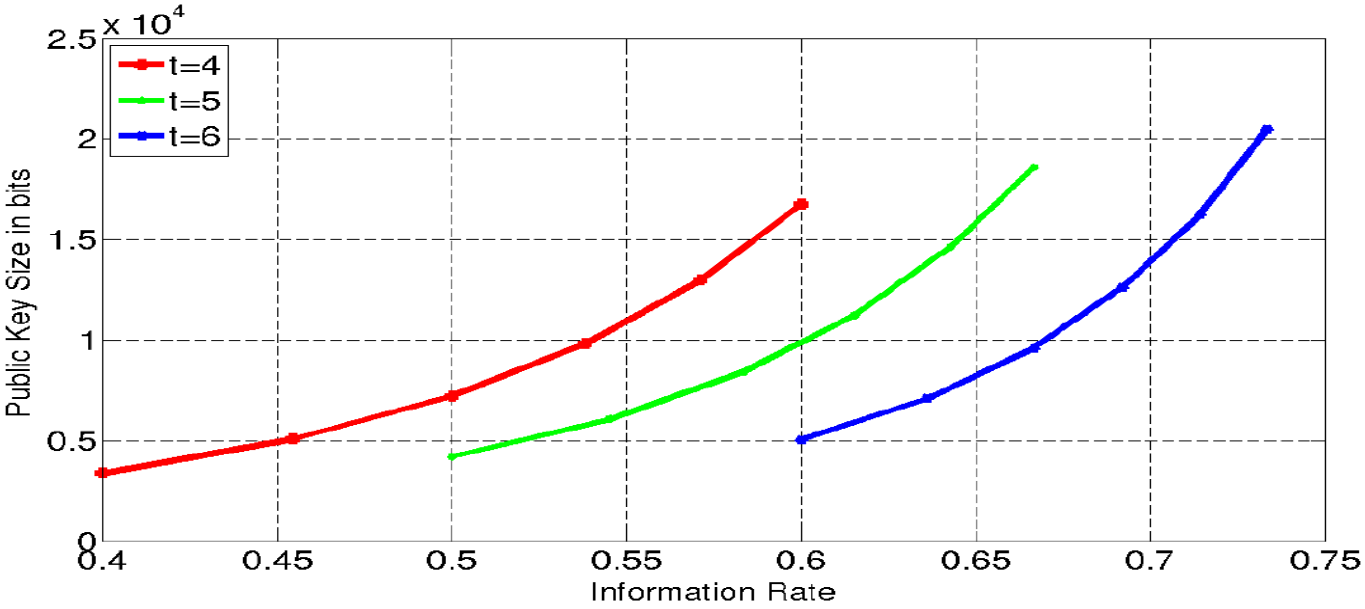


Fig. 5. Key Size vs Information Rate.

TABLE. I. ARITHMETIC OPERATIONS REQUIRED IN FAST DECODING ALGORITHMS

Operations	**MDA**	**DREDA**
Multiplications	$2tn + t!\,(t-1) + 2t(t-1)^2$	$2tn + 2t(2t+1) + t(t-1)^2$
Additions	$t! + 2t^2(t-2) + 3(t-1)n$	$t^2(t + {}^3/_2) + (3t-1)n + {}^t/_2$
Divisions	$2t(t-1)$	$t(t-1) + 2t + 1$
Squares	0	$\frac{t(7t+3)}{2}$
Square Roots	$t(2t+1)$	0

TABLE. II. ARITHMETIC OPERATIONS REQUIRED IN MDA ALGORITHM

K	**T**	**MDA**			
		Mul	Add	Division	SQ
20	4	368	396	24	36
18	5	920	662	40	55
16	6	4236	1484	60	78
14	7	31136	6090	84	105

TABLE. III. ARITHMETIC OPERATIONS REQUIRED IN DREDA ALGORITHM

K	**T**	**DREDA**			
		Mul	Add	Division	SQ
20	4	332	396	21	63
18	5	470	557	31	95
16	6	642	749	43	135
14	7	854	980	57	182

V. CONCLUSION

In this paper, an algorithm for implementing geo encryption using one of the algebraic code based cryptosystem called GPT cryptosystem is proposed. The algorithm proposed a new technique for calculating location based parity check matrix and corresponding public key. Although the key is calculated using geographic location but still it is completely randomized by mixing it with random elements from extension field thus the level of security of the proposed system is equal to that of the underlying GPT public key cryptosystem. This work introduced an idea of encrypting with one public key and decrypted with multiple different private keys but calculating different parity check matrix for each user.

REFERENCES

[1] Scott, L., Denning, D.E., “A Location Based Encryption Technique and Some of Its Applications”, Proceedings of the 2003 National Technical Meeting of The Institute of Navigation, Anaheim, CA, pp. 734-740, January 2003. http://faculty.nps.edu/dedennin/publications/locationbasedencryption-ion2003.pdf.

[2] Peter W. Shor, “Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer”, SIAM Journal on Computing, vol. 26, issue. 5, pp. 1484-1509, October 1997.http://dx.doi.org/10.1137/S0097539795293172.

[3] Nicolas Sendrier, Jean-Pierre Tillich, "Code-based Cryptography: New Security solutions against a quantum adversary", ERCIM News, ERCIM, 2016, Special Theme Cybersecurity (106), July 2016 https://hal.archives-ouvertes.fr/hal-01410068/file/codebased-final.pdf.

[4] Nicolas Sendrier,"Code-Based Cryptography: State of the Art and Perspectives", IEEE Security & Privacy, vol.15, issue 4, pp.44-50, 2017 https://doi.org/10.1109/MSP.2017.3151345.

[5] McEliece, R. J., "A Public Key Cryptosystem Based on Algebraic Coding Theory", JPL DSN Progress Rep. 42-44. https://tmo.jpl.nasa.gov/progress_report2/42-44/44N.PDF.

[6] Niederreiter, H., "Knapsack-Type Cryptosystem and Algebraic Coding Theory", Probl. Control and Inform. Theory, vol. 15, pp. 19-34, 1986.

[7] Gabidulin, E. M., Paramonov, A.V., Tretjakov, O.V., "Ideals over a non-commutative ring and their application in cryptology", advances in

Cryptology, Proc. EUROCRYPT' 91, LNCS 547, D. W. Davies, Ed. Springer-Verlag, 1991, pp. 482-489. https://doi.org/10.1007/3-540-46416-6_41.

[8] Gibson J. K., “Severely denting the Gabidulin version of the McEliece public key Cryptosystem”, Designs Codes and Cryptography, 6(1), 1995, pp.37-45. https://doi.org/10.1007/BF01390769.

[9] Gibson J. K., “The security of the Gabidulin public-key cryptosystem”, U. M. Maurer (Ed.), Advances in Cryptology --EUROCRYPT'96, LNCS vol 1070, Springer, Berlin, 1996, pp. 212-223. https://doi.org/10.1007/3-540-68339-9_19.

[10] Ourivski A. V., Johansson T., “New Technique for Decoding Codes in the Rank Metric and Its Cryptography Applications”, Problems of Information Transmission, 38(3), 2002, pp. 237-246. https://doi.org/ 10.1023/A:1020369320078.

[11] Overbeck, R., “Structural attacks for Public Key Cryptosystem Based on Gabidulin codes”, Journal of Cryptology, 21(2), 2008, pp.280-301. https://doi.org/10.1007/s00145-007-9003-9.

[12] Gabidulin E. M., "Attacks and counter-attacks on the GPT public key cryptosystem", Designs Codes and Cryptography, Springer Netherlands, (48) 2, August 2008, pp.171-177. https://doi.org/10.1007/s10623-007-9160-8.

[13] Gabidulin, E.M., Rashwan, H., Honary,B., "On Improving Security of GPT Cryptosystems", Int. Symposium on Information Theory, pp.1110-1114. 2009 https://doi.org/10.1109/ISIT.2009.5206029.

[14] Rashwan, H., Gabidulin, E. M. and Honary, B. (2011), “Security of the GPT cryptosystem and its applications to cryptography”. Security Comm. Networks, 4: 937–946. 2011 http://dx.doi.org/10.1002/sec.228.

[15] Khan, E., Gabidulin, E.M., Honary, B., Ahmed H., " Modified Niederreiter Type of GPT Cryptosystem Based on Reducible Rank Codes", Designs Codes and Cryptography, Springer, vol(70) 1, pp. 231-239, 2014 https://doi.org/10.1007/s10623-012-9757-4.

[16] Anna-Lena Horlemann-Trautmann, Kyle Marshall and Joachim Rosenthal, "Extension of Overbeck's attack for Gabidulin based cryptosystems", Designs Codes and Cryptography, 2017. https://doi.org/10.1007/s10623-017-0343-7.

[17] Ayoub Otmani, Herve Tale Kalachi, Selestin NDJEYA, "Improved Cryptanalysis of rank metric schemes based on Gabidulin Codes", Designs, Codes and Cryptography, 2017. https://doi.org/10.1007 /s10623-017-0434-5.

[18] Philippe Gaborit, Ayoub Otmani, Herve Tale Kalachi, "Polynomial-time Key recovery attack on the Faure-Loidreau scheme based on Gabidulin Codes", Designs, Codes and Cryptography, 2017. https://doi.org/10. 1007/s10623-017-0402-0.

[19] Pierre Loidreau, "A new rank metric codes based encryption scheme", Post-Quantum Cryptography : 8th International Workshop, PQCrypto 2017, Utrecht, The Netherlands, June 26-28, 2017, Proceedings, pp. 3-17, 2017. https://doi.org/10.1007/978-3-319-59879-6_1.

[20] Philippe Gaborit, Oliver Ruatta, Julien Schrek, Jean-Pierre Tillich, "Rank based cryptography: a credible post-quantum alternative to classical cryptography", NIST workshop on cybersecurrity in a Post-Quantum World 2015. https://csrc.nist.gov/csrc/media/events/workshop-on-cybersecurity-in-a-post-quantum world/documents/papers/session1-gaborit-paper.pdf.

[21] Singh KJ and Gagneja K. Overview of securing multimedia content using efficient encryption methods and modes. International Journal of Advanced and Applied Sciences, 2017; 4(10): 84-96.

[22] Arboleda. ER, Fenomeno CE and Jimenez JZ. KED-AES algorithm: combined key encryption decryption and advance encryption standard algorithm. Int. J. of Adv. in Appl. Sci. 2018; 8(1): 44-53.

[23] Nagavalli S, Ramachandran G. A Secure Data Transmission Scheme using Asymmetric Semi-Homomorphic Encryption Scheme. Int. J. of Adv. in Appl. Sci. 2018;7(4): 369-376.

[24] Pushpa K, Lakshmi L, Sabitha Ch.,Dhana B and Sreeja S.Top-K search scheme on encrypted data in cloud. Int. J. of Adv. in Appl. Sci. 2019;9(1): 67-69.

[25] E. M. Gabidulin, "The theory of codes with maximum rank distance", Problems Inform. Transmission 21 (1), pp. 1-12, 1985. https://www. researchgate.net/publication/235008632_Theory_of_codes_with_maxim um_rank_distance_translation.

[26] E.M. Gabidulin,: ``Public-Key Cryptosystems Based on Linear Codes over Large Alphabets''; Efficiency and Weakness, in:Codes and Ciphers, Editor: P.G. Farrell, pp. 17--32, Essex: Formara Limited, 1995. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.4876&rep =rep1&type=pdf.

Localization of Mobile Submerged Sensors using Lambert-W Function and Cayley-Menger Determinant

Anirban Paul[1], Miad Islam[2], Md. Ferdousur Rahman[3], Anisur Rahman[4]
Department of Computer Science and Engineering
East West University, Dhaka
Bangladesh

***Abstract*—This paper demonstrates a new mechanism to localize mobile submerged sensors using only a single beacon node. In range-based localization, fast and accurate distance measurement is vital in underwater wireless sensor networks (UWSN). The knowledge of exact coordinates of the sensors is as important as the actuated data in underwater wireless sensor networks. Mostly bouncing technique is used to determine the distance between the beacon and the sensors. Moreover, to determine the coordinates, trilateration and multilateration technique is used; where using multiple beacons (usually three or more) is the most common approach. Nevertheless, because of many factors, this method gives less accurate results in distance measurements, which finally leads to determine erroneous coordinates. As TDOA is very ponderous to achieve in underwater environment because of time synchronization; again, using AOA is extremely difficult and challenging; TOA is the most common approach and is widely employed. However, it still needs precise synchronization. So, to determine the distances between beacon and sensor nodes, we have used a method based on Lambert-W function in this study, which is an approach based on RSS, and it avoids any synchronization. Besides, coordinates of the mobile sensors are calculated using Cayley-Menger determinant. In this paper, the method is derived and the accuracy is verified by simulation results.**

***Keywords*—Lambert-W function; Cayley-Menger determinant; submerged mobile sensor; single beacon; localization**

I. Introduction

The underwater wireless communication is still a very challenging term in wireless communications. As radio signals cannot propagate much underwater, acoustic signals are widely used as a substitute. The need of knowing and observing the marine life is increasing rapidly. The data lays underwater could be of great use with precise Information about the location. Moreover, collecting those data is equally important for underwater surveillance, deep-sea exploration etc. Therefore, it is very important to collect those data using submerged sensors. In addition, according to [1], the accurate localization of sensors is vital for proper interpretation of the actuated data. In terrestrial condition, localization of wireless sensors has developed greatly and many mechanisms have been proposed. There are two categories: range-free and range-based schemes. Range-based scheme can give more accurate result than range-free scheme and most of the sensors nowadays have those characteristics. In this paper, we have studied on accuracy of submerged moving sensors coordinate, which has a wide range of application in practical, like pollutant tracking and estuary monitoring [2]. Moreover, as seen in [2, 3], in underwater, acoustic signals are used for range measurements because radio signals cannot propagate much under water.

In many studies regarding UWSN, the main puzzle of computing RSS has been resolved circuitously. According to Patwari [16], most of the studies have presumed that the RSS value can be converted to the distance but the complication of conversion has been ignored. In [6], the authors have proposed two methods for determining the distance of sensors in underwater using the transmission loss (TL), which can be acquired from the RSS. They proved that, the method using Lambert-W function gives significantly better result than the Newton-Raphson method considering the possible environmental constraints. In addition, the simulations result strongly back their claim. The resultant value is also notably close to the actual value. The authors of [5] proposed a method for localization based on sensors anchored to the seabed and the mobile sensors try to communicate directly with these anchored nodes to determine their position. This scheme cannot be applied to dynamic environment.

Rahman [4] has introduced a method to localize underwater sensors using Cayley-Menger determinant. They have used bouncing technique to calculate the necessary distances between the beacon and the sensors. Moreover, they used only a single beacon to localize the sensors. The sensors are considered static and the beacon takes measurements from at least six randomly different positions. However, their proposed model gives significantly accurate results.

The authors of [14] proposed a method to calculate the coordinates of submerged static sensors using a single beacon. They used trilateration to solve the problem and they dealt with multipath fading during distance measurement. In [7], authors solved the equations of multilateral operation. They tried to determine the unknown position using nonlinear square optimization. However, as per [9], in a nonlinear equation system, it does not give surety of a unique solution. For example, in trilateration method, distance is the only data to measure the distance between the nodes.

After analyzing the studies discussed above, we propose a new mechanism to find out the coordinates of mobile submerged sensors, using a single beacon at the water surface. In addition, as in [8], to obtain primary subsets of nodes the precise conditions were vindicated using rigidity principle.

This paper is arranged as follows: solvable configuration and problem domain are described in Section II. Section III explains the technique for distance calculations. In Section IV, the theoretical method to determine the static sensors coordinates explained. In Section V, mechanisms for determining the mobile sensors coordination is explained. In Section VI, analysis part is explained. Section VII discusses simulation results and at last, conclusions and future possible works are explained in Section VIII.

II. Proposed Configuration

A. *Problem Domain*

In the proposed method at least 3sensor node and one beacon node is necessary to determine the coordinates of the mobile submerged sensors. The beacon is floating at the water surface. The distance between the sensors and the beacon are measured using a method based on Lambert function, as described in Section III. Usually a buoy or boat is used as a beacon and the sensors are deployed underwater in aquatic environment such as ocean or river. All the sensors are supposed to be in the same plane in underwater, which is parallel to the plane of the water surface where the beacon is, shown in Fig. 1 and 2.

We assume for simplicity, the sensors are Autonomous Underwater Vehicle (AUV), having static speed; and all sensors are moving in same direction. For six different positions of the sensors, same numbers of random different positions of the beacon are needed to take the measurements of the distance in between the sensors and the beacons. In the proposed model, the sensors generate acoustic signals in a pre-defined frequency. Then the beacon calculates transmission loss from RSS and calculates the distance to sensors. A solvable configuration of three sensors with the beacon is shown in Fig. 2. As stated by [11], the proposed model works in underwater within 1.8-323m depth. Moreover, as specified in [13], for acoustic signals, the method works for a frequency range below 50 kHz.

Our proposed model has a wide range of practical applications as most of the research and explorations of ocean take place in shallow water.

B. *Environmental Constraints*

The environment of underwater is more hostile than terrestrial environment. There are many environmental variables such as corrosion by salt water, the node's movements by the ocean current, attenuation distortions, issues of multi-path and difficulty of sensor nodes' deployment. In [13] we see that, it is quite complex and difficult to process and gather the information of the environment through ocean data communication due to the constraints of underwater environment unlike the terrestrial environment.

Acoustic signal is slower but propagates much further comparing to the radio signal. Again, the transmission loss is affected by temperature, depth, salinity, scattering, diffraction etc. As in [15], how these previously mentioned factors affect the transmission loss is not considered in this study and transmission loss is taken as a variable TL.

III. Distance Determination for Cayley Menger using Lambert-W Function

Assumptions:

- The sensors can generate acoustic signals with a pre-defined frequency.
- While measuring distances, the factors that affect transmission loss is considered.
- Base for all the sensors is same and the base is of tetrahedron shape.
- All sensor nodes will have a fixed ID.

A. *Underwater Acoustic Transmission Loss Calculation*

There are two types of acoustic sound loss in underwater. These are classified as attenuation loss and spreading loss. Spreading loss includes spherical and cylindrical loss. In addition, attenuation loss includes absorption, leakage from ducts, scattering and diffraction. For simplicity, we only consider the transmission medium losses. For a distance D,

$$TL_{sph} = 20\log(D), \text{ Spherical} \tag{1}$$

$$TL_{cyl} = 10\log(D), \text{ Cylindrical} \tag{2}$$

So, total transmission loss we get from (1) and (2) is,

$$TL_{total} = TL_{sph} + TL_{cyl} + 10^{-3}\alpha D \tag{3}$$

Here, α is the absorption co-efficient, as per the Thorp absorption coefficient model.

$$\alpha = 1.0936\,[\frac{0.1f2}{1+f2} + \frac{40f2}{4100+f2}] \tag{4}$$

Here, 1.0936 is multiplied to change the unit it to dBkm^{-1}. As stated by [11], under a wide variety of condition, spherical data fits the measured data. So, by reducing (3) and (1),

$$TL = \frac{20\,ln(D)}{ln\,(10)} + \frac{\alpha D}{1000} \tag{5}$$

We will need to convert (5) into Lambert function to find a solution for the distance *D.*

Here, the Lambert-W function is

$$Y = AXe^{AX} = W(X) \tag{6}$$

We need to find Lambert function *X=W(Y*). Now, considering *X* = *D* from (6), we will have,

$$Y = AXe^{AX}$$

$$\frac{Y}{A} = D.\ e^{A.D} \tag{7}$$

$$ln\left(\frac{Y}{A}\right) = ln\ (D) + A.D$$

Let's consider, $\gamma = ln\ (10)/20$, then,

$$\frac{ln(Y/A)}{\gamma} = \frac{ln(D) + A.D}{\gamma} \qquad (8)$$

To derive (5), we must have these two conditions,

$\left(\frac{A}{\gamma}\right) = \left(\frac{\alpha}{1000}\right)$ and

$$\frac{ln(Y/A)}{\gamma} = \mathrm{TL} \qquad (9)$$

By solving them we get,

$A = (\gamma\alpha/1000)$,

$$Y = Ae^{\gamma TL} \qquad (10)$$

B. Distance Measurements using Lambert-W Function

The Lambert-W function, is the multi valued inverse of $\omega \rightarrow \omega e^{\omega}$ defined by,

$$z = (z)^{W(z)} \qquad (11)$$

Where, z and $W(z)$ can be complex. The sub-domain of both real and positive is used.

Here, z is the transmission loss (TL). There is exactly one $\omega \geq 0$ for each $z \geq 0$, so W returns a single value as distance.

Now,

$$\omega_1 = p - 1, where\ p = \sqrt{2(eY + 1)} \qquad (12)$$

Using Halley Method, iterating toward $W(Y)$ from (12),

$$\omega_j + 1 = \omega_j - \frac{\omega_j e^{\omega_j} - z}{e\omega_j(\omega_j + 1) - ((\omega_j + 2)(\omega_j e^{\omega_j} - z)/(2\omega_j + 2))} \qquad (13)$$

This solves (11) for ω where $z > 0$. Accordingly,

$$Y = AXe^{AX} \qquad \therefore X = \frac{W(Y)}{A} \qquad (14)$$

From (14) and (10), we can write the final equation of Distance (D) via Lambert function,

$$D = \frac{20000 \times W((ln(10)/20000)\ \alpha e^{TL})}{\alpha\ ln(10)} \qquad (15)$$

IV. Coordinates Determination of Static Sensors using Cayley Menger

A. Determining Coordinates of the Sensor Nodes

The goal of localization of the sensor nodes is to determine the precise position of the sensors. The only measurement here is to measure the distance. However, in nonlinear system, the degree of freedom analysis does not guarantee a singular solution. Multilateration or trilateration techniques are some nonlinear system, which are used to localize the sensors in some or full. According to Guevara [10], the convergence of Bayesian methods and optimization algorithms heavily depends on primary conditions used. They linearize the trilateration equations to overcome convergence problem.

In Fig. 1, the initial position of the beacon and the sensors are shown. The position of the beacon is S_j, (j = 4, 5… 9) and three sensor nodes are S_i, (i = 1, 2, 3). Without affecting generality, a coordinate system can be defined with respect to one of the sensor S_i, (i = 1, 2, 3) as the origin (0, 0, 0) of the system. Now the trilateration equation can be formed. The distance between beacon and the sensors are weighed data. Again, inter node distances d_{12}, d_{13}, d_{23} and volume of the tetrahedron V_t, are unknown. We write the equations based on the local positioning system configuration of Fig. 1. For that using Cayley-Menger determinant, the volume of tetrahedron V_t is expressed as followings:

$$288V_t^2 = \begin{vmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & d^2{}_{12} & d^2{}_{13} & d^2{}_{14} \\ 1 & d^2{}_{12} & 0 & d^2{}_{23} & d^2{}_{24} \\ 1 & d^2{}_{13} & d^2{}_{23} & 0 & d^2{}_{34} \\ 1 & d^2{}_{14} & d^2{}_{24} & d^2{}_{34} & 0 \end{vmatrix} \qquad (16)$$

$$d^2{}_{34}(d^2{}_{12} - d^2{}_{23} - d^2{}_{13}) + d^2{}_{14}\left(\frac{d^4{}_{23}}{d^2{}_{12}} - d^2{}_{23} - \frac{d^2{}_{13}d^2{}_{23}}{d^2{}_{12}}\right) +$$
$$d^2{}_{24}\left(\frac{d^4{}_{13}}{d^2{}_{12}} - \frac{d^2{}_{13}d^2{}_{23}}{d^2{}_{12}} - d^2{}_{13}\right) - (d^2{}_{14}d^2{}_{24} + d^2{}_{14}d^2{}_{34} - d^2{}_{24}d^2{}_{34} -$$
$$d^4{}_{14})\frac{d^2{}_{23}}{d^2{}_{12}} - (d^2{}_{34}d^2{}_{24} - d^2{}_{14}d^2{}_{34} + d^2{}_{14}d^2{}_{24} - d^4{}_{24})\frac{d^2{}_{13}}{d^2{}_{12}} +$$
$$\left(\frac{144v_t{}^2}{d^2{}_{12}} + d_{13}{}^2 d_{23}{}^2\right) = (d^2{}_{24}d^2{}_{34} - d^4{}_{34} + d^2{}_{14}d^2{}_{34} - d^2{}_{14}d^2{}_{24})$$

Here the unknown terms are,

$(d^2{}_{12} - d^2{}_{23} - d^2{}_{13})$, $\left(\frac{d^4{}_{13}}{d^2{}_{12}} - \frac{d^2{}_{13}d^2{}_{23}}{d^2{}_{12}} - d^2{}_{13}\right)$, $\frac{d^2{}_{23}}{d^2{}_{12}}$, $\frac{d^2{}_{13}}{d^2{}_{12}}$,

$\left(\frac{144v_t{}^2}{d^2{}_{12}} + d_{13}{}^2 d_{23}{}^2\right)$ and $\left(\frac{d^4{}_{23}}{d^4{}_{12}} - d^2{}_{23} - \frac{d^2{}_{13}d^2{}_{23}}{d^2{}_{12}}\right)$

By grouping and expanding known–unknown variables, we get,

$$d^2{}_{14}X_1 + d^2{}_{24}X_2 + d^2{}_{34}X_3 - (d^2{}_{14} - d^2{}_{34})(d^2{}_{24} - d^2{}_{14})X_4 - (d^2{}_{24} - d^2{}_{14})(d^2{}_{34} - d^2{}_{24})X_5 + X_6 = (d^2{}_{24} - d^2{}_{34})(d^2{}_{34} - d^2{}_{14}) \qquad (17)$$

Equation (17) becomes as the linear shape of $a_1x_1 + a_2x_2 + a_3x_3 + \ldots\ldots + a_nx_n = b_1$. We need at least six measurements as we have six unknowns in (17). And this can be performed by following the same approach described in section earlier, moving the beacon S_j, (j = 4, 5... 9) to six different positions and measuring the distances in the vicinity of P_4. Finally, we get m number of linear equations of the form,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots\ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots\ldots + a_{2n}x_n &= b_2 \qquad (18) \\ a_{m1}x_1 + a_{m2}x_2 + \ldots\ldots + a_{mn}x_n &= b_m \end{aligned}$$

By omitting references to the variables, the system of (18) can be represented by the augmented matrix of the system. Here, the first linear equation is represented by the first row of the array and so on. We can express it in a linear form, which is AX = b. Then the equations can be written as:

$$A = \begin{vmatrix} d^2{}_{14} & d^2{}_{24} & d^2{}_{34} & -(d^2{}_{14} - d^2{}_{34})(d^2{}_{24} - d^2{}_{14}) & -(d^2{}_{24} - d^2{}_{14})(d^2{}_{34} - d^2{}_{24}) & 1 \\ d^2{}_{15} & d^2{}_{25} & d^2{}_{35} & -(d^2{}_{15} - d^2{}_{35})(d^2{}_{25} - d^2{}_{15}) & -(d^2{}_{25} - d^2{}_{15})(d^2{}_{35} - d^2{}_{25}) & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d^2{}_{19} & d^2{}_{29} & d^2{}_{39} & -(d^2{}_{19} - d^2{}_{39})(d^2{}_{29} - d^2{}_{19}) & -(d^2{}_{29} - d^2{}_{19})(d^2{}_{39} - d^2{}_{29}) & 1 \end{vmatrix}$$

$$X = \begin{bmatrix} \frac{d^4{}_{23}}{d^2{}_{12}} - d^2{}_{23} - \frac{d^2{}_{13}d^2{}_{23}}{d^2{}_{12}} \\ \frac{d^4{}_{13}}{d^2{}_{12}} - \frac{d^2{}_{13}d^2{}_{23}}{d^2{}_{12}} - d^2{}_{13} \\ d^2{}_{12} - d^2{}_{23} - d^2{}_{13} \\ \frac{d^2{}_{23}}{d^2{}_{12}} \\ \frac{d^2{}_{13}}{d^2{}_{12}} \\ 144\frac{v_t{}^2}{d_{12}{}^2} + d^2{}_{13}d^2{}_{23} \end{bmatrix} \quad b = \begin{bmatrix} (d^2{}_{24} - d^2{}_{34})(d^2{}_{34} - d^2{}_{14}) \\ (d^2{}_{25} - d^2{}_{35})(d^2{}_{35} - d^2{}_{15}) \\ \vdots \\ (d^2{}_{29} - d^2{}_{39})(d^2{}_{39} - d^2{}_{19}) \end{bmatrix}$$

After finding the values of X (X_1, X_2, X_3, X_4, X_5, X_6) we calculate d_{12}, d_{23}, d_{13} as follows:

$$d^2{}_{12} = \frac{X4}{1 - X4 - X5}, \qquad d^2{}_{13} = \frac{X3X5}{1 - X4 - X5}, \qquad d^2{}_{23} = \frac{X3X4}{1 - X4 - X5}$$

As we assume that the submerged sensors coordinate are $S_1 = (0, 0, 0)$, $S_2 = (0, y_2, 0)$ and $S_3 = (x_3, y_3, 0)$ then with respect to coordinates of the sensors the inter sensor distances could be written as follows:

$$d^2{}_{12} = y_2{}^2, \qquad d^2{}_{13} = x_3{}^2 + y_3{}^2, \qquad d^2{}_{23} = x_3{}^2 + (y_3{}^2 - y_2{}^2)$$

After finding the values above, we can calculate the unknown values as follows [4]:

$y_2 = d_{12}$,

$$y_3 = \frac{d^2{}_{12} + d^2{}_{13} - d^2{}_{23}}{2d_{12}},$$

$$x_3 = \sqrt{\left(d^2{}_{13} - \left(\frac{d^2{}_{12} + d^2{}_{13} - d^2{}_{23}}{2d_{12}}\right)^2\right)}$$

Here, d_{12}, d_{13}, d_{23} are computed distance. The sensors coordinate with respect to S_1 are given in Table I.

B. Determining the Coordinates of the Sensor Nodes

Now, the position of the beacon has to be in origin (0, 0, 0) to determine the sensors coordinates. As we can calculate other sensors coordinates with respect to S_1, we only need to find the coordinate of sensor S_1 with respect to the beacon. Now the coordinates of sensor S_1 with respect to the beacon node can be determined by following these steps.

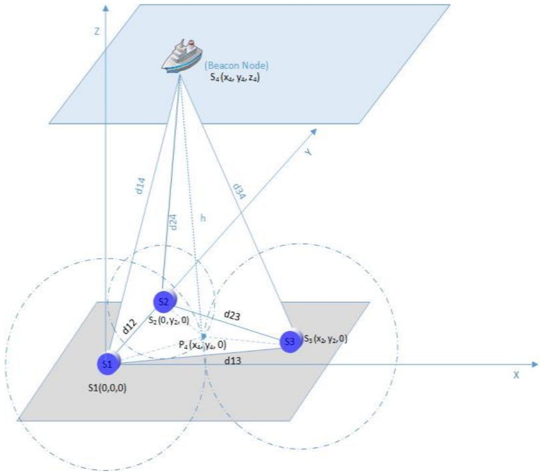


Fig. 1. Coordinate Determinations with Single Beacon.

TABLE. I. COORDINATES OF THE SENSORS WITH RESPECT TO S_1

Node	Coordinates
S_1	(0, 0, 0)
S_2	(0, d_{12}, 0)
S_3	$\left(\sqrt{\left(d^2{}_{13} - \left(\frac{d^2{}_{12} + d^2{}_{13} - d^2{}_{23}}{2d_{12}}\right)^2\right)}, \left(\frac{d^2{}_{12} + d^2{}_{13} - d^2{}_{23}}{2d_{12}}\right), 0\right)$

According to [12], the vertical distance *h* can be measured. After measuring *h*, we assume the projected coordinate of S_4 is P_4 (x_4, y_4, 0) on the plane XY. To find x_4 and y_4, trilateration technique is applied, assuming the distances between sensors S_1, S_2, S_3 and projected point P_4 are D_{14}, D_{24} and D_{34}.

$$D^2{}_{14} = x_4{}^2 + y_4{}^2 \tag{19}$$

$$D^2{}_{24} = x_4{}^2 + (y_4 - y_2)^2 \tag{20}$$

$$D^2{}_{34} = (x_4 - x_3)^2 + (y_4 - y_3)^2 \tag{21}$$

From (19), (20) and (21) we get the coordinates of projected beacon as follows P_4 (x_4, y_4, z_4).

$$X_4 = \sqrt{\frac{1}{2D}(2d_{12}D^2{}_{14} - D^2{}_{14} + D^2{}_{24} + d^2{}_{12})},$$

$$Y_4 = \frac{1}{2d_{12}}(D^2{}_{14} - D^2{}_{24} + d^2{}_{12})$$

As the hypotenuse of $\triangle S_1P_4S_4$, $\triangle S_2P_4S_4$ and $\triangle S_3P_4S_4$ are d_{14}, d_{24} and d_{34} respectively, so the distance D_{14}, D_{24} and D_{34} is possible to obtain by implementing Pythagorean Theorem. Now, the coordinate of the beacon S_4 (x_4, y_4, z_4) will transform as (x_4, y_4, h) where all elements are known.

$$S_4(x_4, y_4, 0) = S_4\left(\begin{array}{c}\sqrt{\frac{1}{2D}(2d_{12}D^2{}_{14} - D^2{}_{14} + D^2{}_{24} + d^2{}_{12})}, \\ \frac{1}{2d_{12}}(D^2{}_{14} - D^2{}_{24} + d^2{}_{12}),\ h\end{array}\right)$$

Applying linear transformation, the coordinate of the beacon node is replaced by the origin of the Cartesian system. The linear transformation would give the coordinates of other sensor nodes as in Table II.

TABLE. II. COORDINATES OF THE SENSORS WITH RESPECT TO S_4

Sensors	Coordinates	Sensors	Coordinates
S_4	(0, 0, 0)	S_2	($-x_2$, y_2-y_4, $-z_4$)
S_1	($-x_4$, $-y_4$, $-z_4$)	S_3	($-x_4$, y_2-y_4, $-z_4$)

V. COORDINATE DETERMINATION OF MOBILE SENSORS

Initially, the distance between the sensors and the beacon are to be calculated with the help of Lambert-W function. Here, d11, d21 & d31 are the distance between sensor S_i's (i = 1, 2, 3) initial position to beacons initial (B_k = 1) position, respectively. As both the beacon and sensors are mobile, the distance between beacons new position to sensors new position is to be calculated using the Lambert function, as mentioned in Section III. For S_1, it is $d1_22$, as shown in Fig. 2. Concurrently, $d2_22$ and $d3_22$ is calculated for S_2 and S_3 following the same process. Here, in dSi_jB_k, S_i (i = 1, 2, 3) is the sensor number, j (j = 1, 2, 3... 6) is the sensors position and B_k (k = 1, 2, 3… 6) is the beacon's position.

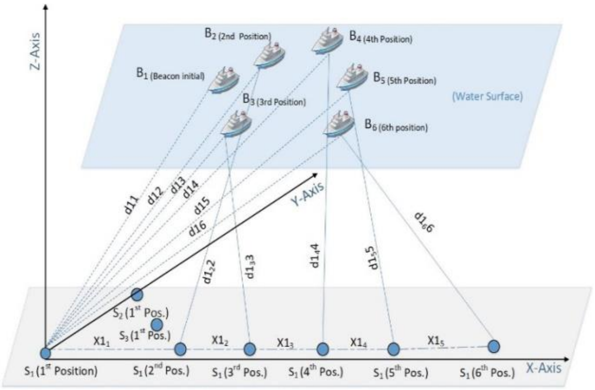


Fig. 2. Coordinate Determination of Mobile Sensors.

Now the distance from sensor's initial position to second position is calculated with the help of the sensors speed and the time beacon took to travel to its new position from previous position. Here, the sensors are moving in a stationary speed and fixed direction (x-axis).

$xSi_j = v_i.t_{m;}$ Here, t_m = time between beacons m-1th and mth measurement and v_i = sensor S_i's speed.

Now, applying Pythagorean Theorem the distance of the beacons new position to the sensors initial position is calculated.

$d_{12} = \sqrt{(d1_2 2)^2 + (x1_1)^2}$; $d1_2 2$ = distance between sensor 1's second position to beacons second position as in Fig. 2.

$d_{22} = \sqrt{(d2_2 2)^2 + (x2_1)^2}$; $d2_2 2$ = distance between sensor 2's second position to beacons second position.

$d_{32} = \sqrt{(d3_2 2)^2 + (x3_1)^2}$; $d3_2 2$ = distance between sensor 3's second position to beacons second position.

This process is repeated six times from six random positions of the beacon with six different positions of the sensors to find the distance from beacon's new position to sensors initial position. For sensor 1, the process is shown in Fig. 2.

Then calculating those distances, the values of augmented matrix is originated, as in Section IV. From that matrix six unknowns ($X_{n,}$ n = 1, 2, 3… 6) of (16) is found. After that, the inter sensor distances of the initial position is generated, as alluded in Section IV. Thereafter, the coordinate of the Projected point P_4 as shown in Fig. 1 and distances from sensors to P_4 is calculated. Then the initial coordinate of the sensors is found as in Table I. In addition, after applying linear transformation with respect to beacon Table II is generated.

By, adding the distance travelled by the sensors from the first position to the sixth position with x-axis; the current coordinates of the sensors are found, as shown in Table III.

$xi = xi_1+xi_2+xi_3+xi_4+xi_5$

TABLE. III. CURRENT COORDINATES

Sensors	Coordinates	Sensors	Coordinates
S_4	(0, 0, 0)	S_2	$(-x_2+x2, y_2-y_4, -z_4)$
S_1	$(-x_4+x1, -y_4, -z_4)$	S_3	$(-x_4+x3, y_2-y_4, -z_4)$

VI. ANALYSIS

Our method is for a specific scenario, where only one beacon is necessary to determine the coordinates of mobile submerged sensors. Most of the localization methods depend on distance measurements and usually lots of sensors and beacons are deployed. Therefore, precise measurement of the distance is one of the most important factors for accurate localization.

In our proposed model, the beacon floats on the water surface and a minimum of three mobile sensors are deployed underwater. Most importantly, our method determines the 3D coordinates of mobile sensors with respect to the beacon node. So the coordinates of the sensors are calculated more accurately as the coordinate of the beacon node can be measured precisely using Global Positioning System (GPS).

A. Distance Measurement Complexity

The limitations of underwater acoustic signal are considered in this model during distance measurements. The method is simple and understandable but it gives accurate results when the transmission loss of the signal is calculated precisely. Considering some of the practical applications, a pragmatic assumption is considered where the beacon should have the capability to receive signal (R_x). On the other hand, the sensors would transmit signal (T_x). In Fig. 3 we see the relation between the distance and transmission loss as the distance is higher, the rate of transmission loss is also high.

The transmission loss depends on several factors like salinity, depth, acidity, temperature, bubble curtain or other damping structure. While measuring TL, these factors must be under consideration. For a constant frequency, the distance increases with the increase in transmission loss and vice versa. Fig. 3 shows the relation between transmission loss and distance.

B. Error Generation

In our method, we have found less error while measuring the distance because the distance measurement method only depends on frequency and transmission loss. The sensors generate the signals initially instead of decoding a message from the RSS, as mentioned in Section III.

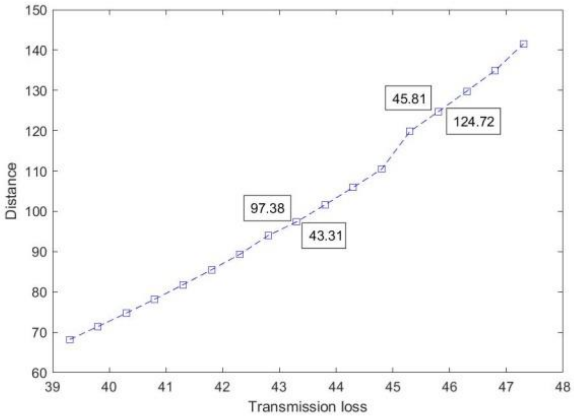


Fig. 3. Relation between the Distance and Transmission Loss.

In acoustic signal propagation, the transmission loss depends on various factors. Therefore, more accurate transmission loss calculation would give a better distance measurement resulting initial and mobile coordinate estimation with less error. In our technique, we have not used any bouncing technique while measuring the distance as the bouncing technique suffers from multipath fading.

VII. Simulation Results and Discussion

The proposed strategy is simulated using MATLAB to validate the mathematical model. The sensors are placed randomly at (0, 0, 0), (0, 70, 0) and (85, 90, 0). The beacon is randomly moved in a plane, parallel to the XY plane. The positions of the beacon are given in Table IV.

One of the sensors is situated at the origin and another one on the y-axis to avoid computational complexity. We have added some Gaussian Noise with the Euclidean distance to find the coordinates of the sensors. After implementing the trilateration, the final coordinates of the sensors are found. Moreover, by using the method based on Lambert function for distance measurement, at a static frequency of 45 kHz, the initial coordinates of the sensors are found. After that, by adding the distance moved by the sensors with respect to the x-axis, the current coordinates are found. Here, the distance travelled by the sensors is 239.2395m.

In Fig. 4, Initial coordinates for sensor S_2 using Lambert function for distance measurements is denoted as S2 and current coordinate of sensor S_2 is denoted as S2′.

In Fig. 5, Coordinates of sensor S_1, S_2 and S_3 using Euclidean distances are denoted as S_1, S_2, S_3 and coordinates using the method established on Lambert function for distance measurements are denoted as S_1', S_2', S_3', respectively.

TABLE. IV. BEACONS COORDINATES

	B_1	**B_2**	**B_3**	**B_4**	**B_5**	**B_6**
x	100	90	80	-10	-20	-30
y	90	80	70	60	-60	-90
z	70	70	70	70	70	70

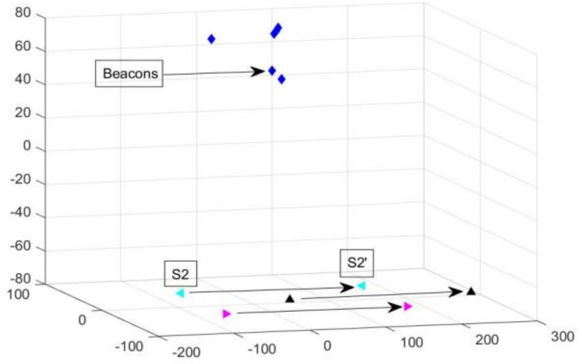


Fig. 4. Current Coordinates of the Sensors.

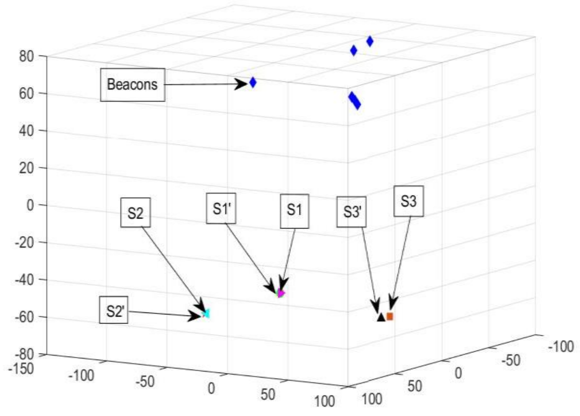


Fig. 5. Comparison between Final Coordinates using Euclidean Distances and Experimental Distances.

Table V compares the error in final coordinates of sensor S_1, S_2 and S_3 as the distances between the beacon and sensors are calculated using the proposed method with when distances between the sensors and beacon are calculated using Euclidean distances. Here, the coordinate of S_3 is showing maximum error. Error in S_1 and S_2 are negligible.

The positional errors of the sensors are given in Table VI. Error for S_1 is negligible where the error is less than a meter. In addition, for S_2 it is a bit above 1.5m. The error for S_3 is comparatively high as it is above 7m.

Positional error of sensors generated with the proposed model is moderate. This also proves the importance of precise evaluation of TL. Table VII compares the error in coordinates at different frequencies. The actual coordinate is measured at frequency 45 kHz, which is denoted at Table V.

The percentage of error increases with the error in frequency.

TABLE. V. COORDINATE ERROR OF SENSORS

Sensors	**Actual Coordinate (x, y, z)**	**Experimental Coordinates (x, y, z)**	**Percentage of Error (%)**		
			x	***y***	***z***
S_1	(-102.87, -88.44, -70)	(102.73, -88.61, -70)	0.16	-0.19	0
S_2	(-102.88, -8.09, -70)	(-102.73, -9.78, -70)	0.14	-20.88	0
S_3	(13.97, -55.19, -70)	(12.19,-8.12, -70)	12.79	12.80	0

TABLE. VI. POSITIIONAL ERROR OF SENSORS

S_1	**S_2**	**S_3**
0.2288m	1.694m	7.287m

TABLE. VII. ERROR COMPARISION AT DIFFERENT FREQUENCIES

Sensors	Frequency	x	y	z
S_1	44.95kHz	0.058%	0.167%	0%
	44.85 kHz	0.174%	0.501%	0%
	44.75 kHz	0.286%	0.829%	0%
S_2	44.95 kHz	0.058%	-10.041%	0%
	44.85 kHz	0.174%	-27.828%	0%
	44.75 kHz	0.286%	-43.122%	0%
S_3	44.95 kHz	12.705%	12.303%	0%
	44.85 kHz	41.279%	32.400%	0%
	44.75 kHz	71.086%	48.042%	0%

VIII. CONCLUSION AND FUTURE WORK

In this paper, a mathematical model is presented to localize submerged mobile sensors using only one beacon node. A method based on Lambert-W function is used to measure the distances between the beacon and the sensors and the coordinates of the sensors are determined using Cayley-Menger determinant. Where all the sensors are moving in the same direction along the x-axis, and the sensors speed are static and known. Moreover, our distance measurement technique contributes less error and does not need any kind of synchronization. Simulation result validates that there are some error between the Euclidian distance and the experimented distance; resulting in erroneous coordinates. However, precise measurement of Transmission Loss gives accurate distance; finally leading to flawless coordinate determination. Therefore, the accurate measurement of Transmission Loss gets utmost priority in this approach.

In future, we plan to localize the sensors, moving in different directions and unknown speed.

REFERENCES

[1]. H.P. Tan, R. Diamant, W.K.G. Seah, and M. Wanldmeyer, "A survey of techniques and challenges in underwater localization," Ocean Engineering, vol. 38, pp.1663-1676, 2011.

[2]. J.H. Cui, J. Kong, M. Gerla, and S. Zhou, "The challenges of building mobile underwater wireless networks for aquatic applications," Network, IEEE, vol. 20, pp. 12-18, 2006.

[3]. P. Xie, J. H. Cui, and L. Lao, "VBF: vector-based forwarding protocol for underwater sensor networks," Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems, pp. 1216-1221, 2006.

[4]. A. Rahman, V. Muthukkumarasamy, E. Sithirasenan, "Coordinates determination of submerged sensors using cayley-menger determinant," IEEE International Conference on Distributed Computing in Sensor System, 2013.

[5]. T.C. Austin, R. P. Stokey, and K. M. Sharp, "PARADIGM: a buoy-based system for AUV navigation and tracking," in OCEANS 2000 MTS/IEEE Conference and Exhibition, 2000, pp. 935-938 vol.2.

[6]. M. Hosseini, H. Chizari, T. Poston, M. Bt. Salleh, A. H. Abdullah, "Efficient underwater RSS Value to distance Inversion Using Lambert Function," Mathematical Problems in Engineering, vol.2014, Article ID 175275, 2014.

[7]. P. Duff and H. Muller, "Auto calibration algorithm for ultrasonic location systems," in Wearable Computers, 2003. Proceedings. Seventh IEEE International Symposium on, 2003, pp.62-68.

[8]. E. Olson, J. Leonard, and S. Teller, "Robust range-only beacon localization," in Autonomous Underwater Vehicles, 2004 IEEE/OES, 2004, pp.66-75.

[9]. J. Guevara, A. R. Jimenez, A. S. Morse, J. Fang, J. C. Prieto, and F. Seco, "Auto-localization in Local Positioning Systems: A closed-form range-only solution," in Industrial Electronics (ISIE), 2010 IEEE International Symposium on, 2010, pp.2834-2840.

[10]. J. Guevara, A. Jiménez, J. Prieto, and F. Seco, "Auto-localization algorithm for local positioning systems," Ad Hoc Networks, 2012.

[11]. "Principles of Underwater Sound," Publishing, 3rd edition, 1983.

[12]. R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert-W function", Advances in Computational Mathematics, vol.5, no. 4, pp.329-359, 1996.

[13]. Yun, Nam-Yeol&NamGung, Jung-Il & Park, Hyun-Moon & Park, Su-Hyeon& Kim, Chang-Hwa, "The underwater environment monitoring system based on ocean oriented WSN (Wireless Sensor Network)," Journal of Korea Multimedia Society, 2010.

[14]. A. Rahman, V. Muthukkumarasamy, E. Sithirasenan, "Localization of submerged sensors using radio and acoustic signals with single beacon," Cichoń J., Gębala M., Klonowski M. (eds) Ad-hoc, Mobile, and Wireless Network, ADHOC-NOW, 2013.

[15]. M. Hosseini, H. Chizari, C. K. Soon, R. Budiarto "RSS-based distance measurement in underwater acoustic sensor networks: An application of the Lambert-W function", International Conference on Signal Processing and Communication System, 2010.

[16]. N. Patwari, P. Agrawal, "Calibration and Measurement of Signal Strength for Sensor Localization, IGI Global, 2009.

A Proposed Method to Solve Cold Start Problem using Fuzzy user-based Clustering

Syed Badar Ud Duja[1], Baoning Niu[2]*, Bilal Ahmed[3], M. Umar Farooq Alvi[4]
Muhammad Amjad[5], Usman Ali[6], Zia Ur Rehman[7], Waqar Hussain[8]
College of Information and Computer, Taiyuan University of Technology, Taiyuan, Shanxi, China

Abstract—**With the elevation of the online accessibility to almost everything, many logics, systems and algorithms have to be revised to match the pace of the trends among the socialized networks. One such system; recommendation system has become very important as far as the socialized networks are concerned . In such paced and vibrant environment of the online accessibility and availability to heavy and large amount of data uploaded to the internet such as, movies, books, research articles and much more. The method of recommendation where provides the socialized networks between the operators, at the same instance, it provides references for the users to asses other users that effects their socialized relation directly or indirectly. Collaborative filtering is the technique used for recommending the same taste of picks to that of the user, and it is accomplished by the user's mutual collaboration, this technique is mostly used by the social networking sites. Nowadays this technique is not only popular but common for recommending the data to the user; meanwhile it also motivates the researchers to find the more effective system and algorithm so that the user's satisfaction can be achieved by recommending them the data according to their search history. This paper suggests the CF (Collaborative Filtering) model that is based on the user's truthful information applied by the FCM (Fuzzy C-means) clustering. This study proposes that the fuzzy truthful information of the user is to be combined with rating of the content by other users to produce a recommender system formula with a coupled coefficient with new parameters. To achieve the results the Data set of Movie Lens is included in the study which shows significant improvement in the recommendation subjected to the condition of cold start.**

Keywords—Recommender system; collaborating filtering; cold start problem; clustering; user based clustering

I. Introduction

Commercial e-business and e-commerce websites or social media networking websites can depicted as the instance of recommender system. These kind of online attractions use recommender systems to suggest the likewise articles and content to the viewers. It is used to filter the information available on the internet related to any particular content to suggest that content to the user. Seeking that a likewise content that has been rated or watched frequently by the similar type of users is to be recommended to other users with same taste. The historical interests of the user plays an important role in recommending the content based on the rating and preferences of other users. Recommendation systems are used in the variety of areas like in recommending videos on YouTube, shopping products on Amazon and similar apps on play stores. There are several approaches in use to recommend particular articles to the user such as collaborative filtering, content-based recommendation or personality based recommendation and knowledge based recommendation system. All these approaches are in use depending upon the necessity of the platform for which it is being used. To recommend the content to the user, different types of engines, components and elements are fed by the recommender systems. These components normally consists of data collectors, whose job is to collect the likewise data available on the internet based on the keyword used for the search or other web analytics tools. The second component is data processor which processes the collected data for the third component that is recommender model. Which then processed for the user interface and again finally to the last component that is processing of the data that is recommended by the recommender by the model. These all components take their input from the first clue is given on the platform from the user[1][2]. Thus, such tools that recommends content to the user or group of users that supports them to explore the content they like available on the internet from their interests taken from their search history that has been rated or watched by the other users in that group to let them watch or go through the explicit set of content. Online item recommendation is achieved by the collaborative filtering technique. The past activities performed by the user such as conducting a purchase or selected by that user or other users, are given the numerical values called rating of the item, worked as the aid for the collaborative recommender model to filter out and suggest the content [3]. This technique has been used in many applications and platforms, showing very efficient results in certain environments. Collaborative Filtering recommender system at first step uses rating of the item by judging the similarities between the users rated that item. Thus, this recommender system worked on the assumption that users who liked this item previously will like the recommended content and if this particular user likes this content then the item recommended will also be liked by the users of same interest. Hence, in this system items recommended to the other users is based on the computation of similarities between users and rating of the content by other users. There are many statistics similarity measures can be implemented to determine the similarity of the rating and other statistics of the users. Pearson's correlation is one such statistics approach to determine the suggestions for the recommendation, for which the value can be somewhere between 1 and -1 to determine the extent to which two variables are linearly related to each other. Similarly, cosine similarity measures the similarity between two vectors and degree between them in the inner space of the product. These

*Corresponding Author.

both techniques are themselves interrelated to each other in the context of mean deduction, but dissimilar when it comes to the measure of coefficients [4].

The generation of the top recommendations the most popular problem i.e. cold start problem which is most popular and commonly faced challenge. Different methodologies by the researchers have been proposed to address the problem of cold start. The occurrence of this problem is faced much when there is not much data is available for particular item or for particular user. Basically, it is the problem in the information system to filter out the data available on the internet based on the data like rating of the item and interest of the user, and this data is not sufficient to infer any results for the recommendation system model. In this study it refers to the inadequate interactions between the users and the architecture of the recommended model and the quality of the recommendations are degraded significantly. In addition this study is focused on the rating of the user items to find in the nearest neighbors.

The mentioned problem then resolved by using the one of collaborative filtering technique. Collaborative filtering techniques as mentioned earlier is achieved on two bases that are Item based collaborative filtering and user based collaborative filtering, the techniques are used to recommend items from the action of other or same user respectively. These both techniques in practice are very effective but when it comes to the dynamic environment of the users or group of users, then these techniques are hard to become scalable subjected to the condition that items does not change too much. However, item based collaborative filtering approach is easy to compute offline and re-training is not required that is why KNN technique is the right thing to go for recommendation model. It also provides foundation to the development of recommendation system. However the KNN's learning method is lazy and also no-parametric, and in this technique the results are compiled for the new systems from the database that includes separated data points that are clustered. It is profound and easy to use technique because of its simplicity, despite this; it can overkill the performance of other complex models especially in forecasting the economics situations and its related demographics and data compression etc. In the recommendation system settings the KNN collaborative filtering algorithm is used to crawl through the whole space of user items to determine the neighbors of the user. In this regard, it is intricate and thorny to provide the recommendation in real-time due to its unpleasant approach towards consumption of time while swarming through the large number of users.

II. Realted Work

A. Literature Review

To overcome the issue of Cold User (cold start) in existing techniques e.g. CF many researches proposed their models, architecture and frameworks. Before conducting this study, around 20 studies has been gone through that identified and addressed the same problem while conducting their work. These studies discussed the problem and also proposed their solution to overcome cold start problem. In this study the solution produced by the researchers are reviewed in detail and sectioned according to their nature. The review of the solutions is carried out based on the categories like, cross domain and social network data system, implementation of association rules on the user profile, similarities in ratings and demographics, behavior and historical interest of the users, deep learning strategies, ICHM, CBCF, OWA and CART.

In 2015, another research addressed the cold start problem and generating recommendation for cross-selling items and products on large e-commerce websites like Amazon and eBay[5]. In this study it is mentioned that the cold start problem hinders the functioning of the recommender system when there is insufficient information about the user or inadequate ratings of the particular item. In this study, knowledge based links among the large e-commerce platforms and the sharing and transfer of knowledge about the user and the item among the domain can mitigate the cold start problem. Cross domain recommender system addressed the cold start problem by sharing and transferring information about the item or user from other similar platforms. This, solution has been implemented and now can be seen in many applications of renowned tourism organizations. Suggesting the tourism places from the picture uploaded and tagged by other tourists on the social media networking websites about the places which are not explored by the particular user is the effective application of the cross-domain recommender system [6] [7, 8].

In this regard, researchers developed an application in which cross-domain knowledge sharing technique is used to suggest the travel destination based on the tourists travel interests and suggests likewise destinations based on the geo-tagged pictures uploaded on the social media with the prediction of the weather[8]. In this study it is mentioned that the problem arises when a new tourist is going on a trip. The recommendations for the new tourists from the system are then generated by the other tourist trip planned for the last location. While matching the interests for the particular user from the likewise social media profiles of other users some of the irrelevant and non-related generation of recommendation occurs which infers in incorrect results thus reducing the efficiency of the whole recommendation system.

Fig. 1 shows the cross domain recommendation model in which recommendation for the new users are generated on the bases of their behavior and from the users of the same group matching their interests from the social media networking profiles. So the knowledge about the user behavior, their likes and dislikes is shared with the targeted domain.

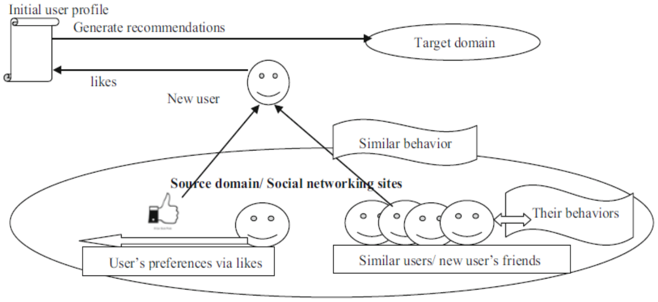


Fig. 1. Cross Domain Recommendation Model.

Shapira, Rokach, and Freilikhman [9] in their research discussed the cold start problem and suggested a solution for this problem. It is suggested that recommendation system can collect data from the Facebook profile of the user and filter out the data related to the particular domain. If there is insufficient information about the interests of the user related to the domain then it may extract the information from the behavior of the user and the reactions posted on the friends profile by that user. This study suggested the cross-domain learning method that is k-nearest neighbor which is source a source aggregation method for collecting the information about the new user in the specific domain. The authors concluded that this aggregation method based k-NN shows significant improvement in the recommended results if the users reacted to the content is not much spaced.

Thus, infers that the cross-domain model functionality is based on the data set density. For resolving the cold start problem the data set density is the significant variable.

Sobhanam concluded in their research that the cold start issue can be addressed by the implementation of association and clustering [10]. The association rules are implemented on the profile of the new user to generate a new enriched profile, also the implementation of the frequency pattern taxonomy based profile aided in the generation of these profiles of the user. This technique can generate the top-N recommendations for the new user.

Park addressed the same issue but in different domain [11]. This is relatively old study but addressed the cold start as it is based on the filterbots algorithm. This study discussed that the naïve filterbot algorithm is used by which the system can be infused with bots or hypothetical user. This algorithm can collect the information about the rating of the item (here item can also be a user). This model extracts the average rating for the item or user that is based on the attributes or association of the specific user or item. It also calculated the similarities between users using demographics of the users and determining its average rate. The infused bots and pseudo users in the user matrix as another user or actual rating of the item are treated by the system by applying algorithms of collaborative filtering for the generation of recommendations. Item based and user based algorithms are used to predict the interests of the user and to determine interest similarities among the users Pearson correlation is used which is given as:

$$S(a,b) = \frac{\sum_{i\in I_a \cap I_b}(r_{a,i}-\overline{r_a})\cdot(r_{b,i}-\overline{r_b})}{\sqrt{\sum_i (r_{a,i}-\overline{r_a})^2}\cdot\sqrt{\sum_i (r_{b,i}-\overline{r_b})^2}}$$

Where:

S = similarity

a, b = users

$(r_{a,i} - r_a) \cdot (r_{b,i} - r_b)$ = difference between the item's rating by user a, b and average rating

$I_a \cap I_b$ = the set of items rated by the user a,b

In recommending the items to the new users based on the demographic similarities and user interest similarities, there is another approach other than the naïve filterbot model that is Triadic aspect model. In the subsequent section some models are discussed with details as part of literature review.

Lam and his team in their research proposed a triadic aspect model which generates the prediction ratings of the item using information of the similar user as its input [12]. It extracts the statistical demographics like gender, age, work and other related information of the user. This model was suggested by the Hoffman to analysis the two-mode and co-occurrence data that has implementations in machine learning, retrieval of information and its filtration. This model predicts the interest of the users based on the triadic aspects includes age, gender and work, through which the interests of the existing users are extracted and generate recommendations for the new users with similar demographics and triadic features. This study concluded that the results are satisfactory when implemented to about 280 different types of new users but where the data set is relatively larger the results produced are not satisfactory and recommendations generated for the new user to address the cold start problem are not precise.

In a research paper it was narrated that predicting users based on their interests and behavior for generating recommendations related to them is like a web intelligence system [13]. The cold start problem discussed and addressed in their studies by suggesting implementation of clustering techniques to framework of item-based collaborative filtering. This research also suggests the integration of the information content into the collaborative filtering. It infers the hybrid approach towards the solution of the cold start problem that is item-Based Clustering Hybrid Method. The features of the content information are clustered through which the preferences of the user rating are complemented. A statistical approach is suggested in the study that linearly combined the likeness between the user ratings. It uses the cosine measures for the rating of clustered content and Pearson measures for the ratings of the users. It is concluded after experimental results that the sparsity in the data is effectively addressed and also significance improvement is noted in the recommendation [14]. It was not the only hybrid model used previously but number of researches suggested hybrid model as one such model was introduced in which collaborative filtering, content based and demographic based models were infused to address the cold start problem [15].

Meng used collaborative filtering method as a different approach to increase the performance of cold start problem [16]. The extortion of the user's interest is achieved by walking through the history interests and cognitive similarity among the alike users to implement the social sub community division through well-known similarity measurement of Pearson and clustering approach based on K-means. The recommendations were generated by building the CART based upon user's static information & distinct group of alike users.

Sakarkar and Deshpande used a dissimilar and unique approach to collect the information about the user. In this study [17] the past educational data of the user was subjected to the k-means clustering technique. It was a somehow like triadic model, but it uses three different features as an input of the recommendation model. Despite educational data, current professional experience and information of the parents are used as an input. However, the classification of the new user is done by the k-nearest neighbor model based on his attendance in the computer based assessment. IMSAA online real dataset was used to carry out the experiment along with movie lens 100k dataset. This technique however, directly addressed the cold start for new user effectively.

Jazayeriy and his team in their study used the approach to effectively used the rating of the items to mitigate the cold start issue for new users [18]. This technique was based on the existing category of the item; it determines the average ranking of the each item in the category and judge against the user's average rating to improve the ranking of the recommended items. It shows an acceptable improvement while tested on the movie lens 100KB, 1MBand 10MB datasets.

B. Existing Recommender Algorithms

The cold start problem is addressed by various researchers in multiple ways however, it is established by now that the mentioned problem have its solution veiled in addressing the effectiveness and efficiency of the algorithms used to determine the input for the recommender model. This lead the study to infer that there are two most common proposed models; collaborative filtering and content based methods. The earlier one according to this study refers that the new user will receive the recommendations based on the likeliness of the item from other similar users with alike preferences. However, it also pointed that the features based analysis of the users is not mandatory. While the later solution according to this study refers that items are analyzed by the said methods (content based) or take in account the features of the users to extract the items from the internet that matches the interests of the new user. However, this study ponders on the distinctive situation of the cold start problem subjected to the condition that new user does not show up with any of the previous ratings. In such situation, both collaborative filtering and content based methods will function to a limited accuracy performing less effectively to address the issue. This way the recommendation will not be generated for the cold user, hence affecting the efficiency of the algorithms. This will elevate the quantity of objects present in systems than amount of objects that have been rated till then by the user especially through thin data availability and this intensify the contention of the cold start problem [19].

The above situation gives rise to two issues; cold start user and cold start item. The earlier issue arises due to insufficient activity turned up from the new user regarding item ratings, this in turns become difficult for the recommended system to find similarity among users. The later mentioned issue arises due to lack of ratings on items from other users, therefore less likely to be picked up by the recommender system.

CF approach will not be proposed keeping the scale of this research in accordance as this technique uses models of deep and machine learning algorithms to predict the rating of the item which were not rated by the users up to adequate level. There number of collaborative filtering algorithms that are based on the Bayesian Networks, semantic models and clustering models. Markov decision process, singular value decomposition and multiple multiplicative factor based models are the examples of model based collaborative filtering [20]. Collaborative filtering accomplished through such algorithms develops the model of the rating of the user at first for providing the recommendation of the item. Probabilistic approach is used for algorithms in this group for visualizing the process of collaborative filtering. In such situations the visualization or imagination of the collaborative filtering process behaves same as if the estimated value of prediction for user is computed based on the rating given by the user on other items.

However, the scope of the study requires that memory based model must be used instead as the contribution of the study is to find the effective way of determining the similarity between the users and among items. The memory based collaborative filtering is used to determine the list of items for the subjected users that is based on the similar behavior among the users. Memory based collaborative filtering can be discussed from both item based and user based perspectives. However, according to the model proposed in this study user-based collaborative filtering will be discussed in the subsequent section.

C. User based Approach

The user based collaborative filtering is adopted to determine the rating of an item received from multiple users, for this purpose the similarity among these users is calculated. Henceforth, the rating for the same item from the subject user can be predicted with the help of computation. As discussed earlier, Pearson correlation coefficient for subjected user will produce the association and connection among user "u" and "v" illustrated in eq 1 as:

$$\text{Sim (u,v)} = \frac{\sum_{i\in I}(r_{u,i}-\overline{r_u})(r_{v,i}-\overline{r_v})}{\sqrt{\sum_{i\in I}(r_{u,i}-\overline{r_u})^2}\sqrt{\sum_{i\in I}(r_{v,i}-\overline{r_v})^2}} \qquad (1)$$

Where;

iϵI and I = over all co-rated items from both u, v users

$r_{u,i}$ = rating of item "i" by the user "u"

$\bar{r}_u$, $\bar{r}_v$ = mean rating of the user "r" and "v"

D. Prediction Computation

The prediction for the subjected user for a specific item can be obtained by the formula given below in eq 2:

$$\text{Predict (a,i)} = \overline{r_a} + \frac{\sum_{u\in U} sim(a,u).(r_{u,i}-\overline{r_u})}{\sum_{u\in U}|Sim\ (a,u)|} \qquad (2)$$

The above computation illustrated that, the summation of similarity of the user "a" and "u" and weights of the rating normalized by the sum of the weights on that specific item "i", then summing up this with the mean rating value of the active user "a" can produce the prediction of recommending that item "i" to the active user "a". Whereas, user "u" belongs to the neighbor user "U" in terms of high similarity index while (a,u) represents the range or set of the users with similar rating weights to which the specified user (a) belongs.

E. Fuzzy Clustering and C-means Algorithms

To determine the prediction of rating of an item for the specified user, computation of similarity of rating from the other similar users is achieved using statistical techniques. For this purpose, this study proposes the clustering technique, in which the grouping of similar objects concealed in multidimensional and multivariate datasets and the partitioning of unordered items is done [21]. Using this technique, the scattered and uncategorized data can be gathered into clusters with similar type of items. This statistical approach is widely known in the machine learning enhancement, thus this study emphasize on the algorithm and technique of clustering in order to analyze the adjacency and likeliness of the users. For this reason, the clustering technique is used in the present study is mainly encompass Fuzzy algorithm. The fuzzy clustering is also known as soft clustering, unlike hard clustering the data points in the fuzzy clustering are not strictly belongs to a distinctive cluster but they can belong to more than one cluster at a time [22]. This is main approach of the study for using soft clustering, because the rating of an item may not be only represents all the similar users, but the item may be rated by the unique user or user from different cluster sometimes. The fuzzy clustering becomes very significant when these clusters do not have distinguishable boundary, and have overlapped behavior with neighboring clusters. Using this technique, proposed recommender system will not fall short when the exact similarity within the cluster is not available especially when the membership function is concerned. To achieve the fuzzy clustering the Fuzzy C-means algorithm is used which is known for identifying the fuzzy cluster with minimum cost function [23].

F. Clustering with Fuzzy C-Mean

Fuzzy C-mean algorithm is very similar to K-means algorithm and is one of the most widely used fuzzy clustering algorithms. This algorithm chooses number of clusters and not only one. In this algorithm each data point is randomly assigned with coefficients for being in the respective clusters. This algorithm is iterative until converged but the change between the coefficients is not more the given threshold i.e. denoted by ϵ. According to the scope of this study algorithm determine the clusters for the data point sets "x". This algorithm tries to partition the limited number of collection of n elements i.e. X= {x1……..xn} if the criteria is given for being into the collection of c fuzzy cluster. In this perspective of given finite data set a partition matrix is returned with a list of c cluster centers i.e. C= {c1…….cc} and illustrated as:

$$\underset{c}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} \left\| X_i - C_j \right\|^2$$

Where:

$$w_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| X_i - C_j \right\|}{\left\| X_i - C_k \right\|} \right)^{\frac{2}{m-1}}}$$

However, for a set of data point $X_j \epsilon R_d$, where j=1……N with minimum cost function to find the partition where data point belongs is given as in eq 3:

$$j(U, M) = \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{i,j}^{m} D_{i,j} \tag{3}$$

Where;

$\mu = [\mu_{i,j}]$, i denotes cluster, j denotes its object and μ overall represents the value of fuzzy membership in given cluster of given object and m belongs to the range of [1, ∞].

However, the value of the distance between xj and mi is denoted by $D_{i,j}= D(x_j,m_i)$.

The algorithm in the continuity of above formulation can be developed as:

Step 1 Value of the m, c and ϵ will be defined along with the assigning step value t = 0 and initializing the mean matrix.

Step 2 When t= 0 it computes the membership matrix μ, else update when t<0 using the below formulation:

$$\mu_{i,j}^{(t+1)} = \frac{1}{\sum_{t=1}^{c} \left(\frac{D_{tj}}{D_{i,j}} \right)^{1/(1-m)}}$$

If i=1,….,c and j=1,….,N.

Step 3 From the preceding step, μi,j will take new values and the mean matrix C will be updated accordingly using:

$$C_i^{(t+1)} = \frac{\sum_{j=1}^{N} \left(u_{i,j}^{(t+1)} \right)^m x_j}{\sum_{j=1}^{N} \left(u_{i,j}^{(t+1)} \right)^m}$$

Step 4 Iteration of step 2 and 3 is required until the function converges and minimum cost value is achieved i.e. difference between the change in mean matrix is smaller than the small number ϵ.".

III. Proposed Model

Two algorithms are proposed as the framework of proposed model. One algorithm is for the training of the recommender system, where movie lens data set is presented to the algorithm as input and expected output of the recommender system will be fuzzy user-based measurement of similarity and briefs as:

START

Step 1 Access Movie Lens website to Load data set, to construct two matrices

Step 2 Construction of rating matrix of user-movie

Step 3 User based similarity matrix is produced from user-movie rating matrix i.e. preceding step, by implementing Pearson similarity measure.

Step 4 Truthfulness matrix of user is constructed

Step 4.1 Computation of activity of user u (count (Mu)= User_activity (u)

Step 4.2 Probity of user computation

User_Propity (u) = $\sqrt{\frac{\sum_{j \in M} (R_{u,j} - \overline{R_u})}{count\ (M_u)}}$

Where;

R is rating

U is user

J is movie

Ru is average rating for the user (u)

Step 4.3 Computation of the sore of user's neighbor

MAX AVG (ui) = User_friend score (u) where i=1,…..,K(no. of nearest user)

Step 5 Similar users are clustered keeping degree of membership in account based on the truthfulness matrix after implementation of fuzzy c –means clustering. Where range of value of membership is [0,1]

Step 6 The fuzzy based matrix which infers the product of two similarity matrices (i,j) then produced by computing the proposed formula i.e. user-based similarity times fuzzy truthfulness similarity measure illustrated as:

Similarity (i,j)fuzzy based= Similarity user based x C (combination Coefficient) + similarity fuzzy truthfulness x (1-C)

Combining and taking the product of two matrices i.e. user based similarity matrix and similarity of fuzzy truthfulness matrix results in the fuzzy based similarity matrix as illustrated above.

END

The results of the first algorithm infers the fuzzy based similarity matrix, whereas, the second algorithm is used for the testing phase of the recommender system in which the new user with few ratings will be taken as input. This algorithm will produce the recommended items for the new user.

START

Step 1 selecting cold start users from the dataset

Step2 information for the new user's truthfulness is computed

Step 3 implementation of Pearson correlation coefficient on the new user's ratings matrix

Step 4 according to the new users truthfulness matrix clustering

Step 5 proposed similarity formula is computed i.e. $j(U,M) = \sum_{i=1}^{c} \sum_{j=1}^{N} \mu_{i,j}^{m} D_{i,j}$

Step 6 Implementation of prediction similarity formula (from prediction computation section)

Step 7 Fetch results of clustering of truthfulness matrix of new user from Step 4 (highest estimated ratings) and recommend to the new users.

END

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

in the title or heads unless they are unavoidable.

IV. Discussion and Conclusion

The model proposed in this study is tested through experiment by providing Movie Lens data to the developed algorithms. This data was consisting of more than 1600 movies and their corresponding 100 thousand ratings from more than 900 users, is used to extract the ratings form the user and process it further after analysis. Cold start users for this model are selected randomly from the data those who have less ratings. Scrutinize intimately Table I, it is translucent that there lies similarity between the ratings of movies from different users. However, some users have rated very few movies from first 10 movies and some of them did not rated at all. The data for the rating of the first 10 movies from 10 users is included in tabular form in Table I.

As an instance, user 3, 4 & 5 did not rate any of the movies, however user 9 rated only one movie. They have no common ratings, and also user 3, 4 and 5 did not rate first 10 movies. On the other hand user 8 and 9 rated only movie 7. For the purpose of computation and analysis user 4 and user 9 will be considered. In Table III, user 9 rated 12 movies and user 4 rated 13 movies. On measuring the similarity matrix based on user data the similarity between the users will be computed and is shown in the Table II.

It is clear from the user based similarity matrix that the value of similarity between user 4 and user 9 is zero. According to the algorithm 1 the processing is achieved offline which will definitely effective in reduction of recommendation time. The model proposed in this study is produced the user-movie matrix of rating as depicted in Table I and Table II is constructed when the Pearson correlation similarity formula is applied on the user-movie matrix.

However, the computation results of user activity and user probity are shown in Table III which illustrated the whole data set, which is obvious demonstration of the truthfulness of each user and their behavior.

In addition, fuzzy c-means clustering plays important role in the performance of prediction. According to the demand of the study the construction of fuzzy matrix is vital because producing clusters of truthfulness of the users to know about the belonging of the user which is illustrated in Table IV.

These values are the membership values of the user belong to the respective cluster and to be used to compute the similarity between users using the formula developed in the equation 3. After this computation the sparsity problem is mitigated, hence improve the prediction accuracy especially with cold start problem. The depicted two clusters are computed with fuzzy c-means which illustrates that, for instance, user 1 have 32% membership of cluster 1 and 68%membership of cluster 2. The accuracy in prediction cannot be achieved if it is being done only on the bases of user-based similarity matrix in user-based collaborative filtering especially for the users who does not bear common ratings for the same item. As for instance, user 4 is neither share ratings as of 6, nor as of 9, hence similarity value is zero. However, if the model proposed in this study is used, then the similarity and prediction can be achieved for the new users using the upgraded equation 3 according to which user truthfulness plays the key role to successfully recommend the item to the cold user. From results and explanation of the suggested model it becomes translucent that mitigation of sparsity and cold user problem are well addresses through the vigorous fuzzy user-based collaborative filtering is effusive with the proposed framework.

TABLE. I. USER-MOVIE RATING MATRIX

USERS	MOVIES									
	1	2	3	4	5	6	7	8	9	10
User1	5	3	4	3	3	5	4	1	5	3
User2	4	0	0	0	0	0	0	0	0	2
User3	0	0	0	0	0	0	0	0	0	0
User4	0	0	0	0	0	0	0	0	0	0
User5	0	0	0	0	0	0	0	0	0	0
User6	4	0	0	0	0	0	2	4	4	0
User7	0	0	0	5	0	0	5	5	5	4
User8	0	0	0	0	0	0	3	0	0	0
User9	0	0	0	0	0	0	4	0	0	0
User10	4	0	0	4	0	0	0	0	4	0

TABLE. II. USER-USER SIMILARITY MATRIX

User	MOVIES									
	1	2	3	4	5	6	7	8	9	10
1	1.0000	0.9545	0.8555	0.9318	0.9285	0.9527	0.9401	0.9754	0.9690	0.9677
2	0.9545	1.0000	0.9522	0.9918	0.9829	0.9565	0.9624	0.9664	0.8907	0.9770
3	0.8555	0.9522	1.0000	0.9484	1.0000	0.8808	0.8721	0.8785	0.0000	0.9214
4	0.9318	0.9918	0.9484	1.0000	1.0000	0.0000	0.9058	0.9816	0.0000	1.0000
5	0.9285	0.9829	1.0000	1.0000	1.0000	0.9355	0.9036	0.9537	0.8807	0.9340
6	0.9527	0.9565	0.8808	0.0000	0.9355	1.0000	0.9579	0.9885	0.9583	0.9796
7	0.9401	0.9624	0.8721	0.9058	0.9036	0.9579	1.0000	0.9645	0.9337	0.9772
8	0.9754	0.9664	0.8785	0.9816	0.9537	0.9885	0.9645	1.0000	1.0000	0.9839
9	0.9690	0.8907	0.0000	0.0000	08807	0.9583	0.9337	1.0000	1.0000	0.9931
10	0.9677	0.9770	0.9214	1.0000	09340	0.9796	0.9772	0.9839	0.9931	1.0000

TABLE. III. USER TRUTHFULNESS MATRIX

	User_activity	**User_probity**	**User_Friend_score**
User1	272	1.47	4
User2	62	1.75	4
User3	41	1.54	4
User4	13	0.33	3
User5	274	2.48	4
User6	200	1.32	5
User7	390	1.76	3
User8	52	02.44	3
User9	12	0.22	2
User10	630	1.37	3

TABLE. IV. CLUSTERING TRUTHFULNESS INFORMATION MEMBERSHIP VALUE

Clusters	USERS									
	1	2	3	4	5	6	7	8	9	10
Cluster1	0.32	0.0094	0.0127	0.028	0.0404	0.113	0.891	0.0107	0.0288	0.0558
Cluster2	0.68	0.9906	0.9873	0.972	0.9596	0.887	0.109	0.9893	0.9712	0.9442

REFERENCES

[1] Samarinas, C. and S. Zafeiriou, Personalized high quality news recommendations using word embeddings and text classification models. 2019, EasyChair.

[2] Zhao, J., H. Wang, and H. Zhang, A Regression-Based Collaborative Filtering Recommendation Approach to Time-Stepping Multi-Solver Co-Simulation. IEEE Access, 2019. 7: p. 22790-22806.

[3] Bharti, R. and D. Gupta, Recommending Top N Movies Using Content-Based Filtering and Collaborative Filtering with Hadoop and Hive Framework, in Recent Developments in Machine Learning and Data Analytics. 2019, Springer. p. 109-118.

[4] Abdelwahab, A., et al. Collaborative filtering based on an iterative prediction method to alleviate the sparsity problem. in Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services. 2009. ACM.

[5] Cantador, I., et al., Cross-domain recommender systems, in Recommender systems handbook. 2015, Springer. p. 919-959.

[6] Arain, Q.A., et al., Intelligent travel information platform based on location base services to predict user travel behavior from user-generated GPS traces. International Journal of Computers and Applications, 2017. 39(3): p. 155-168.

[7] Memon, I., et al., Travel recommendation using geo-tagged photos in social media for tourist. Wireless Personal Communications, 2015. 80(4): p. 1347-1362.

[8] Memon, M.H., et al., GEO matching regions: multiple regions of interests using content based image retrieval based on relative locations. Multimedia Tools and Applications, 2017. 76(14): p. 15377-15411.

[9] Shapira, B., L. Rokach, and S. Freilikhman, Facebook single and cross domain data for recommendation systems. User Modeling and User-Adapted Interaction, 2013. 23(2-3): p. 211-247.

[10] Sobhanam, H. and A. Mariappan. Addressing cold start problem in recommender systems using association rules and clustering technique. in 2013 International Conference on Computer Communication and Informatics. 2013. IEEE.

[11] Park, S.-T., et al. Naïve filterbots for robust cold-start recommendations. in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. 2006. ACM.

[12] Lam, X.N., et al. Addressing cold-start problem in recommendation systems. in Proceedings of the 2nd international conference on Ubiquitous information management and communication. 2008. ACM.

[13] Li, Q. and B.M. Kim. Clustering approach for hybrid recommender system. in Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003). 2003. IEEE.

[14] Wu, Z., Y. Chen, and T. Li. Personalized recommendation based on the improved similarity and fuzzy clustering. in 2014 International Conference on Information Science, Electronics and Electrical Engineering. 2014. IEEE.

[15] Basiri, J., et al. Alleviating the cold-start problem of recommender systems using a new hybrid approach. in 2010 5th International Symposium on Telecommunications. 2010. IEEE.

[16] Meng, C., et al. A method to solve cold-start problem in recommendation system based on social network sub-community and ontology decision model. in 3rd International Conference on Multimedia Technology (ICMT-13). 2013. Atlantis Press.

[17] Sakarkar, G. and S. Deshpande, Clustering based approach to overcome cold start problem in intelligent e-learning system. International journal of latest trends in engineering and technology (IJLTET), 2016. 7(1).

[18] Jazayeriy, H., S. Mohammadi, and S. Shamshirband, A fast recommender system for cold user using categorized items. Mathematical and Computational Applications, 2018. 23(1): p. 1.

[19] Xu, Z. and Q. Fuqiang. Collaborative filtering recommendation model based on user's credibility clustering. in 2014 13th International Symposium on Distributed Computing and Applications to Business, Engineering and Science. 2014. IEEE.

[20] Su, X. and T.M. Khoshgoftaar, A survey of collaborative filtering techniques. Advances in artificial intelligence, 2009. 2009.

[21] Javed, K., M. Iqbal, and S. Mehak, Competitive Analytics from Social Media for Five Leading Sportswear Stores. INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS AND SOCIAL SCIENCES, 2019. 9(4).

[22] Kruse, R., C. Döring, and M.-J. Lesot, Fundamentals of fuzzy clustering. Advances in fuzzy clustering and its applications, 2007: p. 3-30.

[23] Tidke, B., R. Mehta, and D. Rana, A novel approach for high dimensional data clustering. International Journal of Engineering Science and Advanced Technology (IJESAT), 2012. 2(3).

Significance of Electronic Word of Mouth (e-WOM) in Opinion Formation

Javaria Khalid[1], Aneela Abbas[2]
Muhammad Qasim Mahmood[4], Arslan Tariq[6]
Madiha Khatoon[7], Samreen Azhar[9]
Department of Computer Science and IT
University of Lahore
Gujrat Campus, Pakistan[1, 2, 4, 6, 7, 9]

Rida Akbar[3], Rafia[5], Ayesha Akbar[8], Asra Meer[10],
Muhammad Junaid Ud Din[11]
University of the Punjab, Pakistan[3]
University of Gujrat, Pakistan[8]
Gift University, Pakistan[5, 10]
Pakistan Engineering Council (Electrical Engineering)[11]

***Abstract*—In the realm of interconnected digital world, social ranking systems are readily used in different sections of society, for several reasons. The private and public sectors both are making use of social ranking systems as a tool to engineer human behavior, and crafting a digitally stimulated social control. Online reviews and ratings are one of the significant marketing strategies of online sellers to steer out consumers' opinion and ultimately their purchasing decisions. Buyers usually go through these reviews and ratings while purchasing online product or hiring online services. Online consumer reviews, recommendations for product and services, and peer viewpoints play a significant role in the customer's opinion formation. Different online forums of product reviews, ratings and recommendations differ in their objectives, functions, and characteristics. This paper focuses upon a systematic literature review and comparative study of the influence the positive and negative reviews and ratings of the products, automobile services, movies, restaurants, products and services on OLX & eBay, etc. have on opinion formation. Moreover, how these reviews influence others opinions of buying and using the products, services and apps will be analyzed.**

Keywords—Component; E-WOM (Electronic word of mouth); opinion formation; positive reviews; negative reviews

I. Introduction

The growth of online social platform has radically altered the way people find information about other people, products and services and how people interact with others [1]. Social media, the internet-based digital and social networking platforms, offer users the opportunities to link with others, form groups, interact with people, produce messages, share content, write comments, respond and follow one another in a cybernetic community" [2]. In the era of social media, individuals generate and circulate content and try to have a thoroughgoing impact upon masses. Consequently, the information traded in digital social networking platforms is termed by many as "user-generated content" [3].

Electronic word-of-mouth communication (e-WOM) is defined by Goldsmith (2006) as "word-of-mouth communication on the Internet, it can be dispersed by social applications as online forums, reviews sites, and social networking sites". It is considered an indispensable source of information that affects human behavior. Customers are encouraged to give their reviews of the product since 1995 in a large number of 10 million and it was enabled by Amazon.com [4]. Online word-of-mouth (e-WOM) platforms have become one of the most vital sources of information for modern consumers. A large variety is given such as "explosion in the number, range, completeness, and general availability of online reviews" [5]. Researchers found that e-WOM had a positive impact on the decision of the consumers. It affected the opinions of the people of different ages and genders and helped them in making further decisions. For making buying and usage decisions, advertisement is highly preferred by the people rather than word of mouth but for opinion formation, word of mouth is considered more reliable. A bad experience can develop a bad impression of a product which results in negative impression whereas positive word of mouth can give a positive imprint of products and services [6].

As per a report by a research firm 70% of consumer's favor and trust online product reviews [7]. Online views affect consumer choices in different ways. Airlines, telephone companies, resorts, movies, products and services all are adjudged by online reviews and recommendations. Moreover, 5,000 shoppers across five countries identified their three frequently used resources they use for opinion formation. Online reviews on retailer websites (52%) were considered as one of these three frequently used informative resources. Furthermore, recommendations by friends and family members (49%) and advices from store employees (12%) were also critical in influencing the opinions of the masses [8].

In this research paper, will analyze the literature review of different services, products, and applications along with their positive and negative reviews and how they influence upon opinion formation of an individual.

II. Literature Study Related to Products, Services and Apps Reviews

Smita Dayal (2016) proposes that companies put online social platforms in the zones of societal advertising, communal client associations and organizations, and innovative professional mock-ups. The social media is an eminent forum, which help businesses to work and interact with prospect clients, personnel and other investors [9]. The buying decision of a person is strongly affected by a number of available reviews on social platforms that can be either good or bad. Visual, descriptive and collective reviews have a noticeably optimistic effect on purchasers' decisions [10], hence the apparent menace of clients can be abridged to a great extent. As suggested by Prabha Kiran and Vasantha S. (2015) E-WOM can instigate the buying intents of consumers while they purchase online [11] [12].

Following tables (I-V) give a comprehensive insight of the online products, services and apps reviews and ratings and how they influence the opinions and viewpoints of the other people.

These tables represent comparative literature study of different products like EBAY and OLX as it is elaborated in Tables I and II. Description, results and methodology are also explained in different tables of automobiles, Movies and restaurant rating systems respectively in Tables III, IV and V.

TABLE. I. COMPARATIVE STUDY OF ONLINE PRODUCTS REVIEWS AND RATINGS

Paper title	Description	Results		Methodology
		Factors	**Statistical Results**	
Research on Product Review Analysis and Spam Review Detection (2017) [13]	• Product Review • Preprocessing • Abusive removal • Sentiment Score • Spam Review • Feature Tagging • Popularity Analysis • Results	1. Removal of special characters and punctuations, like (#, ^, *, etc.) 2. Removal of irrelevant and malicious curse words 3. Removal of repeating letters (stemming) E.g. ("happyyy, hungryyy") 4. Abbreviations extending	Not Applicable	Product Rating
Brand awareness research (2016) [14]	The purpose is to know: • The level of brand understanding of the case company • Online travel agency TravelBird. • How customers view the case company and either they see it as an appealing brand.	The important factors when buying travel products. • The Frequency of purchase. • The type of travel product. • Living status of the respondents. • The awareness levels of the logo, brand name, and Images of Travel Bird	Not Applicable	In this study the variables are: **Higher Purchase Rate:** • Travel Product. • Good Quality. • Under 25 age. • Advertisement. • Living Status.
An Analysis Study of Improving Brand Awareness and Its Impact on Consumer Behavior Via Media in North Cyprus(2015) [15]	• It accentuates upon the significance of these dimensions (brand knowledge, brand fidelity, brand image, and end-users behavior) it influences consumer's discernment.	• Brand Loyalty • Brand Image • Customer Behavior • Media **Dependent Variable:** • Brand Awareness	T-Test: 0.322 0.395 0.346 0.334	A linear regression model is highlighted and discussed. The maximum coefficient is acquired by Brand Image which is equivalent to 0.395 with maximum *t*-statistic (6.583).

TABLE. II. COMPARATIVE STUDY OF EBAY AND OLX ONLINE REVIEWS AND RATINGS

Paper title	Description	Results		Methodology
		Factors	**Statistical Results**	
E-bay or Craigslist?: Explaining Users' Choice of Online Transaction Community [16]	• It examines consumers' base for selecting a specific virtual transaction community. Based on transaction cost economics (TCE), we assume business expenses influence buyers' choice. • We look at dissimilar institutional mechanisms and compare them eBay's well repute systems vs. Craigslist's face-to-face native business. • Resource-based view (RBV) model states that the mass of target spectators in online business municipal troubles and it has an influence upon decision.	• Institutional mechanisms • Assist in trust-building • Eradicate intermediate such as PayPal • No contract fee • Free shipping cost	-----	• EBay's recognized systems provides a safe online business atmosphere, while Craigslist's face-to-face indigenous operation causes consumers' security apprehension. • Enlarging RBV study, higher viewers of an item will create a higher purchase on eBay than on Craigslist. • It summarizes and differentiates two kinds of the business community.

TABLE. III. COMPARATIVE STUDY OF AUTOMOBILES ONLINE REVIEWS AND RATINGS

Paper title	Description	Results		Methodology
		Factors	Statistical Results	
Brand Awareness and Customers Satisfaction towards OLA Cabs in Bengaluru North and South Region (2015) [17]	• Firstly, awareness of brand knowledge is important. • End user's satisfaction relates to their emotional reaction correlated with products and services. • It investigates brand knowledge and consumers' liking OlaCabs.	• Convenient • Brand • Quick and Safe • Easy to Book • Less Cost • At time Pick-up and Drop	• Chi-Squrae:182.731 • Degree of Freedom: 4	The methodology which is used is Chi-square, Asymy. Sig and Degree of freedom.
A study of customer satisfaction level of ola and uber paid taxi services with special reference to Pune city (2018) [18]	• Observing about consumers' level of liking and satisfaction of the folks who are using OLA and UBER cabs in Pune city. • Learning and studying different aspects like pricing, market share, revenue models, app convenience, etc.	Taxi Preferred due to these motivational factors: • Satisfaction Level • Complaints • Recommendation • Discount • Payment options	---	The questionnaire and interviews are conducted using OLA and UBER. **RESULTS:** • "Safety" as the most important factor while choosing OLA/UBER. • 54% of respondents favor it.

TABLE. IV. COMPARATIVE STUDY OF MOVIES ONLINE REVIEWS AND RATINGS

Paper title	Description	Results		Methodology
		Factors	Statistics	
Dynamic Effects Among Movie Ratings, Movie Revenues, and Viewer Satisfaction. [19]	• It attempts to explore how ratings of a movie from proficient critics, proletarian communities, and spectators themselves have an impact upon key movie performance measures. • They also come across that high advertising generates high ratings and movie revenues. • This research draws attention to how spectators' viewing, rating histories and movie communities' collective view describes viewer satisfaction.	• Six genres • Sequel • Eight major studio distribution • MPAA rating • 7 major holiday release	• Six genres Thriller (35, 14%), romance (25, 10%), action (51, 21%), drama (50, 20%), comedy (74, 30%), animation (11, 4%) • Holiday release Theatre release (34, 14%), video release (36, 15%)	• This study reveals that ratings are linked with movie performance, as calculated by both movie revenues and spectator satisfaction. • For movie rental companies, movie ratings are an effectual calculation of a member's satisfaction. • Our movie-level data examination entails that marketers should assign more ad dollars to movies that gather early high ratings by professional reviewers.

TABLE. V. COMPARATIVE STUDY OF RESTAURANTS ONLINE REVIEWS AND RATINGS

Paper title	Description	Results		Methodology
		Factors	Statistical Results	
Reviews, Reputation, and Revenue: The Case of Yelp.com [20]	• Do online customer reviews affect restaurant demand? • This question is explored by using a fresh dataset mingling reviews from the website Yelp.com	• Revenue ($) • Rating • Reviews • Elite Reviews • Friends of Reviewers	Reviews: Mean: 3.6 Obs: 14,593 Standard Deviation: 0.9	The impact of customers' reviews: 1. A one-star boost in Yelp rating leads to a 5-9 percent amplify in revenue. 2. Customers are more receptive to more noticeable quality changes. 3. Customers react more strongly when a rating holds more information.
An analysis of online reviews of upscale Iberian restaurants [21]	• It examines the relationships between service quality, food quality, consumer satisfaction, and consumer retention restaurants. • A questionnaire-based survey was held among 400 students served at 10 limited-service restaurants.	**Independent Variable:** • Reliability • Responsiveness • Assurance • Empathy • Food Quality **Dependent Variable:** • Customer Satisfaction	-----	• The purpose of the study was to scrutinize the connection between service and food quality dimensions and consumer satisfaction. • The study supposed that both service and food quality would have an affirmative impact on consumer satisfaction, which in turn would certainly influence consumer retention.

Online Customer Reviews on Restaurants and Expert Opinions: An Integrated Approach [22], [23]	• Online reviews have an impact on consumers' restaurant choices. • Online reviews are the most prominent when it comes to consumers' decision making. • This study attempts to explore the relative significance of online reviews in customers' restaurant choices.	-----	In Customer satisfaction: • Food Quality is Ranked 1st with Mean: 1.8 SD: 1.518. • Restaurant Rating is ranked with Mean: 4.05 S.D: 1.646	• This research endeavors to fill the gap in terms of the association between online reviews and restaurant characteristics that are professed as significant aspects while choosing a restaurant. • The results of the study accentuate that food quality and overall restaurant rating of the restaurant have the maximum influence on consumer's restaurant choices.
The Impact of E-Word-of-Mouth on the Online Popularity of Restaurants: A Comparison of Consumer Reviews and Editor Reviews [24][25]	• It determines the number of online reviews consumers partake a constructive and positive impact upon restaurant performance. • A cafeteria that has a superiority certificate and a good number of reviews promote income and popularity contrasted with the other one.	**Independent Variables:** • No. of Reviews • Restaurant Ranking **Control Variables:** • Food • Service • Atmosphere **Dependent variables:** • Total sales • No of customers • Average check	The number of reviews: Mean: 113.36 S.D: 104.03	• Online consumer remarks determine the significant role in E-WOM and young ones choose to use online platforms instead of outdated offline media. • Consequently, clients who have no dining experience with the restaurants, they should significantly make use of platforms like Facebook, Twitter, YouTube, and Instagram.

III. Significance of Study

Online reviews have now become quite prevalent, influential and a key source of obtaining knowledge and information. Many types of researchers have organized influential studies to comprehend two main queries: "why do consumers have faith and then use the information which is given by comparatively unfamiliar and unidentified persons" and "how do consumers infer the reviews to obtain the information they wish for and are ready to put faith into it". It includes archival data gathered from websites like Amazon.com and Ebay.com and attitudinal data which is gathered using surveys and experiments [26].

In recent times, Mudambi and Schuff (2010) investigated consumers' perception of reviewing assistance while deriving data from Amazon.com. The authors experimented on the impacts of many variables, like the extent of the textual piece in reviews and numerical ratings of products, on customers' discernment of reviews. Quite intriguingly, the authors detected that reviews having acute affirmative or negative ratings are typically found unsupportive by other persons [27]. The reviews by the consumer had an influence on sales of video games and uncovered that negative reviews generally have a better impact on sales, and the number of reviews linked with a product is usually considered as heuristics by customers to measure the common features of the product [28].

Researchers consider that in online shopping mall, online positive and negative reviews of the products are sold [29]. It is indicated that online reviews are appraisal data of products and services put on third-party and trader's sites, which is shaped by the consumers and shape up the prospect consumers' behavior. Online reviews are evaluation information about the various features of customer commodities [30]. On the whole, a quantitative study of how online reviews influence consumer purchasing conduct can be changed into a study of how online reviews influence merchandise sales. Thorough research on online reviews and movie box office revenues, various researchers discovered employing sales to determine customer's buying behaviors was viable in the quantitative measure [31].

Hence, numerous findings demonstrate how online reviews influence consumers purchasing behavior through monthly commodity deal proceedings. Watson, the originator of behavioral psychology, presented the "stimulus-response" model. Based on it, Mehrabian proposed the consumer behavior model, Stimulus-Organism- Response Model (SOR model) [32] as shown in Fig. 1.

In Fig. 1, SOR model depicts that the situation can stimulate purchasers' mind-set and then affects consumer conduct, having the influence of online reviews on buyers' judgment and purchase decision. Below are given Different applications and there positive and negative reviews are elaborated in tabular form and how they impact customer decision making, is discussed.

Many Online Doctor Service Apps and their positive and negative reviews are discussed in Table VI. A survey was conducted about online rating and score of the app and patient reviews and it included responses from more than 800 people. The main responses concluded the following points:

- 74.6% of respondents had investigated doctors, dentists, or medical care online.
- 69.9% said an affirmative online standing is extremely or really important in choosing a healthcare supplier.
- 51.8% of patients who had submitted negative online reviews about a medical practice had never been in touch with the address of their concerns.
- Patient satisfaction doubles when a negative review is addressed [33].

NRC Health's research discovered that 92.4% of clients make use of online reviews to get direction about most of their everyday purchasing decisions.

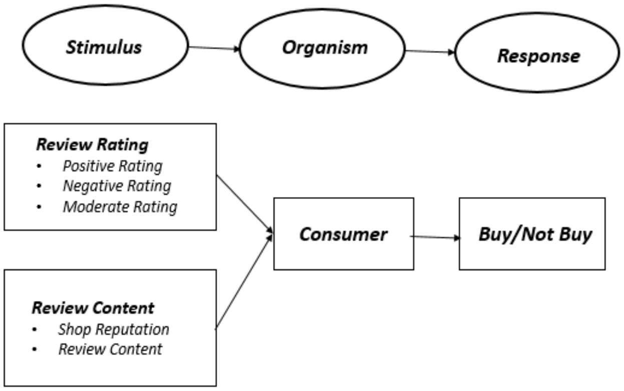


Fig. 1. SOR Model [32].

As said by a Local Consumer Review Survey organized by Bright Local, 97% consumers from the age group of (18-34) study online reviews to judge a local business [34].

Many Online Lawyer Service Apps and their positive and negative reviews are discussed in Table VII. Reviews permit customers to acquire real feedback about the firm to determine if it's good and fit for their needs. A report from Reviewtrackers.com determines that almost **83%** of people checked lawyer reviews at the initial stage to find an attorney. In the 2015 survey done by Bright Local with 2,354 participants was found that 92% people are concerned with online reviews for judging the products and services. In a 2013, A survey of 1,046 individuals was conducted by Zen desk and it concluded that positive reviews affect **90%** of the participant's decisions [37].

Google's search engine algorithm gives preference to legal websites having positive reviews among a diversity of platforms, putting websites of law firm having more reviews higher in the search results.

- 90% customers consult reviews before going to or communicating a law firm.
- 84% of clients have reliance and confidence upon online reviews.
- 74% of respondents say that positive reviews develop believe and trust upon native law firms [38].

Online Automobile Service Apps and their positive and negative reviews are discussed in Table VIII. As everybody knows, Uber is comfortable to ride but most people give them less than five-star rating [39]. Due to the negative rating, the uber driver will get fewer rides and earn less than other drivers who have positive ratings and reviews.

But a new study from New York University found that "the value of rating systems like Uber's decreases over time because of public pressure to give another person a high rating, which continually pushes the average up and up until it becomes fairly irrelevant hence unreliable" [40].

After listing different reviews of lawyers, doctors and automobiles apps, the last Table IX represents names of different online home services applications and their ratings and positive or negative reviews. Furthermore, how these reviews and ratings influence the people opinions and decisions are also evaluated in the above given tables.

TABLE. VI. ONLINE DOCTOR SERVICES APP REVIEW

App Name [33][34][35]	Rating	Positive Review	Negative Review	Impact Upon Customer Decision
I Online Doctor	4.3	"Excellent Application to find and consult with best doctors"	"It's a waste of time fake online doctor"	Yes
		"I refer to my child's pediatric Its balance between my work and family"	"Very immoral customer service"	Yes
Consult By Doc Dr	5.0	"It is very beneficial, Easy to cooperate between Doctor and Patient"	"No one attends call even after entering your credit card detail"	Yes
		"Best online doctor consultation app accomplished by a team of expert doctors"	"No Wicked Comments"	Yes
Call Doc App – Consult Indian Doctors Online	4.8	"Best online consultation app for both patients and doctors"	----	Yes
		Excellent	----	Yes
My Live Doctors - Online Doctor Consultation	4.0	"Great app to provide medical help remotely"	"Fraud app hai. Doctor ki fees 500 le Liya doctor ne koi contact nahi kiya"	Yes
		"This is a good app. You can virtually visit the doctor through video call and can communicate whenever you want"	"May be fake app unable to register as a Dr and no response from the support team"	Yes
Hello Lyf -Online Doctor Consult	4.5	"My Live Doctors is a mobile application that will help you find a doctor and book a free online doctor consultation"	"I wrote my medical condition. I thought I have a reply in minute or hour but am waiting"	Yes
		"Helpful in times of emergencies. I had an emergency situation in an unknown place, this app helped me find doctors nearby"	"Don't work. Give error message every time you try to post something"	Yes

TABLE. VII. ONLINE LAWYER SERVICES APP REVIEW

App Name [36][37][38]	**Rating**	**Positive Review**	**Negative Review**	**Impact Upon Customer Decision**
24 Justice Online Lawyers and Legal Services	5.0	“I strongly approve and commend to use this app for all of your legal affairs & have excellent services.	“No Bad Comment”	yes
		That’s a great app helpful for every man”	----	yes
Lawyers Online	4.5	“This is an awesome piece. It's an app ever Nigerian should have, not just lawyers. Wow”	“I can't open it after updating it”	Yes
		“Free and helpful indeed. Keep informing as you know that laws are not fixed. Appreciations”	“It's OK but it does not carry every decision of the appellate courts”	Yes
Legal Services Link	4.6	“Excellent app”	“This app was absolutely useless in the fact that I was unable to search for an attorney”	yes
		“Outstanding App”	“Kept checking boxes that I didn't mark”	Yes
Got My Legal Help	4.4	“Easy and convenient. It's an all in one app for legal and Emergency situations”	“Useless app. this app take your information to provide legal advisor but nobody responds”	Yes
		“Great piece of mind knowing Love this App!”	“Not too good”	Yes
Introduction to Law	4.5	“It’s okay and not too difficult to understand”	“Not Good enough”	Yes
		“This is the app people of today should download”	-----.	Yes

TABLE. VIII. ONLINE AUTOMOBILES SERVICES APP REVIEW

App Name [39][40]	**Rating**	**Positive Review**	**Negative Review**	**Impact Upon Customer Decision**
Uber	4.0	"My driver was very "professional and polite". Uber is faster, economical and much friendlier service than any of the local taxi"	"When I order a uber, my phone says the driver will arrive in 5 minutes, and ten minutes later, they still in the same place"	Yes
		“Overall, Uber is a great service, they handle all complaints promptly”	“Couldn't create an account because my 'number was previously in use' even though my phone is new”	Yes
Careem - Car Booking App	4.5	“Excellent service, excellent and outstanding help via call service. Complaints are responded quickly”	“Very disappointed to use this service in Pakistan. No option to give feedback on Careem”	Yes
		“Great app!!!! But Not able to buy a package now... There is an issue in payment”	"Extremely wretched app. The service is unpleasant. No call center help is accessible. You have to chat with a representative if u have an issue"	Yes
TAXI Booking -Cab Booking App	3.1	“Very good app”	“A dirty app that I have ever seen Friends don't download this app. It is just for show. I cannot book any cab every time”	Yes
		“Like this app”	"Don't download this app full waste of time and it takes time to open"	Yes
AutoWala	3.1	“It is very good it is very well I like this because where I want to go the auto Wala goes very good”	"Not at all user-friendly. Not useful at the time of need. I wish there should be zero ratings. ask plenty of questions while registering and a problem with location"	Yes
		“Best app for auto”	"Please don't download this app, it's wastage of time. This app is just meant to collect your info & data that's it"	Yes
Cab Booking Online All In One	3.7	“Cars are very beautiful and wonderful and clean”	“Is there any option where I can put rating this is time waste app”	Yes
		“Great Cab booking online app! It simple and has a nice UI and this has also saved my device space”	"Worst of time fake app"	Yes

TABLE. IX. ONLINE HOME SERVICES APP REVIEW

App Name	Rating	`Positive Review	Negative Review	Impact Upon Customer Decision
Doorstep Services - Variety of Home Services	3.8	“Awesome app for door services, easy to use”	“Wanted to fix the plumbing issue at my home. My Bookings section does not show the booking. Uninstalling it”	Yes
		“Good app. with best services”	“I booked for refrigerator repairs, still it shown as pending no response. Pathetic app and services”	Yes
Service Guru: Electrician, Plumber, AC Repair App	3.8	“Provide best services in all area & create customer care section”	"Waste no one is using this app"	Yes
		"Your services are just awesome but please provide your customer care number so customers can do an inquiry about rates of work"	“Not good app”	Yes
House joy-Trusted Home Services	3.6	“Good for getting maintenance and cleaning work at home. Best, smooth and appropriate”	"Booked for Salon service (Mehendi) on the 4th of august itself for 11 august did not get any update. They call on 11 of August at 4:00 PM and inform that they cannot fulfill the request"	Yes
		"Provides salon services. Very Glad about service. All the products used were hygienically sealed and good quality"	"Too bad experience. Raised a complaint about bad repair service. And support informed that they will get back in touch with me but no us"	Yes

IV. DISCUSSION

For marketing different products and services different firms use social media platforms as it has more significant influence upon clients’ decisions and judgments [9]. Online customer reviews and ratings are the prevalent sources of facts for clients who purchase goods from an unfamiliar or even from a familiar website [41]. The primary purpose of this study is to know the influence of products and services’ reviews and scoring upon opinions of the people. The people consider the reviews more reliable and truthful than the recommendations by the professionals and paid experts. For this reason, the major contribution of this study is to explore the impact of the online consumer reviews and e-WOM, on purchasing decisions and hiring services. Fig. 2 elaborates the proposed framework of purchasing decision-making process including different steps and how these are influenced by social identity, optimism, undesirable E-WOM, trust and various individual and environmental factors.

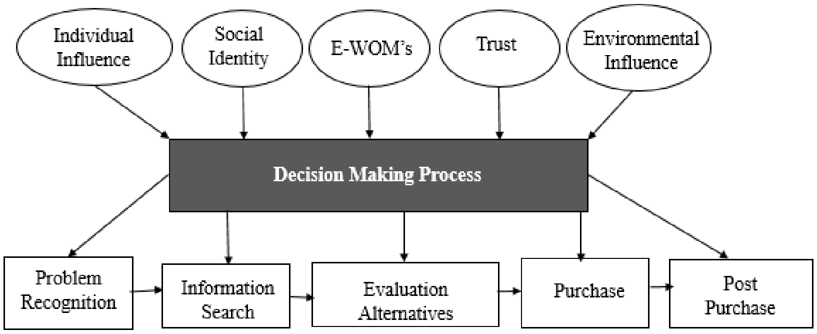


Fig. 2. Decision Making Process [42].

V. CONCLUSION

In nutshell, advancement of technology emerges a new advert of marketing phenomenon of online product review systems, which play a pivotal role in user’s purchase decisions. This research work aims to evaluate the impact of e-WOM (electronic word-of-mouth) and online reviews on the user's buying behavior. A framework SOR is also discussed for evaluating the relation between consumer’s purchase decisions and online reviews. According to this framework, the number of positive reviews impacts positively on the user's decision, neutral reviews do not influence the user and lastly, bad reviews put a negative impact on users’ purchasing decisions. Furthermore, rapidly growing interconnected digital world brings an extensive pool of information via online reviews and consumers preferably rely on this information to eliminate the vulnerabilities of purchasing in a virtual environment. Nowadays, platforms like eBay and OLX have earned a great reputation from user's reviews and transparency in their systems. Customers mitigate the chance of uncertainty in purchasing by checking the past performance of sellers. However, this research work highlights the significance of user reviews across various categories of products and shown facts of different products, services, and applications. Lastly, in future, this research work can be extended and used for evaluating human social behavior on social networking platforms.

REFERENCES

[1] Baptista, J., A. D. Wilson, R. D. Galliers, and S. Bynghall. 2016. “Social Media and the Emergence of Reflexiveness as a New Capability for Open Strategy.” Long Range Planning 50 (3): 322–336.

[2] Ngai, E. W. T., S. S. C. Tao, and K. K. L. Moon. 2015. "Social Media Research: Theories, Constructs, and Conceptual Frameworks." International Journal of Information Management 35 (1): 33–44.

[3] Saboo, A. R., V. Kumar, and G. Ramani. 2016. "Evaluating the Impact of Social Media Activities on Human Brand Sales." International Journal of Research in Marketing 33 (3): 524–541.

[4] Kristopher Floyd, Ryan Freling, Saad Alhoqail, Hyun Young Cho and Traci Freling, "How Online Product Reviews Affect Retail Sales: A Meta-analysis", Journal of Retailing, 2014.

[5] Pradeep Racherla, Wesley Friske,"Perceived 'usefulness' of online consumer reviews: An exploratory investigation across three services categories", Electronic Commerce Research and Applications 11(2012)548–559.

[6] Nawaz Ahmad, Jolita Vveinhardt, Rizwan Raheem Ahmed, "Impact of Word of Mouth on Consumer Buying Decision", European Journal of Business and Management, Vol.6, No.31, 2014.

[7] Kaynar, O., & Amichai-Hamburger, Y. (2008). The effects of cognition on Internet use revisited. Computers in Human , 24(2), 361-371.

[8] Ali Yayli ,Murat Bayram,"Ewom: The effects of online consumer reviews on purchasing decision of electronic goods".

[9] Smita Dayal, "An analysis of social media influence on online behaviour of Indian customers". XVII International Seminar Proceedings,2016, 887-906.

[10] Zan Mo, Yan-Fei Li & Feng Fan , "Effect of online reviews on consumer purchase behaviour", Journal of Service Science and Management,2015, pp. 419-424.

[11] Prabha Kiran, Vasantha S," Review article- Exploring the impact of online reviews on purchase intentions of customer", American International Journal of Research in Humanities, Arts &Social Sciences, 2015, pp.211-214.

[12] Simona Vinerean, Iuliana Cetina, Luigi Dumitrescu & Mihai Tichindelean (2013). The effects of social media marketing on online consumer behaviour. International Journal of Business and Management, 8(14), PP.66-78.

[13] Shashank Kumar Chauhan, Anupam Goel, Prafull Goel, Avishkar Chauhan and Mahendra K Gurve, "Research on Product Review Analysis and Spam Review Detection", 2017.

[14] Antila, Sara.(2016). Brand awareness research Case: TravelBird.

[15] Karam,A.A. Saydam,S.(2015,January). An Analysis Study of Improving Brand Awareness and Its Impact on Consumer Behavior Via Media in North Cyprus (A Case Study of Fast Food Restaurants). International Journal of Business and Social Science Vol. 6, No. 1.

[16] Thomas Ngo-Ye, "Ebay or Craigslist?: Explaining Users' Choice of Online Transaction Community",Issues in Information Systems Volume 14, Issue 2, pp.382-392, 2013.

[17] Prof.Manjunath.G.(2015). Brand Awareness and Customers Satisfaction towards OLA Cabs in Bengaluru North and South Region. Research journal of social science and management.

[18] Khade,A.A. & Dr. Patil,V.(2018). A study of customer satisfaction level of ola and uber paid taxi aservices with special reference to pune city.International Journal of Management, Technology And Engineering.

[19] Sangkil Moon, Paul K. Bergey, & Dawn Iacobucci "Dynamic Effects Among Movie Ratings,Movie Revenues, andViewer Satisfaction",Journal of Marketing Vol. 74 (January 2010), 108–121.

[20] Michael Luca, "Reviews,Reputation and Revenue:The case of Yelp.com", 2016.

[21] Luís Pacheco, "An analysis of online reviews of upscale Iberian restaurants", Multidisciplinary e- Journal, 2018.

[22] Nefike Gunden, "How Online Reviews Influence Consumer Restaurant Selection",2017.

[23] Dr. Ramazan GÖRAL , Simge TOKAY, " Online Customer Reviews on Restaurants and Expert Opinions: An Integrated Approach", European Journal of Interdisciplinary Studies May-August 2015 Volume 1, Issue 2.

[24] Woo Gon Kim, Jun (Justin) Li, Robert A. Brymer, "The impact of social media reviews on restaurant performance:The moderating role of excellence certificate", International Journal of Hospitality Management 55 (2016) 41–51.

[25] Julie Zhang, Qiang Ye, Rob Law, Yijun Li , " The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews",International Journal of Hospitality Management · December 2010.

[26] Chevalier, J., and Mayzlin, D. The effect of word of mouth online: online book reviews. Journal of Marketing Research, 43, 3, 2006, 345–354.

[27] Mudambi, S., and Schuff, D. What makes a helpful online review? MIS Quarterly, 34,1, 2010, 185–200.

[28] Yang, J., and Mai, E. Experiential goods with network externalities effects: an empirical study of online rating system. Journal of Business Research, 63, 9–10, 2010, 1050–1057.

[29] Park, C. and Lee, T.M. (2009) Information Direction, Website Reputation and eWOM Effect: A Moderating Role of Product Type. Journal of Business Research, 62, 61-67.

[30] Mudambi, S.M. and Schuff, D. (2010) What Makes a Helpful Review? A Study of Customer Reviews on Amazon. com. MIS Quarterly, 34, 185-200.

[31] Duan, W.J., Gu, B. and Whinston, A.B. (2008) The Dynamics of Online Word-of-Mouth and Product Sale An Empirical Investigation of the Movie Industry. Journal of Retailing, 84, 233-242.

[32] Mehrabian, A. and Russell, J.A. (1974) An Approach to Environmental Psychology. The MIT Press, Cambridge.

[33] Christopher Cheney,April 29, 2019.[online].Available at "https://www.healthleadersmedia.com/clinical-care/70-patients-call-online-reviews-crucial-selecting-healthcare-providers", Accessed on 02, Feb.2020.

[34] Andrew Ibbotson, 26 Nov,2018. [online].Available at "https://nrchealth.com/patients-trust-online-reviews/", Accessed on 02, Feb.2020.

[35] Shannon Woodworth And Tuesday Wilson,17 Dec,2018[online].Available at "https://www.nextech.com/blog/online-reviews", , Accessed on 02, Feb.2020.

[36] https://broadly.com/blog/online-reviews-for-lawyers/

[37] Cristopher Bryant, 17 Dec,2016,[online].Available at "https://www.martindale.com/marketyourfirm/blog/the-importance-of-client-reviews-for-attorneys/", Accessed on 02, Feb.2020.

[38] Online Reviews For Lawyers, [online].Available at "https://www.acceleratenow.com/law-firm-review-management/". Accessed on 02, Feb.2020.

[39] Caraoline O' Donovan, 11 April,2017,[online].Available at "https://www.buzzfeednews.com/article/carolineodonovan/the-fault-in-five-stars", , Accessed on 02, Feb.2020.

[40] Aric Jeniks, 5 April,2018,[online].Available at "https://fortune.com/2018/04/05/uber-negative-ratings-stars/", Accessed on 02, Feb.2020.

[41] Feng Zhu & Xiaoquan Zhang (2010). Impact of online consumer reviews on sales: The moderating role of product andconsumer characteristics. Journal of Marketing, 74, pp.133-148.

[42] Nawaz Ahmad, Jolita Vveinhardt, Rizwan Raheem Ahmed, "Impact of Word of Mouth on Consumer Buying Decision", European Journal of Business and Management,Vol.6, No.31, 2014.

Design and Development of Autonomous Pesticide Sprayer Robot for Fertigation Farm

A.M. Kassim[1], M. F. N. M. Termezai[2], A. K. R. A. Jaya[3], A. H. Azahar[4]
S Sivarao[5], F. A. Jafar[6], H.I Jaafar[7], M. S. M. Aras[8]
Centre of Excellence for Robotic, and Industrial Automation (CERIA)
Department of Mechatronics Engineering, Faculty of Electrical Engineering
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, Durian Tunggal, 76100 Melaka, MALAYSIA

***Abstract*—The management of pest insects is the critical component of agricultural production especially in the fertigation based farm. Although the fertigation farm in Malaysia has advantages in the fertilization and irrigation management system, it still lacking with the pest management system. Since almost the insect and pests are living under the crop's leaves, it is difficult and hard labor work to spray under the leaves of the crop. Almost agricultural plants are damaged, weakened, or killed by insect pests especially. These results in reduced yields, lowered quality, and damaged plants or plant products that cannot be sold. Even after harvest, insects continue their damage in stored or processed products. Therefore, the aim of this study is to design and develop an autonomous pesticide sprayer for the chili fertigation system. Then, this study intends to implement a flexible sprayer arm to spray the pesticide under the crop's leaves, respectively. This study involves the development of unmanned pesticide sprayer that can be mobilized autonomously. It is because the pesticide is a hazardous component that can be affected human health in the future if it exposed during manual spraying method especially in a closed area such as in the greenhouse. The flexible sprayer boom also can be flexibly controlled in the greenhouse and outdoor environment such as open space farms. It is expected to have a successful pesticide management system in the fertigation based farm by using the autonomous pesticide sprayer robot. Besides, the proposed autonomous pesticide sprayer also can be used for various types of crops such as rockmelon, tomato, papaya. pineapples, vegetables and etc.**

Keywords*—Pesticide spryer; autonomous robot; fertigation; farm; under crop leaves

I. INTRODUCTION

The agriculture industry is growing from time to time as the demand on its yield abruptly rising in conjunction by the end of 2050, the agriculture yields are expected to be able to support the rapid population growth. From this forecast, the dependency on the agriculture yields to meet the population growth is a concern because the world population is expected to grow by over a third, or 2.5 billion people, between 2009 and 2050 [1]. For the countries that in the phase of developing, the population grew significantly faster compared to the countries that already developed hence result in the requirement on the feedstock that came from the agriculture yields. Agriculture industry become vastly practice all around the globe by make use of the prosperous motherland with the diversity natural resource and geographical advantages, the agriculture become applicable and acceptable to the certain countries because of its promising a good returns but there is a resemblance of the problem faced by all these countries which are in term of pest control.

In chili fertigation farms, pests such as mites, snails, and maggots are a common type of pest that can be found in this farm by making the plants as their source of food and breeding ground. In this case, pest invasion is an unavoidable circumstance but can be controlled by having pesticide spraying periodically. Normally, the worker needs to manually spray the pesticide while wearing protective gear and walking from crop to crop. This method indeed inefficient practice and hazardous chemicals used in spraying can be fatal to the worker even wearing protective gear because research conducted found that the protective gears do not stop the chemical but only reduce the amount of exposure [2]. Based on the studies conducted also, The World Health Organization (WHO) estimates approximately about 3 million cases regarding pesticide poisoning which happened every year, thus causing the death of 220,000 people who especially live in developing countries [3].

II. RELATED WORKS

With flourishing technology that is introduced in this 21st century, there is numerous types of robots been used in agricultural activity starting from the cultivation process to the production process. The autonomous robot had been introduced in various application such is in underwater [4], rescue[5], line following robot based on metal detection[6]. In agriculture field, the usage of robotics in agriculture operation able to help to increase the production process and improve efficiency[7]. One of the types of the robot used in agriculture is for the purpose of pesticide spraying with the ability to navigate in the farm, recognize the target and regulate the spraying mechanism[8]. The use of autonomous robot pesticide sprayer as the substitution of the worker who used conventional pesticide sprayer can be applicable.

Besides, the demand for the agriculture robot also stimulates the consciousness of how important its role in the current and future generations. The survey conducted shows that the demand for robots and drones in agriculture will be expected to be rose from 2018 to 2038. Hence, the usage of the autonomous robot is assumed to rise thus replacing the current labor worker. This granular 20 years market forecast covers all the aspect of the agricultural robots and drones for 16 market categories with the expectation by the end of 2038 [7], the market of the robots and drones in these categories is predicted will close to 35 billion with the viable technology and ongoing market demand by considering its technology and application .

Nevertheless, the common problem with an autonomous robot use in agricultural activity is the navigation method used to able the robot fully-operated with decision making capability. In order to navigate through all the field, there are some research has been done [8-11]. It can be done through infrstructure ready or to be without infrastructure. Some research on RFID based navigation are conducted to be implemented as navigation tools [12-13]. As artificial intelligence (AI) starts to emerge, the current robot should be able to navigate the next movement by the adaptation of the surrounding environment and decide which path it will take. The typical method used in the detection is based on the targeted object orientation or repelled signal emits from the sensor itself then calculates the distance in between it [13-18]. Other than that, there is also the robot that uses the vision observation then accumulates all the acquired data to generate the data fusion that enables the robot to navigate itself through the farm [19].

The second problem with the agricultural robot is due to the dissemination of the pesticide to the crops. Unregulated spraying during the disposition of the pesticide to the crop can lead to the low rate of coverage on leaves, wastage of pesticide and hazardous exposure to workers due to disperse pesticide to the desired target [20]. With regulated spraying by the pump, the higher coverage of dissemination to the crops can be achieved whereby the positioning of each crop was varied from one another in the farm. Furthermore, instead of hiring the workers to do miscellaneous work on the farm which can affect themselves, it can be done by an autonomous agriculture robot thus save the expenses on the labor worker [21].

Lastly, the designed robot used in agriculture having the difference performance index depends on the variable they want to achieve. Certain researchers may focus on UAV based pesticide spraying, localization and motion control of agriculture mobile robot, pest image identification and else [22]. This also same goes to the type of the plant being used as the target which differs from one another in terms of size, leaves density and height. Hence, it would be difficult to decide which designed robot was most successful at the time being.

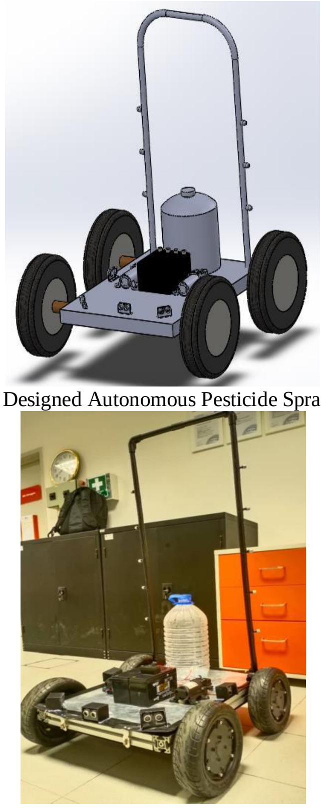
In this research, the aim of this study is to design and develop an autonomous pesticide sprayer for the chili fertigation system. Then, this study intends to implement a flexible sprayer arm to spray the pesticide under the crop's leaves, respectively. This study involves the development of unmanned pesticide sprayer that can be mobilized autonomously. It is because the pesticide is a hazardous component that can be affected human health in the future if it exposed during manual spraying method especially in a closed area such as in the greenhouse. The flexible sprayer boom also can be flexibly controlled in the greenhouse and outdoor environment such as open space farms. It is expected to have a successful pesticide management system in the fertigation based farm by using the autonomous pesticide sprayer robot. Besides, the proposed autonomous pesticide sprayer also can be used for various types of crops such as rockmelon, tomato, papaya. pineapples, vegetables and etc.

III. Hardware Configuration

A. System Construction

The overall design of the autonomous pesticide spraying robot is illustrated in Fig. 1. The design is done using Solidwork software and the development of the autonomous pesticide spraying system based on the design. The specification of the autonomous pesticide spraying robot is shown in Table I.

The dimension of autonomous pesticide spraying robot is determined to be 122 cm (2 feet) because the size of the row for fertigation farm is about 3 feet. In addition, the height autonomous pesticide spraying robot is determined to be 2 m because the normal height of the chili fertigation farm is below 2 m. The system overview for the autonomous pesticide spraying system is illustrated in Fig. 2 shows an overall connection between two different systems that will be combined inside of the autonomous pesticide spraying robot. The development of the autonomous pesticide sprayer prototype consists of two parts where the navigation system and the spraying system. The interconnection between the selected components in the designed robot is crucial and plays a major role to make sure the robot function as desired. Misconnection between the electronic components can lead to malfunction of the designed system thus deviated the operation from achieving the project objective.



(a) Designed Autonomous Pesticide Sprayer.

(b) Developed Autonomous Pesticide Sprayer.

Fig. 1. System Construction.

TABLE. I. AUTONOMOUS PESTICIDE SPRAYING ROBOT SPECIFICATION

Item	***Specification***
Robot dimension	122 cm x 122 cm x 200 cm (L x W x H)
Robot weight	12 kg without payload
Drive system	4-wheeled drive system
Power supply	24V DC lead-acid rechargeable battery
Ground clearance	12 cm from the ground
Payload	Max: 20 kg

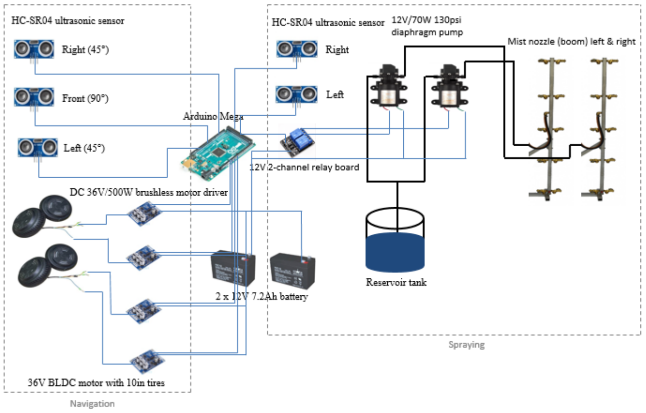

HC-SR04 ultrasonic sensor
Right (45°)
Front (90°)
Left (45°)
Arduino Mega
DC 36V/500W brushless motor driver
36V BLDC motor with 10in tires
Navigation
HC-SR04 ultrasonic sensor
Right
Left
12V/70W 130psi diaphragm pump
Mist nozzle (boom) left & right
12V 2-channel relay board
2 x 12V 7.2Ah battery
Reservoir tank
Spraying

Fig. 2. System Overview of Acceleration-based Movement Detection.

B. Navigation System

The navigation system consists of some ultrasonic sensor, microcontroller, four units of brushless DC motor with a motor driver for each motor, and a 24 V DC rechargeable battery. The microcontroller is the heart of the system where the designer can write and load the program into it to control the sequence and operation of the peripheral that connected to its pin 12 in the microcontroller. Using the programming software which has been predetermined, the coding will be uploaded into the microcontroller which will determine how the designed robot will be operated. In this project, the Arduino Mega 2560 will be used as shown in Fig. 3 because has adequate I/O pins for input and output either analog and digital I/O.

On the other hand, there are eight units of ultrasonic sensor which is mounted at the edge of the frame and the center of each frame. The sensor is an important component in designing and developing the robot with the necessity to move and navigate itself without human intervention. It acts as eyes and ears which will retrieve the data or information from surrounding before sending it to the brain, microcontroller to be processed. [12], where all data were accumulated altogether to generate more accurate and consistent data. The ultrasonic sensor operations are based on the distance calculated from the time interval taken by the emitted sound wave to repel back to the receiver. The ultrasonic sensor which mounted at the center of the frame is fixed perpendicularly 90° facing forward while the ultrasonic sensors mounted at the edge, right and left having 45° deflections each. This concept is referred from the previous design which has been implemented in the wearable device for the visually impaired person [13]. Other than navigation purposes, the ultrasonic sensors will be used to activate the spraying system when the plants were detected in range. Fig. 4 shows the type of ultrasonic used in this prototype.

Fig. 3. Arduino Mega 2560.

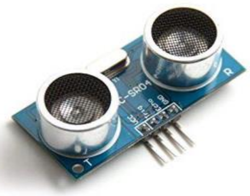


Fig. 4. HC-SR04 Ultrasonic Sensor.

Besides, the four units of brushless DC motor are used for the four-wheeled driving system. The brushless DC motor which is used is manufactured together with the tire that normally applied in a hoverboard. The diameter of the tire which is selected is 25.4 cm to have higher ground clearance about 12 cm. The higher ground clearance is important to pass the irregular surface such as stone or rock along the path. To facilitate the process of BLDC motor rotation and change its direction of rotation, the motor driver is used. With the method used for changing the direction of the motor wheels by the motor driver, it also can be implemented to change the heading and direction of the robot platform with the concept of the hoverboard drive. Hoverboard drive, in essence, use both tires left and right to change the heading and direction of the robot by manipulating the rotation of both tires.

For example, if the robot wants to turn to the right, the left tire is rotating forward while the right tire is rotating backward. This allowed the robot to have curve turning thus change its direction. The 3-phase supply and hall sensor of the BLDC motor will be connected to the out pin and hall pin on motor driver respectively. The 12V battery will provide the supply to BLDC motor by connecting it to the VCC pin of the motor driver. After that, the ZF and VR pin on the motor driver will be connected to the Arduino Mega digital input pins to control the rotation and Pulse Width Modulation (PWM) to the BLDC motor. Fig. 5 shows the hoverboard wheel with the brushless DC motor and the motor driver.

(a) 3-Phase Brushless DC Motor (b) Motor Driver.

Fig. 5. Three-Phase Brushless DC Motor with Motor Diver.

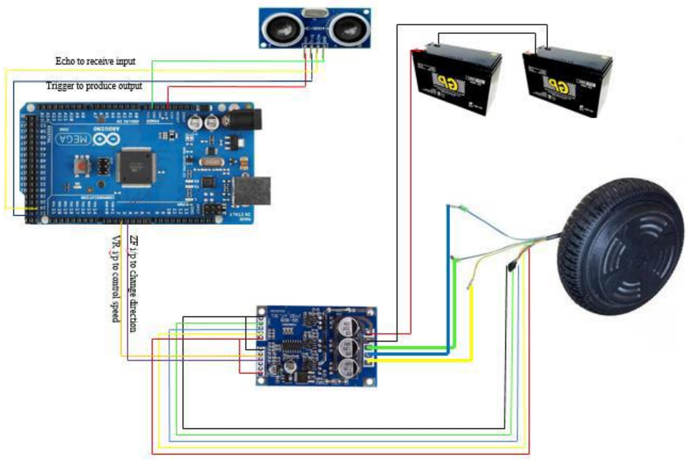


Fig. 6. Connection for Controlling Motor after Ultrasonic Sensor Detection.

The connection of the microcontroller, Arduino Mega 2560, brushless DC motor through 36V/500W brushless motor driver and received the power supply from 24V batteries (2x12V battery in series) as shown in Fig. 6. The motor drivers are able to manipulate the rotation of the motor using its phase connected to the gate driver MOSFET on its circuit.

C. Spraying System

While the microcontroller executing the condition in the navigation part, the condition for the spraying system also will be considered. As the autonomous pesticide spraying robot needs to be able to execute both of the operations simultaneously, the sequence inside of the programming code plays a critical role in the designed project. The main components consist of the spraying system are reservoir tank, pesticide pump, 2-channel relay circuit, tube and some mist nozzles for spraying under the crop leaves. The reservoir tank which is used in the autonomous pesticide spraying robot will be filled with pesticide incapacity of 10L although the maximum of 20 kg of the payload can be carried out.

In order to supply the pesticide from the reservoir tank to the end of the spraying nozzle, the use of the 12V/ 70W 130 psi diaphragm pesticide pump is selected is shown in Fig. 7. The selection of a pesticide pump is crucial because the pump needs to be eligible to push the pesticide out with desired pressure. With the help of the pump, spraying can be directly allocated to the desired targeted plants especially under the crop leaves, by only giving electrical input to the pump which procured by sensor upon detection of the plants.

In term of connection, the microcontroller and the 12V/ 70W 130 psi pump was interconnected through 12V 2-channel relay board where the relay will receive the input signal from the microcontroller to change its contact thus closed the circuit connection from battery to pump hence activating pump. Fig. 8 shows a connection between microcontroller to pump.

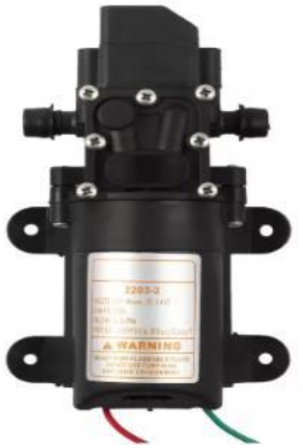


Fig. 7. 12V/70W 130psi Diaphragm Pesticide Pump.

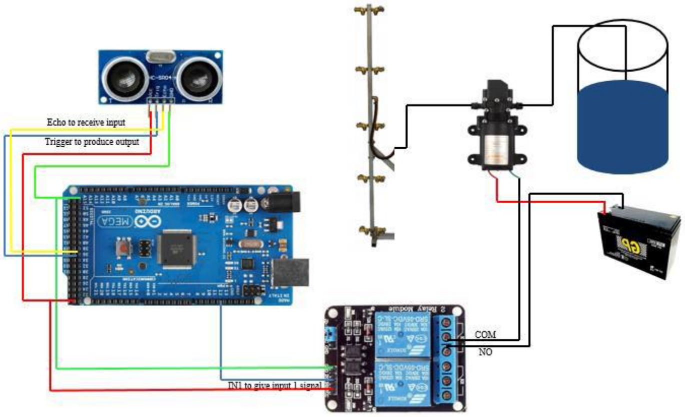


Fig. 8. Connection for Activating after Ultrasonic Sensor Detection.

IV. Autonomous Pesticide Sprayer Operation Flowchart

A. Navigation System Flowchart

Based on the autonomous pesticide sprayer operation, the designed project is divided into two sub-disciplinary part which is the navigation and spraying system. Once the autonomous pesticide sprayer robot activated, it will mobilize through the farm while considering all the operation which is executed simultaneously. The designed project will be regarded to be close to the success after all the execution of the operation undergoes seamlessly and then the robot will be evaluated based on its performance in measurable engineering variables. In order to allow the robot to follow the instruction in the programming code, it is important to identify each step that wants to be executed by the autonomous pesticide sprayer robot step by step as the robot will consider the condition in the top step before moving to the bottom step. Fig. 9 shows the process flowchart inside of the navigation system to make the robot move autonomously throughout the field.

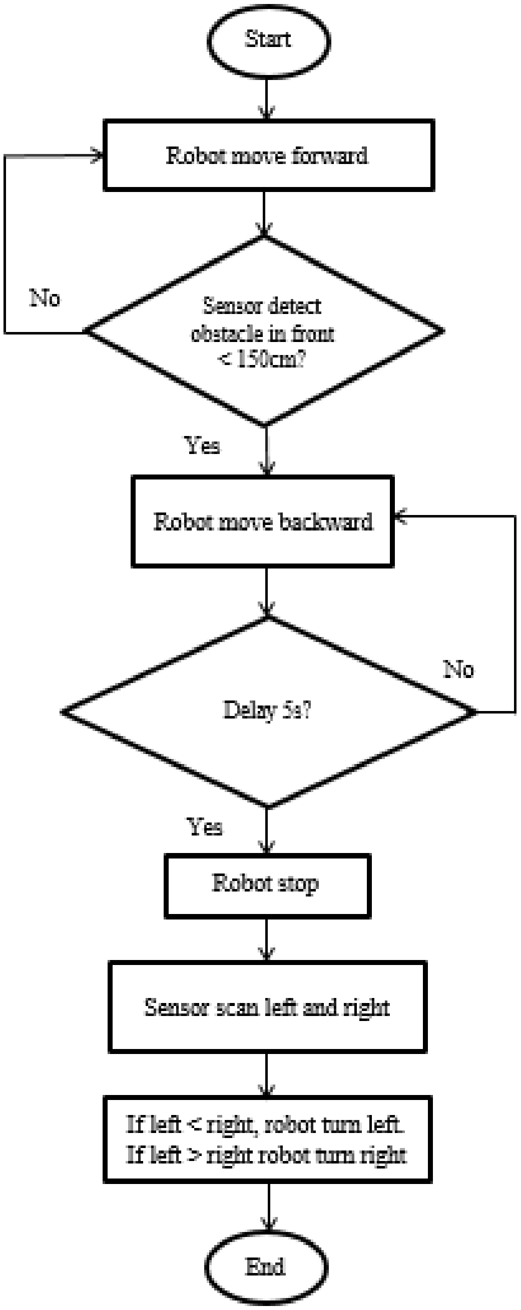

Start
Robot move forward
No
Sensor detect obstacle in front < 150cm?
Yes
Robot move backward
Delay 5s?
No
Yes
Robot stop
Sensor scan left and right
If left < right, robot turn left.
If left > right robot turn right
End

Fig. 9. Navigation System Flowchart.

B. *Spraying System Flowchart*

On the other hand, while the microcontroller executing the condition in the navigation system, the condition for the spraying system also will be considered. As the autonomous pesticide sprayer robot needs to be able to execute both of the operations simultaneously, so the sequence inside of the programming code plays a critical role in the designed project. Fig. 10 shows the process flowchart inside of the spraying system including the.

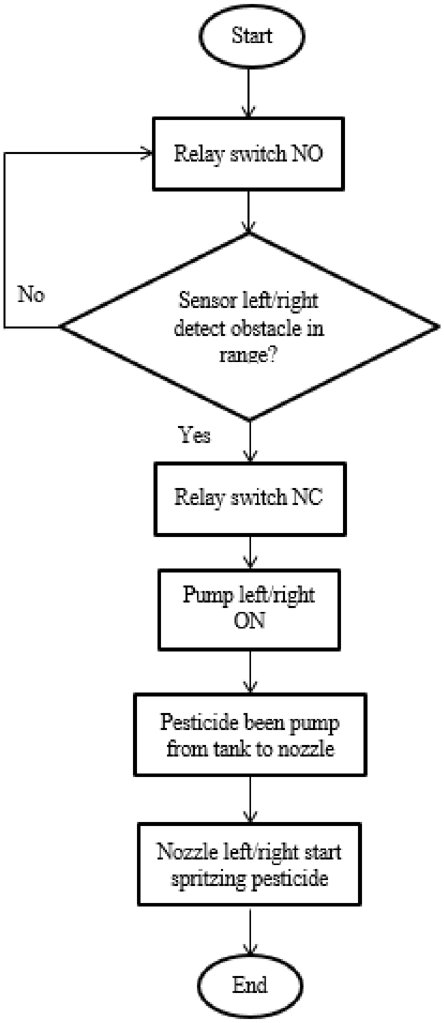

Start
Relay switch NO
No
Sensor left/right detect obstacle in range?
Yes
Relay switch NC
Pump left/right ON
Pesticide been pump from tank to nozzle
Nozzle left/right start spritzing pesticide
End

Fig. 10. Spraying System Flowchart.

V. Experimental Setup

As the autonomous pesticide sprayer robot needs to be tested in the real working environment, the fertigation farm by planting the chili using as the experimental setup environment. The experiment setup in the fertigation farm with approximately 100 chili plants is set. Fig. 11 shows the experimental setup in the chili fertigation farm.

This experiment is performed to know the capability of the sensor to detect the presence of the obstacles in front of it then decide which way it will turn as its next route. The ultrasonic sensor basically emits the soundwave thus received the repelled soundwave as the signal. In this experiment, the sensor distance will be recorded when the autonomous pesticide sprayer robot is moving. The value for front, right and left sensor when there is an obstacle versus the distance took by the robot along the path. The conceptual function for obstacle detection to turn the navigation platform into left direction with ultrasonic affixed 90°, 45° and 135° respectively is shown in Fig. 12.

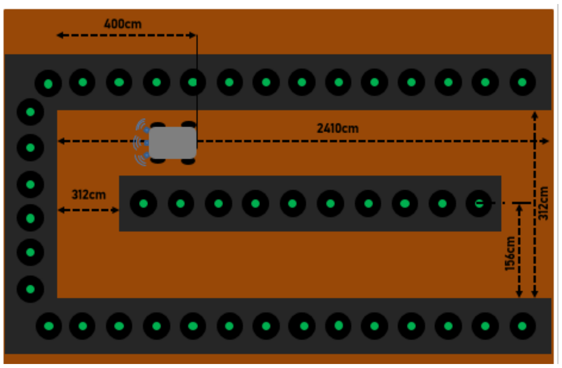


Fig. 11. Experimental Floor Layout in Chili Fertigation Farm.

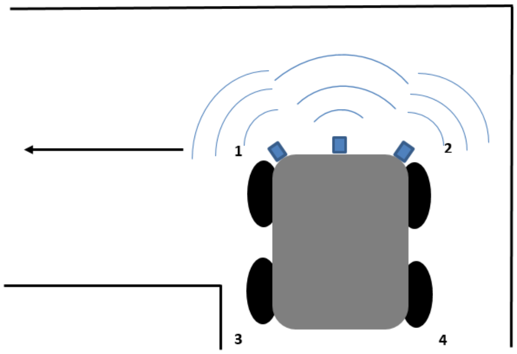


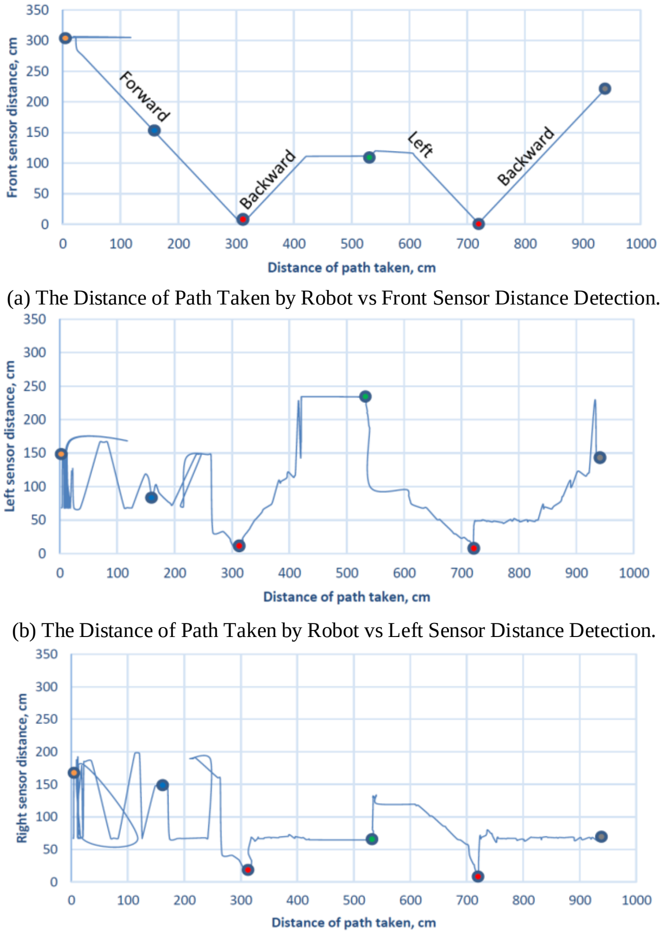
Fig. 12. The Conceptual Function of Sensor Detection in Turning.

VI. Experimental Results

All the measurement data obtained from the ultrasonic sensor on the autonomous pesticide sprayer robot in this experiment are recorded and the moving path of the autonomous pesticide sprayer robot is plotted into the graph as shown in Fig. 13. The graph will be divided into three graphs where represent the front, right and left sensor distance versus the distance of path taken mutually. The experimental method which is applied has been referred from previous works conducted [14]. As shown in Fig. 13. the starting point for the autonomous pesticide sprayer robot started at 3000 cm from the end of the junction. So, the max detection from the front sensor should be below the 300 cm which will become closer as the robot moving forward and farther as the robot moving backward. In Fig. 13, only at point of condition's occurred will be shown as the data was too many to display such as in here it got starting, detection, stopping, turning and ending point in highlighted color which is orange for starting point of the detection, blue for stop detection point, red for actual stopping point, green for start left-turning point and grey for ending point of the detection.

The developed autonomous pesticide sprayer robot was tested to stop at the point of detection by front sensor which is below 150 cm but due to BLDC motor can not instantaneously stop its rotation due to its characteristic of brushless that do not have braking system and also cause by inertia acts upon it, there will be overshoot of autonomous pesticide sprayer robot movement before it was fully stopped at distance 305.03 cm. After fully stop, the robot will take a backward step with delay 5 s until it reaches 227.28 cm from the stopping point. Then, at this point, the value of distance for the left and the right sensor was compared by the given condition in programming code whereas based on which distance was the most farthest detect an obstacle whether left or right sensor. If the left sensor distance was highest means farther from obstacle compared to the right one, the function for turning left will be called out in the programs looping then executed or otherwise.

However, since the autonomous pesticide sprayer robot was tested out to take a left turn in this experiment, the measurement data between left sensor distance is 233.9 cm and right sensor distance, 64.44 cm at this point. Later, the motor will manipulate its direction through gate driver in motor driver to take left turn and basically, the method used to change the heading direction of the autonomous pesticide sprayer robot was based on the combination of motor drive with differential drive when to take left direction, the motor 1 and 3 will turn backward with PWM speed of 80, while motor 2 and 4 will turn forward with PWM speed of 200. Thus, this could allow the autonomous pesticide sprayer robot to have some sort of gliding effect during changing direction. The operation for right-turning can be vice versa to left turning in terms of motor direction turning and its speed. After the left turning with 90° curve, the robot will stop at distance 717.91 cm due to overshoot during turning and then take a backward step once again for 5 s.



(a) The Distance of Path Taken by Robot vs Front Sensor Distance Detection.

(b) The Distance of Path Taken by Robot vs Left Sensor Distance Detection.

(c) The Distance of Path Taken by Robot vs Left Sensor Distance Detection.

Fig. 13. Measurement Data from the Ultrasonic Sensor for Autonomous Pesticide Sprayer Robot.

Based on the experimental results shown, the navigation system for an autonomous pesticide sprayer robot is successfully conducted. The autonomous pesticide sprayer robot could navigate autonomously throughout the experimental field.

VII. Conclusions and Future Tasks

As a conclusion, in order to design and develop an autonomous pesticide spraying for a fertigation farm has successfully conducted. All the subsystem such as navigation systems and spraying systems are included. Although the navigation part has been tested, the autonomous pesticide sprayer robot can be self-navigate by turning at the junction by using the obstacles detection concept inside the fertigation farm. The ultrasonic sensors were used which for front sensor it was adjacently facing forward in 90° while the other two left and right both facing forward with deflection 45°. The ultrasonic sensor could detect the obstacles and stop without hitting the obstacles, respectively.

For future works, the spraying pressure of the autonomous pesticide sprayer robot will be tested and the electronic circuits need a waterproof structure since the autonomous pesticide sprayer robot deals with a pesticide which is fluid. Therefore, the isolation of the electronic component should be done well by separating each electronic component in the container box to prevent it from being damaged if the flooding or leakage happened inside the robot. On the other hand, the pest monitoring system should be developed to be an auto-monitoring device while spraying the pesticide.

Acknowledgment

This project is fully funded by Universiti Teknikal Malaysia Melaka under Short Term Grant Scheme no. PJP/2019/FKE(2C) S01666.

References

[1] FAO, “Global agriculture towards 2050,” FAO. (2009). Glob. Agric. Toward 2050. High Lev. Expert Forum-How to Feed World 2050, 1–4. https//doi.org/http//www.fao.org/fileadmin/templates/wsfs/docs/Issues_papers/HLEF2050_Global_Agriculture.pdfHigh Lev. Expert Forum-How to Feed world, pp. 1–4, 2009.

[2] S. Singh, T. F. Burks, and W. S. Lee, “Autonomous Robotic Vehicle Development for Greenhouse Spraying,” Trans. ASAE, vol. 48, no. 6, pp. 2355–2361, 2013.

[3] P. D. P. R. H. V., “Development of Automated Aerial Pesticide Sprayer,” Int. J. Res. Eng. Technol., vol. 03, no. 04, pp. 856–861, 2015.

[4] Aras, M. S. M., Sulaiman, M., Keong, Y. E., Kasno, M. A., Kassim, A. M., & Khamis, A. (2017). Performance analysis of PID and fuzzy logic controller for unmanned underwater vehicle for depth control. Journal of telecommunication, electronic and computer engineering (JTEC), 9(3-2), 59-63.

[5] Rashid, M.Z.A.; Aras, M.S.M.; Radzak, A.A.; Kassim, A.M.; Jamali, A. Development of hexapod robot with manoeuvrable wheel. Int. J. Adv. Sci. Tech. 2012, 49, 119–136.

[6] M.Z.A Rashid, H.N.M Shah, M.S.M Aras, M.N.Kamaruddin, A.M. Kassim, H.I.Jaaafar,” Metal Line Detection: A New Sensory System For Line Following Mobile Robot”, Journal of Theoretical and Applied Information Technology. ISSN: 1992-8645.

[7] K. Ghaffarzadeh, “Agricultural Robots and Drones 2018-2038: Technologies, Markets and Players,” pp. 1–20, 2018.

[8] G. Zaidner and A. Shapiro, “A novel data fusion algorithm for low-cost localisation and navigation of autonomous vineyard sprayer robots,” Biosyst. Eng., vol. 146, pp. 133–148, 2016.

[9] Anuar bin Mohamed Kassim, Takashi Yasuno, Hazriq Izzuan Jaafar, Mohd Aras Mohd Shahrieel, “ Development and Evaluation of Voice Recognition Input Technology in Navigation System for Blind people”, Journal of Signal Processing, Vol.19, No.4, pp.135-138, July 2015.

[10] M. S. A. Mahmud, M. S. Z. Abidin, and Z. Mohamed, “Localization and motion control implementation for an agricultural mobile robot,” J. Teknol., vol. 79, no. 7, pp. 31–39, 2017.

[11] I. N. Lee, K. H. Lee, J. H. Lee, and K. H. You, “Autonomous greenhouse sprayer navigation using automatic tracking algorithm,” Appl. Eng. Agric., vol. 31, no. 1, pp. 17–21, 2015.

[12] A.M. Kassim, H. I. Jaafar, M.A. Azam, N. Abas, T.Yasuno, "Design and Development of Navigation System by using RFID Technology ", 3rd IEEE International Conference on System Engineering and Technology (ICSET), 2013, pp. 258–262.

[13] A. M. Kassim, T. Yasuno, H. Suzuki, H. I. Jaafar, M. S. M. Aras, "Indoor navigation system based on passive RFID transponder with digital compass for visually impaired people", Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 2, 2016.

[14] M.R Yaacob, N.S.N Anwar, A.M Kassim ,“ Effect of Glittering and Reflective Objects of Different Colors to the Output Voltage-Distance Characteristics of Sharp GP2D120 IR”ACEEE International Journal on Electrical and Power Engineering. 3 (2). pp. 6–10, 2012.

[15] A.M. Kassim, H. I. Jaafar, M.A. Azam, N. Abas, T.Yasuno, " Performances study of distance measurement sensor with different object materials and properties", 3rd IEEE International Conference on System Engineering and Technology (ICSET), 2013, pp. 281–284.

[16] Anuar bin Mohamed Kassim, Takashi Yasuno ; Hazriq Izzuan Jaafar ; Mohd Shahrieel Mohd Aras ; Norafizah Abas, Performance analysis of wireless warning device for upper body level of deaf-blind person, 2015 54th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), (2015), pp. 252–257.

[17] Anuar Mohamed Kassim, Takashi Yasuno, Hiroshi Suzuki, Mohd Shahrieel Mohd Aras, Hazriq Izzuan Jaafar, Fairul Azni Jafar, Sivarao Subramonian, "Conceptual design and implementation of electronic spectacle based obstacle detection for visually impaired persons", Journal of Advanced Mechanical Design, Systems, and Manufacturing, vol. 10, pp. JAMDSM0094, 2016.

[18] Anuar , Mohamed Kassim and Shukor, Ahmad Zaki Performance Study of Developed SMART EYE for Visually Impaired Person. Australian Journal of Basic and Applied Sciences. pp. 633-639, 2013.

[19] Anuar Bin Mohamed Kassim, Takashi Yasuno, Hiroshi Suzuki, Mohd Shahrieel Mohd Aras, Ahmad Zaki Shukor, Hazriq Izzuan Jaafar and Fairul Azni Jafar (November 5th 2018). Vision-Based Tactile Paving Detection Method in Navigation Systems for Visually Impaired Persons, Advances in Human and Machine Navigation Systems, Rastislav Róka, IntechOpen.

[20] R. Oberti et al., “Selective spraying of grapevine’s diseases by a modular agricultural robot,” J. Agric. Eng., vol. 44, no. 2s, pp. 149–153, 2016.

[21] C. Xia, Y. Li, T. S. Chon, and J. M. Lee, “A stereo vision based method for autonomous spray of pesticides to plant leaves,” IEEE Int. Symp. Ind. Electron., no. ISlE, pp. 909–914, 2009.

[22] B. S. Faiçal et al., “An adaptive approach for UAV-based pesticide spraying in dynamic environments,” Comput. Electron. Agric., vol. 138, pp. 210–223, 2017.

Task Sensitivity in Continuous Electroencephalogram Person Authentication

Rui-Zhen Wong[1], Yun-Huoy Choo[2*], Azah Kamilah Muda[3]
Computational Intelligence and Technologies (CIT) Research Group
Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM)
76100 Durian Tunggal, Melaka, Malaysia

***Abstract*—This research investigates on the task sensitivity in multimodal stimulation task for continuous person authentication using the electroencephalogram (EEG) signals. Pattern analysis aims to train from historical examples for prediction on the unseen data. However, data trials in EEG stimulation consists of inseparable cognitive information that is difficult to ensure that the testing trials contain the cognitive information matching to the training data. Since the EEG signals are unique across individuals, we assume that multimodal stimulation task in EEG analysis is not sensitive in train-test data trials control. Data trial inconsistency during training and testing can still be used as biometrics to authenticate a person. The EEG signals were collected using the 10-20 systems from 20 healthy subjects. During data acquisition, subjects were asked to operate a computer and perform various computer-related tasks (e.g.: mouse click, mouse scrolling, keyboard typing, browsing, reading, video watching, music listening, playing computer games, and etc.) as their preferences, without interruption. Features extracted from Welch's estimated Power Spectral Density in different frequency bands were tested. The designed authentication approach computed intra- and inter-personal variability using Mahalanobis distance to authenticate subject. The proposed EEG continuous authentication approach has succeeded. Data collected from multimodal stimulus disregard of task sensitivity able to authenticate subject, where the highest verification performance shown in the low-Beta frequency band. Evidence found that effective frequency region on the middle band was anticipated due to the data collected was based on subject voluntary actions. Future research will focus on the effect of subject voluntary and involuntary actions on the effective frequency region.**

***Keywords*—Electroencephalogram; continuous authentication; task sensitivity; multimodal stimuli; Mahalanobis distance**

I. Introduction

The conventional biometrics use human physiological traits such as fingerprint, iris, face, etc. for authentication. However, one of the limitations of these physiological traits is easily prone to forgery. This is due to the involvement of human exposed body parts that are easy to obtain and replicate. Besides, the system that commonly available in the market is prone to security mistakes such as invasion by hackers or misconduct of authorized personnel. Therefore, alternative biometrics is required due to the needs of increasing level of the information technology security of a system. The human brainwaves that measure in Electroencephalogram (EEG) has proven to fulfill all the biometric requirements (universality, uniqueness, constancy, collectability) [1]. Besides, it has proven to be unique across individuals and is allowed for person authentication [2]. One of the characteristics of EEG is non-stationary or quasi-stationary over time, where this rises the problem of template permanence that often discussed by the research community [3]. However, such characteristics do make possible for spoof resistance, liveness detection, cancellability [4] and etc. which is beneficial for person authentication.

Like most of the security implementation, the common brainwaves based biometric authentication is allowed for one-time authentication only. The major drawback of static authentication (SA) is the system will unaware of the user anymore once the access has been granted to the client. Thus, this provides chances to an intruder for a spoofing attack. In such cases, continuous authentication (CA) is believed to provide security awareness against imposters. CA involves repetition verification process along the time while the user is still logged on to the system. Users whoever using the system will be monitored in the complete session and ensure the authority only given to the correct person. Since CA requires repetition obtaining the data, thus it is more practical if enough user behavioral data can be acquired passively without the consciousness of the user to improve practicality [5]. The excellent time resolution with the continuous nature of EEG signals enables precise detection of brain activity. This allows more possibility for continuous person authentication in real-time which able to detect an imposter and response in just seconds.

EEG signals analysis is proven for person authentication because the cognitive response is individuating in different persons when response towards similar cognitive tasks (e.g.: resting, visual stimulus, mental imaginary, and etc.). Although experiments involved multimodal stimulus that fulfilled unconscious set-up has introduced in several studies to improve the practicality for brainwave continuous authentication [6], [7]. However, these experiments involve only a consistent EEG stimulus set throughout the whole recording session, where the same set of EEG cognitive tasks is possible to occur in every data trial for both data training and testing sets. Thus, a non-specified set of multimodal stimuli of EEG recording is expected to improve the robustness of the experiment for person authentication. This is because the cognitive tasks are indistinguishable and different between data trials during the segmentation of EEG signals. Also, this makes possible for testing trials to contain EEG tasks that may not be experienced in the training set.

*Corresponding Author.

Thus, this paper aims to propose a flexible EEG recording approach that able to cater to the task sensitivity for multimodal stimulus. Besides, the proposed approach should enable unconscious data collection which suitable to be used for continuous authentication. Therefore, EEG signals will be acquired from the user performing random tasks by themselves without limitations. We hypothesized that the data trials consisting of inseparable cognitive information that possibly mismatch during training and testing can still be used as biometrics to authenticate a person. Questions that addressed in this work: (1) Would the multimodal EEG stimulus able to authenticate person disregard of task sensitivity? (2) What is the effective frequency region in Power Spectral Density (PSD) for the multimodal EEG stimulus disregard of task sensitivity?

This paper is structured as follows: Section II describes the overview of EEG based biometrics including EEG characteristics, the process flow of EEG authentication, EEG protocols, and the related works. Section III presents the proposed solution from the experiment paradigm design until the performance evaluation. Section IV presents the results and discussion where Section V draws conclusions and suggests the direction for future work.

II. Overview of EEG based Biometrics

A. EEG Characteristics

EEG measured the spontaneous electrical changes inside the brain that can be obtained by placing sensors along the human scalp. The acquired raw brainwaves often plotted in the amplitude-time graph that explains the voltage fluctuation over a period of time, in a specified brain region. Generally, the informative brain activities lie on several frequency bands that categorized as follow:

1) Delta, δ waves (0.5-4Hz). Waves with the highest amplitude and slowest activity appear in the deep sleep and unconscious state.

2) Theta, θ waves (4-8Hz). Slow activity that appears in the deepest state of meditation.

3) Alpha, α waves (8-13Hz). Appears during relaxation or dreaming and disappear while human during thinking and alert state.

4) Beta, β waves (13-30Hz). Low amplitude waves and appear in a waking state with high attention.

5) Gamma, γ waves (>30Hz). Appears when information processing, decision making, or multimodal sensory processing.

B. Process Flow of EEG Authentication

The brain biometrics has two types of applications: authentication or identification [8]. The decision mechanism for authentication (or is often called verification) involved one to one matching only where the results will be either accept or reject the user. However, identification comparing one-to-many options in the databases where an identified label of the subject will be the output. However, several steps must be fulfilled to complete brainwaves recognition as depicted in Fig. 1: EEG signals collection, signal pre-processing, feature extractions, template matching/classification and classified output.

C. EEG Protocols

Typically, the EEG signal acquisition protocol can be grouped into three (3) categories based on the review from [8]–[10], which are: resting-state/relaxation, event stimulation, and mental imaginary. The resting-state protocol requires subject to sit with relaxing in eye open or eye closing condition, EEG signal will be acquired during this human quiescent state. In event stimulation protocol, cognitive stimulus in different forms will be presented to the subject (e.g.: visual, audio, somatosensory, etc.) because the evoked potential that triggered from the presented stimulus able to differentiate individual. In the mental imaginary protocol, brain signals will be captured while the subject was performing a certain mental task (e.g.: imagining hand movement, rotation, solving arithmetic problem, etc.). Overall, the EEG of a person is recorded from their non-volitional or volitional responses in an engaging session. This time frame will be selected and to be used for further analysis to authenticate individuals. However, several problems would like to address in the CA point of view based on the existing EEG collection protocol.

First, although the relaxation protocol able to acquire prolonged continuous EEG signals, however, it is impossible to expect people will be kept resting all the time. In real life, the human physical and mental state will tend to be active, where uncertainty may rise due to the occurrence of unknown experience that unable to measure in resting EEG data. Second, cognitive recording only involved a single task. To the extent of this, most of the EEG authentication scheme is based on single task training and testing, in which the template is generated from the single and distinguished type of brain task and later to be tested using the same brain task (e.g.: training and verification using left-hand motor imaginary task only), but in real life, we cannot expect human mental activity will always in a regular state. Whereas, multiple task studies have received attention later where EEG recordings from different combinations of brain tasks were used for training and testing (verification). The different design of the EEG experiment was as shown in Table I.

Studies found that the fusion of different EEG tasks in training/testing able to provide significant outcomes as compared to when evaluated individually [11], where the extensive review of multi-task study for EEG subject identification can be found in [12]. As for subject authentication, a study in [13] has conducted several. experiments to evaluate the performance using one type of task for training and tested with another task. Results show the performance remains when mismatch between training and testing tasks compare to using the same task. Also, system performance does improve if the training data involved more tasks in training and tested with another task. Thus, this gains confidence in flexibility for the design of the EEG data collection protocol. However, the above claims only applicable to mismatch training/testing between motor or imaginary tasks only. The author also tried to include resting tasks in the test set, but the performance obtained was very poor, where this highlights the first problem that we have addressed previously in this section.

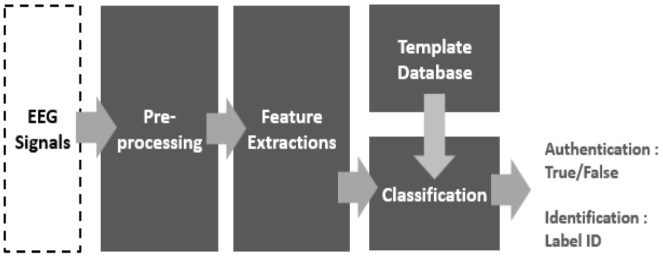


Fig. 1. EEG Authentication Process Flow.

TABLE. I. EEG EXPERIMENT

Stimulus Mode	EEG Experiment	Training Set	Testing Set	Past Research
Unimodal	Single Task	Task A	Task A	[11], [14]
		Task B	Task B	
	Multiple Task	Task A	Task A or Task B	[13]
			Other Tasks	[13], [15]
		Task A and Task B	Task A or Task B	[13], [16], [17]
			Task C	[13]
			Task A and Task B	[16], [11]
Multimodal		Task A+B	Task A+B	[6], [7]
		Task A+B	Task A, Task B, Other tasks, Various combination of Task A+B+Other tasks	(Proposed Work)

Next, the third problem to be addressed was, the conscious response of the user is required during data collection. The event stimulation and mental imaginary protocol. For example, in the visual stimulation protocol, images are presented to the subject to register the Event-Related Potential (ERP) as their template, where the relevant image needs to be presented again during verification. However, it is less suitable to let users aware of CA by kept displaying images to them, where practical CA should allow passive verification as mentioned in [5].

D. Related Works

The common EEG experiments record the user's brain wave through perceiving unimodal stimulus only, in which the single mode of EEG stimulus was presented at a time [18]. However, multiple sensory cognitive processing is more often to happen in a real-world scenario. Meanwhile, the brainwaves through EEG authentication can also be recorded without any controlled stimulation to the user to obtain continuous signals. Attentive tasks such as driving and computer operating involve multimodal stimulus where humans will expose to more than one type of stimulus from different sensory fields (e.g.: visual, auditory, spatial, tactile, and etc.) simultaneously. Study in [6] records continuous EEG from only Fp1 electrode in the simulated driving environment and achieved the best of 27% EER, the recording lasts for three (3) minutes per trial and collected twice a day for five (5) days from thirty (30) subjects. Apart from EEG, a study in [7] records continuous brainwaves in near-infrared spectroscopy (NIRS) for 60s while the user was doing typing tasks. Only a single probe placed on the subject forehead was used to minimal interruption, where this study able to obtain 0.40% EER.

III. PROPOSED SOLUTION

A. Experiment Paradigm Design and Data Collection

A total of 20 students in Universiti Teknikal Malaysia Melaka (UTeM) comprised 10 males and 10 females aged between 20 and 29 (mean age: 23.78 ± 1.93 standard deviations) has participated voluntarily in this experiment. All of them were healthy adults, right-handers, and had normal or corrected to normal vision. Procedures were approved by the Ministry of Health Malaysia under the National Medical Research Register (NMRR-19-2372-50333). Participants signed a printed consent form after being briefed on the overall purpose of the research study and the experimental procedure before participating.

Fig. 3(a) illustrated the experimental set-up for the proposed EEG recording approach, where the arrangement in the actual scenario is as shown in Fig. 3(b). The subject was first asked to wear a wireless EEG head cap and sit in front of a computer with a screen size of 15.6 inches and distance approximately 45 centimeters. The brainwave signals were measured in EEG with 20 dry electrodes sampled at 500Hz frequency. All channels positioned following 10-20 international placement systems which include P7, P4, Cz, Pz, P3, P8, O1, O2, T8, F8, C4, F4, Fp2, Fz, C3, F3, Fp1, T7, F7, and Oz. A reference electrode was placed on the left or right of the subject earlobe, A1 or A2 as illustrated in Fig. 2. The EEG device used was Neuroelectrics Enobio 20 which is a wireless and portable headset that transmits data via Bluetooth. Distance between the headset and the Bluetooth dongle was approximately 100 centimeters.

To authenticate users unconsciously, the experiment design should allow transparent monitoring. For EEG data collection, the subject was asked to operate the computer and perform any computer tasks as their preferences. Examples of computer tasks include mouse scrolling, mouse-clicking, keyboard typing, browsing (reading), video watching, music listening, playing computer games, and any other computer-related tasks. To ensure practicality, two (2) conditions were allowed. First, no restriction on the number of types of computer tasks per recording. Subjects were free to perform any computer tasks at any time as their preferences. Second, no restriction on the number of computer tasks at one-time. Subjects were free to perform different computer tasks concurrently (e.g.: listen to music while reading). However, the subject is informed to minimize their body movement such as avoid excessive hand, head, body, and face movement to reduce the captured of noise signals in the recording.

While the device was placed on the subject scalp, the obtained EEG signals were monitored in the complementary NIC v1.4 software. Color indicators were provided for every electrode to observe signal quality is good (green), moderate (orange), and bad (red) conditions. The indicator is not an impedance check but is guidance checking for line noise level, main noise level, electrode drift, and offset [19]. Thus, it is not necessary to stop the data collection if the indicator turns red. However, to reduce the capture of noise, EEG signals collection begins while the indicator for all electrodes appears green and prolonged for at least five (5) seconds. The experiment runs for one (1) time only for each subject and the total duration for tasked recording lasted for ten (10) minutes. The experiment was done in a quiet and enclosed room dedicated to the EEG experiment.

B. *Signal Pre-Processing*

In his study, open-source API, MNE v0.17.1 is used for data preparation and analysis [20], [21]. This tool is widely used for EEG or MEG data analysis in Python. The block diagram of the proposed EEG-based biometric system used for continuous authentication is depicted in Fig. 4, where all the processes will be discussed in this section hereafter.

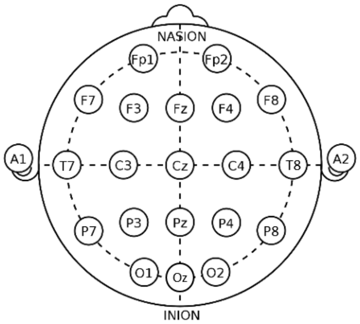
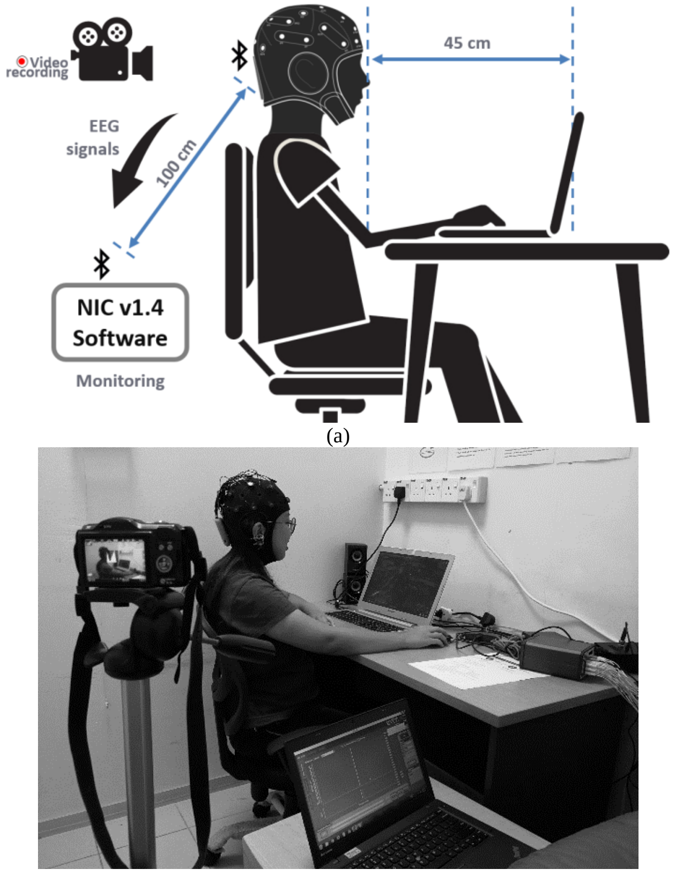


Fig. 2. The International 10-20 Electrode Placement with 20 Channels and 1 Reference Electrode (A1 or A2).



(a)

(b)

Fig. 3. Set-up for EEG Experiment of Proposed Method (a) as Illustrated (b) in Actual Scenario.

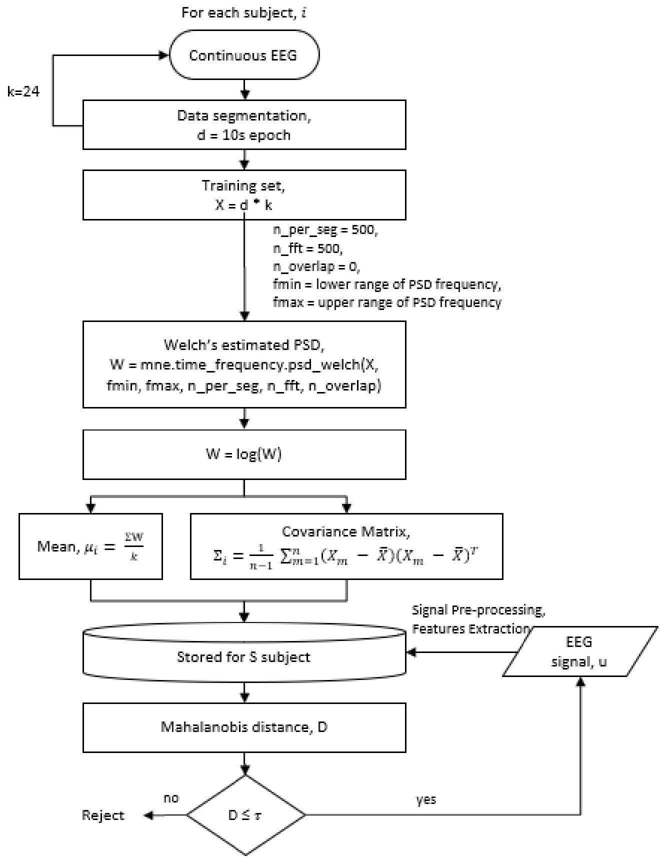


Fig. 4. EEG-based Continuous Authentication Process.

Each of the S=20 subjects in this experiment was given a numerical label to differentiate between subjects. To analyze only quality signals, the first one (1) minute of data were removed and only eight (8) minutes of data in the middle of ten (10) minutes recording were used. Thus, the time-series EEG signals from one electrode contribute to 240,000 data points (8 mins * 60s * 500Hz). Next, the data was segmented equally in 10s epoch without overlapping [7], for each subject and each electrode. Therefore, each user possesses a total of 2k=48 data trials with respective subject labeled, where all electrodes are concatenated in dimension. Data were split to a portion of 50:50, all the trials were still periodically arranged without shuffling, where the first half (k=24 trials) used for template generation (training) and another half for performance verification (testing).

C. *Features Extraction*

Features in frequency domain able to extract dominant brain activity in the specified frequency range. Power Spectral Density (PSD) that measured the signal power in relative frequency band was extensively used to extract features in EEG biometrics analysis. PSD provides fast computation and suitable to process continuous EEG data from simple sources which possibly contains more artifacts [22], this is suitable because the dry electrode used in the experiment has lower signal quality as compared to the wet or gel-based electrode.

Thus, we employ PSD as the feature extraction method and considered Delta δ (0.5-4 Hz), Theta θ (4-8 Hz), Alpha α (8-13Hz), low-Beta β (13-20 Hz), high-Beta (20-30 Hz) to Gamma γ (30-50 Hz) band. Although there is still no agreement on standard reference for the specific value of each frequency band should be, however, the cut-off between alpha and beta at 13 Hz is based on our preliminary experiment, and the gamma band to stop at 50Hz is based on [15]. The selection of the mentioned frequency band is because this range consisting of dominant brain activities that able to recognize person disregard of brain tasks [15], [16]. The frequency band will later be tested in a combined and separated manner to identify the effective region for better efficiency [14].

Thus, the power spectra of the processed EEG signals in each trial, each electrode, and each subject were transformed using Welch's estimation method, using 500 Fast Fourier Transform (FFT) length (this number is set based on the EEG device sampling rate and resulting to 1 resolution point for the power spectral frequency bin) with no overlapping information. Next, the logarithm of power spectra was computed.

D. Authentication Approach

Mahalanobis distance, introduced in [23], is a simple multivariate metric that measures the distance between a point to a distribution. It has been proven efficient in [7] for brain signal continuous authentication. The equation of Mahalanobis Distance is as follows:

$$D = \sqrt{(u-\mu_i)^{-1}{\Sigma_i}^{-1}(u-\mu_i)} \qquad (1)$$

Where D is the Mahalanobis distance, u is the vector of observations (test data trial), μ and Σ^{-1} is the vector of mean and the vector of the inverse covariance matrix of the claimed subject, i respectively. For each registered subject i, the mean vector μ and inverse covariance matrix Σ^{-1} were first computed using the training set. This information will be stored in the memory and to be used to check the distance with the testing trials. It is important to note that the covariance matrix here must be a positive definite matrix because the square root can only take a positive value of the inner product. To authenticate person, value D of the test data, j to the cluster of i-th subjects were computed. This procedure was iterated for every testing trial and every subject. The calculated distance, D_{ij} indicating individual variability that can be explained by intra-individual distance (when $i = j$) and inter-individual distance (when $i \neq j$). A person will be authenticated if $D \leq \tau$, where τ is a pre-specified threshold.

E. Performance Evaluation

To access the performance of the proposed EEG authentication scheme, we employed several widely used evaluation metrics such as Equal Error Rate (EER), False Acceptance Rate (FAR), and False Rejection Rate (FRR). Since authentication is a binary class problem (e.g.: true/false, or accept/reject), it will produce two (2) types of error which are: FAR when an imposter being accepted (false class classified as true); and FRR when a client is being rejected (true class classified as false). However, EER is a point when (FAR = FRR) in a threshold frequency distribution graph, where the value falls under the EER often taken as an optimal point for decision threshold to reject an imposter.

IV. Results and Discussion

As a result, the calculated intra- and inter-individual distances were denoted by blue and red indicators respectively as shown in Fig. 5. The blue cluster is the collection of intra-individual (self-to-self) distances that need to be treated as a client, as opposed to the red cluster which consisting inter-individual (self-to-others) distances that represent imposter. A sliding threshold on the horizontal axis in Fig. 5 able to obtain the FAR and FRR. The combination results in the respective threshold produced the curve as illustrated in Fig. 6, where EER is the intersection point of two (2) curves shown in the graph.

Fig. 7 shows the ERR results that tested in different frequency band specification. Ten (10) different combinations of frequency regions were tested in a separated and combined manner to identify the effective band. The results reveal the band selection based on multimodal EEG stimulus task sensitivity. The lower the EER value indicates better authentication performance due to lower false classification rate. From the results, it is quite appealing that the proposed CA approach is effective. Overall, each frequency band specification has a different verification performance. The combined frequency region from the literature (α+low-β [6], δ+θ+α+β [16], θ+α+β+γ [15]) able to authenticate person, but results show there is separated region which able to authenticate individuals more effectively.

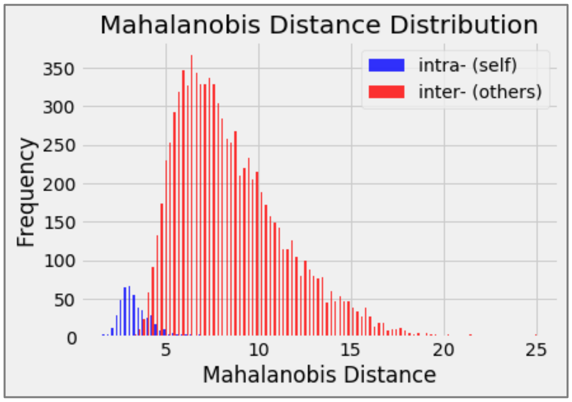


Fig. 5. Mahalanobis Distance Distribution in Low-Beta.

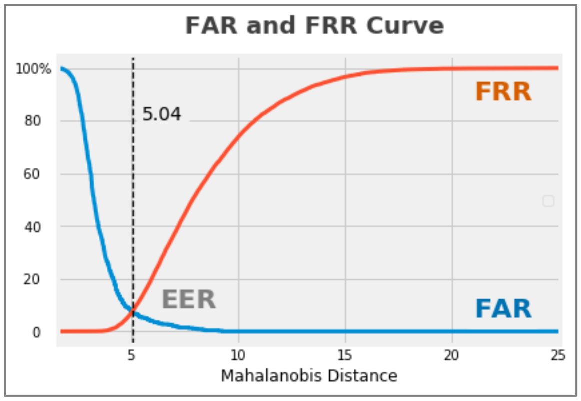


Fig. 6. FAR and FRR Curve in Low-Beta Band.

The authentication performance is good in the order of low-β, high-β, β, a + low-β, α, γ, δ+θ+α+β, θ+α+β+γ, θ, and δ frequency band. Generally, features in β region able to provide good verification results. Specifically, it is clear from Fig. 7 that low-β frequency achieving best verification performance which able to authenticate subject well for the random multimodal EEG tasks disregard of task sensitivity. This results in similar to [6] where the simulated driving scenario was the EEG task. The study compares results between α, low-β, and α+low-β, where features in low-β band are providing the best authentication performance. In this study, the best verification performance achieved was is as shown in Fig. 6, which can be formulated by:

$$EER = FAR(5.04) = FRR(5.04) = 7.29\% \quad (2)$$

Besides, when we look only into the separated frequency sub-band (δ, θ, α, β, and γ), the verification results getting better from lower to higher frequency region except for γ. The rising trend can be associated with the brain activity in relevant frequency regions as discussed in Section II.A, where the informative frequency band will be in the higher region as the human mental state changes from deepest relaxation to highly attentive. The computer operating task involved in this study requires human attention, thinking, decision making, cognitive response, and simple motor movement. Thus, we expect the γ band can provide a result in a higher rank, but evidence shows its performance ranked after β and α band.

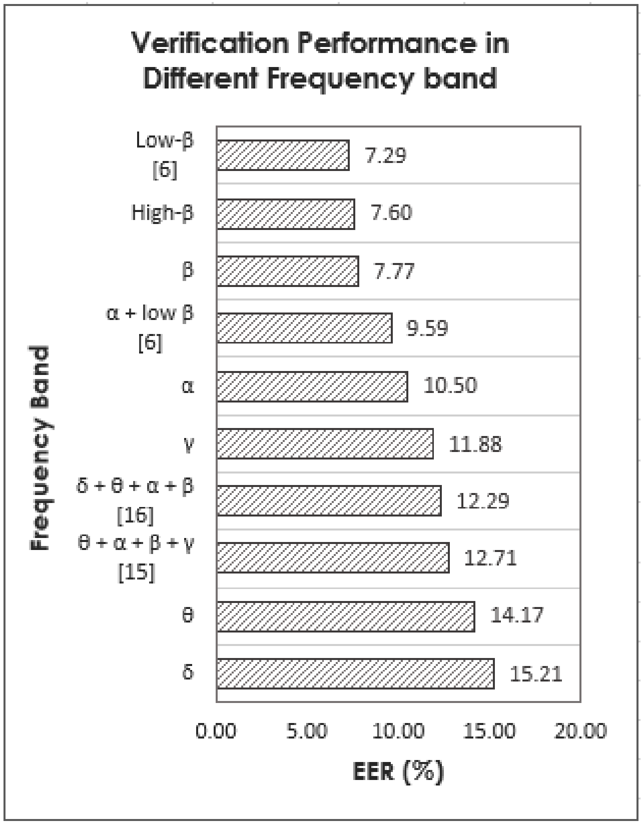


Fig. 7. Verification Performance in different Frequency Band.

However, this is anticipated due to the tasks that performed during data collection is based on user preferences (voluntary action), thus they are comfortable and relax while engaging in the experiment but not in a highly stressed and unknown situation that will pay higher attention to perform the EEG tasks. Another reason where higher frequency bands able to authenticate subject better is because the EEG multimodal stimuli involved attentive tasks. Thus, frequency in the lower band (δ and θ) will not give better results as compared to the higher region (α, β, and γ).

V. Conclusion

This study embarked on the motivation to propose a flexible EEG recording approach for continuous person authentication that able to cater to the task sensitivity for multimodal stimuli. During the data collection experiment, EEG signals are recorded while subject operating a computer and performance random computer tasks such as mouse scrolling, mouse-clicking, keyboard typing, browsing (reading), video watching, music listening, playing computer games, and any other computer-related tasks, based on user preferences. The obtained continuous EEG signals containing inseparable and mismatch cognitive tasks in data trials during training and testing able to authenticate person successfully. We determine to suggest that the low-β band has better separation ability as compared to other frequency band specifications which able to achieve the lowest EER of 7.29%, no matter the task sensitivity for multimodal EEG tasks.

Based on the results, frequency especially located in the middle region was more effective as compared to the lower and higher region. This is because the multimodal stimulus task requires subject attention. Besides, such evidence also anticipated due to the tasks performed during data collection were based on subject voluntary actions, less stress and more pleasant incur the effective frequency region lies in the middle part but not the higher region. Thus, future research may investigate the effect of subject voluntary and involuntary actions on the effective frequency region.

Acknowledgment

The research work was supported by the Fundamental Research Grant Scheme FRGS/1/2017/ICT05/FTMK-CACT/F00346 from the Ministry of Higher Education and University Teknikal Malaysia Melaka (UTeM).

References

[1] A. K. Jain, A. Ross, and S. Prabhakar, “An Introduction to Biometric Recognition,” in IEEE Transactions On Circuits and Systems for Video Technology, 2004, vol. 14, no. 1, pp. 4–20.

[2] S. Marcel and J. del R. Millan, “Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 4, pp. 743–748, 2007.

[3] M. V. Ruiz-Blondet, Z. Jin, and S. Laszlo, “Permanence of the CEREBRE brain biometric protocol,” Pattern Recognit. Lett., vol. 95, pp. 37–43, 2017.

[4] F. Lin, K. W. Cho, C. Song, W. Xu, and Z. Jin, “Brain Password : A Secure and Truly Cancelable Brain Biometrics for Smart Headwear,” ACM Int. Conf. Mob. Syst. Appl. Serv., pp. 296–309, 2018.

[5] I. Traore and A. A. E. Ahmed, Continuous Authentication Using Biometrics: Data, Models, and Metrics. IGI Global, 2012.

[6] I. Nakanishi, H. Fukuda, and S. Li, “Biometric Verification Using Brain Waves Toward On-Demand User Management Systems,” in Proceedings of the 6th International Conference on Security of Information and Networks, 2013, pp. 131–135.

[7] Y. Matsuyama, M. Shozawa, and R. Yokote, “Brain signal’s low-frequency fits the continuous authentication,” Neurocomputing, vol. 164, pp. 137–143, 2015.

[8] M. Abo-Zahhad, S. M. Ahmed, and S. N. Abbas, “State-of-the-art methods and future perspectives for personal recognition based on electroencephalogram signals,” IET Biometrics, vol. 4, no. 3, pp. 179–190, 2015.

[9] Q. Gui, M. V. Ruiz-Blondet, S. Laszlo, and Z. Jin, “A Survey on Brain Biometrics,” ACM Comput. Surv., vol. 51, no. 6, pp. 112:1-112:38, 2019.

[10] H. Yap, Y. Choo, and W. Khoh, “Overview of Acquisition Protocol in EEG Based Recognition System,” in International Conference on Brain Informatics, 2017, pp. 129–138.

[11] Y. Ishikawa, C. Yoshida, M. Takata, and K. Joe, “Validation of EEG Personal Authentication with Multi-channels and Multi-tasks,” in Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), 2014, p. 1.

[12] M. Del Pozo-Banos, J. B. Alonso, J. R. Ticay-Rivas, and C. M. Travieso, “Electroencephalogram subject identification: A review,” Expert Syst. Appl., vol. 41, no. 15, pp. 6537–6554, 2014.

[13] S. Yang, F. Deravi, and S. Hoque, “Task sensitivity in EEG biometric recognition,” Pattern Anal. Appl., vol. 21, no. 1, pp. 105–117, 2016.

[14] C. Han et al., “Contrast between Spectral and Connectivity Features for Electroencephalography based Authentication,” in World Congress on Medical Physics and Biomedical Engineering, 2015, vol. June 7-12, pp. 1224–1227.

[15] S. Altahat, M. Wagner, and E. M. Marroquin, “Robust Electroencephalogram Channel Set for Person Authentication,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 997–1001.

[16] O. Attallah, “Multi-tasks Biometric System for Personal Identification,” in 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2019, pp. 110–114.

[17] Shiliang Sun, “Multitask learning for EEG-based biometrics,” 2008 19th Int. Conf. Pattern Recognit., pp. 1–4, 2008.

[18] E. C. Hames et al., “Visual, auditory, and cross modal sensory processing in adults with autism: An EEG power and bold fMRI investigation,” Front. Hum. Neurosci., vol. 10, no. 167, pp. 1–18, 2016.

[19] Neuroelectrics, Neuroelectrics User Manual: P3.NIC, vol. 2.0. 2016.

[20] A. Gramfort et al., “MEG and EEG data analysis with MNE-Python,” Front. Neurosci., vol. 7, pp. 1–13, 2013.

[21] A. Gramfort et al., “MNE software for processing MEG and EEG data,” Neuroimage, vol. 86, pp. 446–460, 2014.

[22] S. Yang and F. Deravi, “On the Usability of Electroencephalographic Signals for Biometric Recognition : A Survey,” IEEE Trans. Human-Machine Syst., vol. 47, no. 6, pp. 1–12, 2017.

[23] P. C. Mahalanobis, “On the generilised distance in statistics,” Proc. Natl. Inst. Sci. India, vol. 2, no. 1, pp. 49–55, 1936.

Exploiting White Spaces for Karachi through Artificial Intelligence: Comparison of NARX and Cascade Feed Forward Back Propagation

Shabbar Naqvi[1], Minaal Ali[2], Aamir Zeb Shaikh[3], Yamna Iqbal[4], Abdul Rahim[5], Saima Khadim[6], Talat Altaf[7]
Department of Computer Systems Engineering, Balochistan University of Engineering and Technology Khuzdar, Pakistan[1]
Department of Electronic Engineering, NED University of Engineering and Technology Karachi
Pakistan[2, 3, 4, 5]
Department of Telecommunications Engg, Dawood University of Engineering and Technology, Karachi. Pakistan[6]
Department of Electrical Engineering, Sir Syed University of Engineering and Technology Karachi, Pakistan[7]

Abstract—**Marriage of Internet of Everything (IoE) and Cognitive Radio driven technologies seems near under the umbrella of 6G and 6G+ communication standard. The expected new services that will be introduced in 6G communication will require high data rates for transmission. The learning based algorithms will play a key role towards successful implementation of these novel technologies and evolving next generation wireless standards for providing ubiquitous connectivity. This paper investigates performance of two artificial neural network (ANN) based algorithms for Karachi. These include Nonlinear autoregressive exogenous Algorithm (NARX) and cascade feed forward back propagation neural network (CFFBNN) scheme. A dataset for Karachi is also developed for 1805 MHZ. The results of the two algorithms are compared that show Mean Square Error (MSE) for CFFBNN is 6.8877e-5 at epoch 16 and MSE for NARX is 3.1506e-11 at epoch 26. Hence, exploiting computational performance, NARX performs much superior than the classis CFFBNN algorithm.**

Keywords—6G; cognitive radio; NARX; cascaded feed forward neural network; learning

I. Introduction

Internet of Things (IoT) typically refers to interconnection of various networks. The basic network may include some or all of the sensing entities such as kitchens, personal computers etc. into sensing entities [1], [2]. Internet of Everything (IoE) refers to the concept of people to machine connection, people to data and data to things. Hence, it is expected to require huge amount of bandwidth as the data requirement is towards higher side. After the roll out of 5G communication standard, researchers are exploring and investigating different tools and technologies so that the requirements could be fulfilled for next generation wireless standard i.e. 6G. Hence, many applications are envisioned such as wireless brain computer interactions, autonomous systems and connected robotics, block chain and distributed ledger technologies, multisensory RF applications [3], big data for 6G, AI enabled closed loop and intelligent wireless communications [4]. These services are expected to be provided through technologies such as above 6G for 6G, transceivers with integrated frequency bands, Edge Artificial Intelligence, Integrated terrestrial airborne and satellite networks, energy transfer and harvesting and beyond 6G technologies [3]. Additionally, a paradigm shift from software to network intelligence is also expected into the next generation of wireless networks [4]. The evolution towards intelligent connections will result in data rates of up to 1 Tb/s, highly energy efficient with a battery-free IoT device option, massively low latency control i.e. less than 1 msec i.e. end- to end latency, broad frequency bands 73GHz-140 GHz and 1-3 THz, ubiquitous connectivity through integration of global cellular and satellite systems and connected intelligence with machine learning capability[4].

Artificial intelligence algorithms can be distributed between two set of schemes i.e. machine learning and deep learning techniques[5], [6]. These algorithms are computationally efficient algorithms that observe the action rather than computing complex equations. So, based on the observations these algorithms predict the future results. As the wireless communication scenarios are mostly random in nature, hence, these algorithms will help in making the system autonomous. And the algorithms are selected with least MSE. Hence, a list of algorithms are available to automate various levels of wireless communications[5]. In this paper two algorithms are investigated for Karachi RF spectrum measurements. These are NARX and CFFBNN. The results show that the NARX algorithm performs better than CFFBNN.

The rest of the paper is as under. Section II discusses the NARX and CFFBNN, Section III presents the related work, Section IV shows the simulation Results and discussion Section V presents the conclusion of the paper.

II. NARX and CFFBNN

Literature shows that the ANNs have proved to be very efficient for time series and modeling data in various fields including financial, communication network traffic prediction, chaotic time prediction and also for noisy data.

A. Nonlinear Autoregressive Exogenous Scheme

NARX, a type of nonlinear model proposed by Leontaritis and billings is a nonlinear model. It is used for estimating the future values of the time series based on its outputs and exogenous input.

A feature of nonlinear auto regressive models with exogenous inputs (NARX) recurrent neural network is that their architecture has limited feedback, which is dependent only on output neurons and not on hidden neurons. NARX networks are computationally powerful. That is why NARX models are considered to predict a wide class of nonlinear behaviors.

NARX is a discrete-time nonlinear system [7].

The typical architecture of NARX dynamic neural network contains three layers namely input, hidden and output. Unlike other types of dynamic networks like Elaman and Layer recurrent network, there is no context layer. Also that the output is feedback to input in this type of network [8]. In terms of configurations used for NARX network, Fig. 1 Shows two commonly used configurations as found in the literature. In series parallel architecture, an open loop mechanism is deployed which means that only true inputs are used to predict the future values of time series data and estimated output values are not fed back as inputs for this purpose. In completely parallel architecture, output values are fed back as input for prediction of output of time series data. In general, series parallel architecture is used during the training phase of the system and Parallel architecture is used for multistep ahead prediction [9].

B. Cascade Feed forward Backpropagation Neural Network

ANN functions like human neuron which is interconnected to one another. A generic neural network consists of many layers namely input layers, hidden layers and output layers. CFFNN is one of the design types used to develop neural networks. In CFFNN, neurons are connected with both preceding layers as well as all the neurons in layers. The theory behind the vast number of connections is that more connections provide better learning capability for the proposed ANN setup. These networks also used back propagation algorithm for updating weights. In this aspect they are similar to feed forward network. Different algorithms are used to change the weights of the networks. Common examples are Bayesian and Marquardt algorithms. Cascade feed forward network is a type of network which is considered to be efficient as well as flexible. Fast learning is also a property associated with Cascade feed forward network [10].

Another example of Cascade feed forward network is shown in Fig. 4. It can be seen that input neuron is connected to hidden neuron and hidden neuron has connection not only with input neuron but also with output neuron [12].

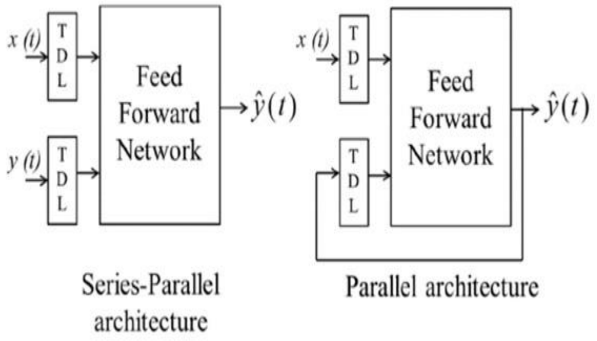


Fig. 1. Series Parallel Architecture of NARX Model [11].

III. Related Work

Cascade feed forward networks have also been used in Cognitive Radio technology.

Iliya et al. have used Cascade feed forward network in a series of experiments along with other algorithms for the prediction of real world RF power within the GSM 900, Very High Frequency (VHF) and Ultra High Frequency (UHF) FM and TV bands. Back propagation algorithm was used in Cascade feed forward network. Authors used sensitivity analysis in order to reduce the input vectors of the prediction models. Experiments showed that Cascade Feed forward network outperformed in terms of Average Mean Square Error (AMSE) for 30 independent runs and their standard deviation [13].

In [14], authors explore the secondary use of RF spectrum under LTE cellular environment. Power spectral density (PSD) of randomly chosen primary user signals is calculated. The prediction performance of the proposed setup is compared using NARX and Auto-regressive integrated moving average (ARIMA). Mean Square Error (MSE) is taken as performance criteria. Additionally, authors also proposed hybrid system comprising of sequential combination of ARIMA followed by NARX. Authors compare the performance of proposed hybrid scheme with ARIMA and NARX. The results of the simulation setup show that hybrid model performs better than NARX by 15.15% and 9.68% than ARIMA in some cases while 40% better than NARX and 33.33% than ARIMA in other cases. The proposed study recommends that the behavioral modeling of licensed activity under LTE setup can prove to be highly useful for secondary users. Especially, due to the fact that no a prior information regarding primary users is required. Hence, it is recommended that the proposed setup can be used by opportunistic users to exploit licensed bands in secondary fashion for optimal use of RF spectrum.

IP based connections typically show a dynamic traffic requirement behavior. This is due to the fact that different applications such as live video conferencing, audio communication, image transmission etc. require different amount of bandwidth. Some of these applications require higher bandwidth requirement than others such as Quality of Service (QoS) based transmissions and video communications. Hence, to resolve the imminent issue of bandwidth requirement, authors in [15] proposed smart VPN bonding scheme. In this scheme, more channels are bonded together to fulfill the requirements of end user. Additionally, the proposed scheme also provides better load balancing. Throughput of the proposed setup is measured that shows promising results.

Cognitive Users initiate the transmission in secondary fashion by following cognitive cycle. That includes spectrum sensing, decision making to shift the transmission, interference testing and etc. In [11], authors propose a sensing-transmission scheduling scheme. The performance of various artificial neural network based architectures is compared under periodic data as well as non-repeating data. The algorithms considered are NARX, feed forward, focused time delay, Elman, distributed time delay, cascade feed forward back propagation and layer recurrent network. The comparison between cascade feed forward back propagation neural network (CFBPNN) and

NARX neural network is presented in the paper. For periodically repeating data, NARX algorithm takes 17.32 seconds in comparison to CFBPNN that takes 95.79 seconds to complete the operation. The optimal percentage achieved through NARX is 97.86% and CFBPNN 95.79%. For non-repeating data, NARX achieves 98.16% using 7.65 seconds in comparison to achieving 100% from 3.83 seconds by CFBPNN. Additionally, MSE for CFBPNN is achieved at 0.0091743 in comparison to 0.088502 by NARX.

Cognitive Radio is a novel concept that promises to solve the imminent issue of spectrum scarcity through opportunistic use. Cognitive Radio may operate on many different schemes. These include underlay, overlay and interweave radio schemes [16].

In underlay cognitive radio environment, primary users and unlicensed cognitive users both exist at the same time. However, the coexistence can only be achieved successfully if the users are following the set rules. Two of such issues are addressed by the authors in [17]. These include the interference limit from primary network for the secondary user network and the amount of interference; secondary users are creating for primary users. The issue of interference can be addressed through a learning algorithm such that the sensor should continually detect the amount of RF power and report to central station. And the central station should have the rights to allow or prevent the unused spectral spaces based on the interference limit.

In [18] authors use NARX based classifier to detect the tumors. The feature extraction is done through Principal Component Analysis (PCA). Additionally, back propagation is used to train the proposed network and NARX is used to make classifications. The results of the proposed setup suggest that the NARX has great potential to detect tumors.

In [19], authors use NARX to predict various time series cases. Additionally, a comparison between real and artificial chaotic data from various experiments is presented. Three conclusions are drawn from the proposed simulation of network. The NARX has a potential to capture the nonlinear dynamic behavior of proposed setup. Another concluding remark is that these algorithms also have some limitations such as challenges in learning long time dependencies due to vanishing gradient. Additionally, these algorithms have limitation of optimizing the embedded memory. Furthermore, Recurrent Neural Network (RNN) model highly affects the performance of prediction. Hence, it is suggested that the user should always avoid the over fitting and saturation condition because too many hidden layers lead to poor prediction.

In this paper, NARX and CFFBNN are compared for Karachi dataset. The algorithms will help in improving the opportunistic activity for Karachi city based on the permission from the concerned departments in the country. For this purpose, a dataset is also developed by using NI 2901 USRP devices. The secondary activity is already allowed by FCC in USA in broadcast bands. The next section explains the details of the experimentation and analysis of the simulation results.

IV. Simulation Results

This section presents the results of the two algorithms selected and simulated for Karachi RF spectrum measurements. The frequency of operation is selected as 1805 MHz. Fig. 2 shows the performance metric for cascade feed forward back propagation neural network. This includes training, validation, test and best results for the algorithm. The MSE for the proposed setup comes out to be is 6.8877e-5 at epoch 16. Fig. 3 shows the performance of learning based algorithm i.e. NARX for the proposed setup. The MSE results for the proposed setup come out to be 3.1506e-11 at epoch 26. Thus, showing a 61% increase in the epochs for achieving an excellent MSE results. The results clearly suggest that the performance of NARX algorithm supersedes its opposite algorithm; however, this performance is achieved on the basis of more training time. Epoch represents the number of passes the algorithm sees the data set. One epoch is equal to one forward and one backward pass. Hence, it can be concluded that MSE performance and no. of epochs are showing a trade-off.

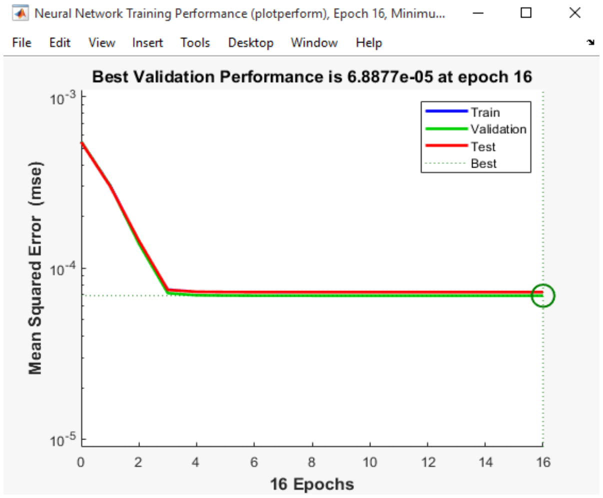


Fig. 2. Shows the Performance Results of Cascade Feed forward back Propagation Neural Network.

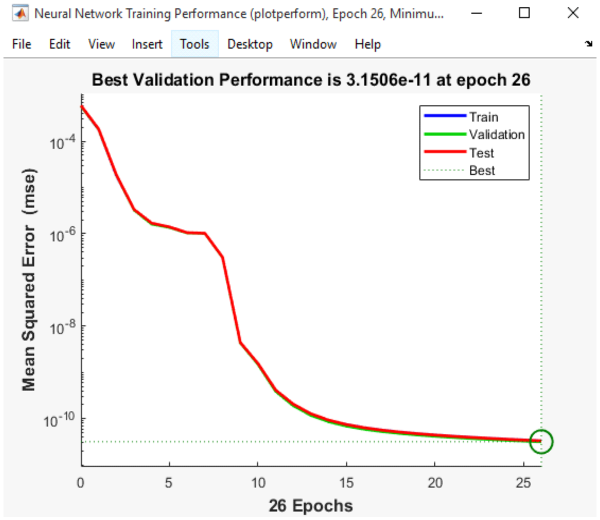


Fig. 3. Shows the Performance Results of NARX.

V. Conclusion

Performance analysis of two machine learning algorithms i.e. NARX and CFFBPNN is analyzed and compared for Karachi city. The results of the proposed network will allow the future secondary users of the wireless network to exploit the RF spectrum in opportunistic fashion such that a ubiquitous connectivity could be provided even under the usage of existing RF network. The proposed algorithm will be highly useful especially for 6G and 6G+ wireless communication standards. NARX algorithm takes 26 epochs for producing the best possible statistics regarding training and validation while CFFBNN produces the best statistics n 16 epochs. However the CFFBNN results in MSE of 6.8877e-5 in comparison to MSE produced by NARX that comes out to be 3.15 e-11. Hence, the prediction results advocate the use of NARX algorithm in future generation wireless radios with autonomous capabilities in comparison to CFFBNN.

The future work will be to investigate Deep Learning algorithms for prediction of spectral holes in Karachi. Additionally, the task will be to compare the performance between NARX, CFFBNN and Deep Learning based algorithms. The focus will be to recommend best possible algorithms, so that secondary use of RF spectrum can be successfully implemented in Karachi.

Acknowledgment

The authors would like to thank the Administration of NED University of Engineering & Technology, Karachi, Pakistan, for providing resources to complete this research.

References

[1] M. H. Miraz, M. Ali, P. S. Excell, and R. Picking, "A review on Internet of Things (IoT), Internet of everything (IoE) and Internet of nano things (IoNT)," in 2015 Internet Technologies and Applications (ITA), 2015, pp. 219–224.

[2] A. A. Khan, A. Z. Shaikh, S. Naqvi, and T. Altaf, "A Novel Cognitive Radio Enabled IoT System for Smart Irrigation," J. Inform. Math. Sci., vol. 9, no. 1, pp. 129–136, 2017.

[3] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," IEEE Netw., 2019.

[4] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," IEEE Commun. Mag., vol. 57, no. 8, pp. 84–90, 2019.

[5] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," IEEE Wirel. Commun., vol. 24, no. 2, pp. 98–105, 2016.

[6] M. A. Khan, A. Z. Shaikh, S. Naqvi, S. Khadim, and T. Altaf, "Deep Learning Enabled Spectrum Sensing Radio for Opportunistic Usage," IJCSNS, vol. 19, no. 11, p. 179, 2019.

[7] H. Xie, H. Tang, and Y.-H. Liao, "Time series prediction based on NARX neural networks: An advanced approach," in 2009 International conference on machine learning and cybernetics, 2009, vol. 3, pp. 1275–1279.

[8] M. Saberivahidaval and S. Hajjam, "Comparison between performances of different neural networks for wind speed forecasting in P ayam airport, I ran," Environ. Prog. Sustain. Energy, vol. 34, no. 4, pp. 1191–1196, 2015.

[9] J.-A. Ryu and S. Chang, "Data Driven Heating Energy Load Forecast Modeling Enhanced by Nonlinear Autoregressive Exogenous Neural Networks," 2019.

[10] I. Khan, H. Zhu, D. Khan, and M. K. Panjwani, "Photovoltaic Power prediction by Cascade forward artificial neural network," in 2017 International Conference on Information and Communication Technologies (ICICT), 2017, pp. 145–149.

[11] I. K. Aulakh, "ANN Application in Sensing-Transmission Scheduling in Cognitive Radio."

[12] G. Renisha and T. Jayasree, "Cascaded Feedforward Neural Networks for speaker identification using Perceptual Wavelet based Cepstral Coefficients," J. Intell. Fuzzy Syst., no. Preprint, pp. 1–13, 2019.

[13] S. Iliya, E. Goodyer, J. Gow, J. Shell, and M. Gongora, "Application of artificial neural network and support vector regression in cognitive radio networks for RF power prediction using compact differential evolution algorithm," in 2015 federated conference on computer science and information systems (FedCSIS), 2015, pp. 55–66.

[14] R. T. Fleifel, S. S. Soliman, W. Hamouda, and A. Badawi, "LTE primary user modeling using a hybrid ARIMA/NARX neural network model in CR," in 2017 IEEE Wireless Communications and Networking Conference (WCNC), 2017, pp. 1–6.

[15] G. Capizzi et al., "Available bandwidth estimation in smart VPN bonding technique based on a NARX neural network," in 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), 2017, pp. 601–606.

[16] A. Z. Shaikh and L. Tamil, "Cognitive radio enabled telemedicine system," Wirel. Pers. Commun., vol. 83, no. 1, pp. 765–778, 2015.

[17] F. S. Mohammadi and A. Kwasinski, "Neural network cognitive engine for autonomous and distributed underlay dynamic spectrum access," ArXiv Prepr. ArXiv180611038, 2018.

[18] H. P. H. Anh, "Medical Image Classification and Symptoms Detection Using Fuzzy NARX Technique," in 4th International Conference on Biomedical Engineering in Vietnam, 2013, pp. 335–342.

[19] E. Diaconescu, "The use of NARX neural networks to predict chaotic time series," Wseas Trans. Comput. Res., vol. 3, no. 3, pp. 182–191, 2008.

Comparative Study of Truncating and Statistical Stemming Algorithms

Sanaullah Memon[1]
Department of Information Technology, Shaheed Benazir Bhutto University, Shaheed Benazirabad
Sindh, Pakistan

Prof. Dr. Ghulam Ali Mallah[2]
Department of Computer Science
SALU, Khairpur, Sindh, Pakistan

K.N.Memon[3]
Department of Mathematics and Statistics, QUEST, Nawabshah,
Sindh, Pakistan

AG Shaikh[4], Sunny K.Aasoori[5]
Department of BSRS
QUEST Nawabshah
Sindh, Pakistan

Faheem Ul Hussain Dehraj[6]
Department of Business Administration
Shaheed Benazir Bhutto University
Shaheed Benazirabad
Sindh, Pakistan

***Abstract*—Search and indexing systems bear a significant quality called word stemming, is lump of content excavating requests, IR frameworks and natural language handling frameworks. The fundamental topic in the search and indexing through time is to upgrade infer via robotized diminishing and fussing of the words into word roots. From index term by evacuating any connected prefixes and postfixes, Stemming is done to proceeding piece of work of index word, and more extensive idea than the real word is spoken by trunk. In an IR framework, the numeral of recovered archives is expanded by stemming process.**

Keywords*—Stemming; truncating; statistical; NLP; IR; Lovins; Porters; Paice/Husk; Dawson; N-gram; HMM; YASS

I. Introduction

In present days, indexing and search systems support the word stemming and are in twist chunk of Natural Language Processing (NLP) systems, Information Retrieval (IR) systems and Text Mining applications. The principal concept is to ameliorate recollect by lessening the words to their root's word [1]. Before the function of the index word, stemming is done and the concept of stem is broader than the actual term. The number of completed forms is enlarged through stemming operation in IR systems. Before any related algorithm is actually applied, the summary, categorization and text clustering is also needed as chunk of the pre operation.

Generally the data castoff to stock besides get search calls within IR refers to the identification lists identified as lead words. Work in the IR systems involves only recent Pakistani language interests. The evolution of such structures is constrained by the dearth of inaccessibility of language assets and utensils in these languages.

Endeavors are made to create more powerful explore drives for stemmer. Nearly all IR systems are used to lessen the structural alternatives of a word to its root [2]. Generally IR systems used to tokenize printed archives and use stemming to lessen the quantity of marks and catch semi-identical standings resulting from the root [3].

II. Problem Statement

Stemmer is unique method that uses the information retrieval system to lessen a word's structural alternatives to its stem. In recent years, the huge growth in the content of the Urdu and Sindhi web has enlarged the necessity for active algorithms and stemming methods. Stemmer allows us to lessen a word to its root. The execution of stemming algorithms in information retrieval has been an extended-ranking issue.

Now many algorithms for stemmers of different languages embracing language founded on Arabic Script have been designed and suggested. Yet small search is documented in the prose regarding the Sindhi and Urdu languages. The aim of study is to suggest generic stemmers, thorough explanations, deliberations and assumptions of the numerous procedures of recycled for abbreviating and arithmetical mechanisms, approaches and devices that are in trend and have been recycled and applied before.

III. Objectives

- To study numerous methods from the literature that were recycled and executed for the Stemming of the numerous languages
- To analysis the algorithmic mechanisms of automatic stemmers' Truncating and Statistical approaches.
- For stemmers of various languages, the tests of the methods used Truncating and the statistical algorithms are previously performed.
- To examine cause and issues behind intended consequences based on stated results of certain algorithms.
- To suggest a more suitable stemmer algorithm for Pakistani languages i.e. Urdu and Sindhi which could yield near outcomes satisfactory.

IV. Overview of Stemming

Stemming is the procedure of reducing the amount of modulated words to find root and is the support mechanism for numerous NLP applications with the IR method meanwhile the search process is based only on the word's stem. Stemming gives an IR system two significant advantages. First, it improves the system's recall as the query words are harmonized in the documents with their morphological versions and second it lessens catalogue scope resulting in noteworthy advantages in haste and retention necessities. Following Table I displays the stemming system.

TABLE. I. Stemming System

Prefix	**Stem**	**Suffix**	**Complete Word**	**Meaning**
نا	امید	ي	نا اميدي	Hopelessness
اٹ	لک	یا	اٹلیکیا	Unwritten
وڈ	ڦڙ	و	وڈڦڙو	Heavy Rain
پر	دیس	ي	پرديسي	Foreigner

V. Arrangement of Stemming Algorithms

It is possible to classify stemming algorithms into classes. There is a typical way for each of these groups to find the stems of word variants. Fig. 1 displays the arrangement of stemming algorithms.

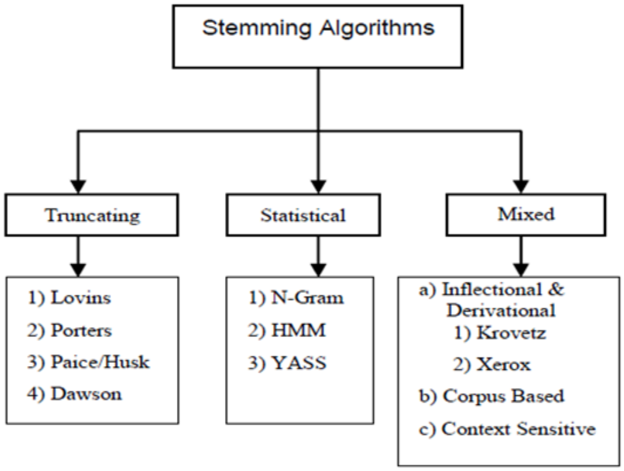


Fig. 1. Arrangement of Stemming Algorithms.

VI. Truncating Methods (Affix Removal)

The approaches are associated to eliminating a words' suffixes or prefixes (usually referred to an affixes), as clearly suggested by name. This was a simplest stemmer which shortened a term at the nth sign. Terms shorter than n are held as they are in this system. Over stemming chances increase when the length of the word is short.

Another modest method was the S-stemmer, a procedure that combines singular and plural noun shapes. Donna Harman suggested this algorithm [4].

The algorithms have rules for the deletion of plural suffixes so that they can be transformed to the unique forms. Truncating algorithms are the most widely used stemmer.

A. Lovins Stemmer

This was Lovins' initially prevalent and operative stemmer in 1968. It achieves a lookup on 294 end tables, 29 conditions and 35 rules for transformation. The stemmer of Lovins eliminates from a phrase the longest suffix. Upon removal of the ending, the term is recoded by a dissimilar table that requires different adjustments to translate these trunks into valid words. Unpaid to its flora as a single license procedure it always eliminates an extreme of single suffix from a word.

B. Porters Stemmer

Porters stemming algorithm was proposed in 1980 as the most popular stemming method. Many modifications and improvements on the basic algorithm were made and suggested [5]. It has five phases and directional are functional in each step until the criteria are passed by one of them. If a rule is adopted, the suffix will be deleted and the next step will be taken.

C. PAICE / HUSK Stemmer

This stemmer indexed to approximately 120 directions by the preceding memo of a suffix [6]. On each repetition, the past character of the term tries to search an appropriate law. Every law defines a termination, deletion or replacement. If no such law exists, it will stop.

D. Dawson Stemmer

It provides a large additional complete gradient of around 1200 suffixes. It has a one-pass stemmer too, so it's pretty quick. The suffixes are classified by their length and last letter in the reversed order indexed.

VII. Statistical Methods

Another solution to suffix striper is suggested by Prasenjit [7]. These stemmers depend on strategies and factual investigation. For example, most statistical stemmers used the Hidden Markov Model approach based on N-Gram. Melucci proposed a model using automatons of finite-state where the function of probability regulates transitions between states.

A. N-Gram Stemmer

An N-gram is a sequence of n characters, typically contiguous, extracted from a continuous text segment to be precise, a N-gram is a traditional of n consecutive characters dig out from a word. The key clue behind this approach is that a high proportion of N-grams will be shared by similar words.

B. HMM Stemmer

Melucci and Orio suggested this model [8]. In this method, it is probable to calculate the possibility of each track and invention the most likely path by the Viterbi coding in the automatic graph. To apply HMMs for stemming, the product of a concatenation of two subsequences can be viewed as a arrangement of letters that makes a word, a prefix and a suffix.

C. YASS Stemmer

The presentation of a stemmer produced by groups a wordlist without any dialect input is corresponding to that achieved using ordinary rule based stemmers like Porter's according to the authors. Groups are recognized using tiered approach and space measurements. The resulting clusters are then measured to be classes of equivalence and their centroids to be the stems.

VIII. Literature Review

Since the decade, several stemmers have been produced and available on the computer and internet market. It is noted that stemmer is necessary part of any culture's process of gathering knowledge. Stemmer is the fundamental component of the IR process. Stemmers have been found to be the simplest type of all morphological systems. Since the absolute starting point of the information recovery period, the main group focus was established to support various languages and efficient algorithms. Majority of the stemmers work achieved in advance changes into based totally on regulations. The linguistic inputs based on the preparation of rules are very complex and hard work. In addition, it calls for excellent linguistic understanding to design such stemmers. Earlier stemmers were designed for the English language on a rule based approach. The first rule based stemmer was developed by Lovins [9]. Around 260 language rules have been mentioned for this purpose in order to curb the English language. Lovins' approach was the heuristic iterative longest match. Martin porter offered the most outstanding effort in the field of rule based stemmer [10]. He condensed Lovin's laws to roughly 60 guidelines. Porter stemmer algorithm, he has developed. This algorithm is very simple, effective and commonly used for search engine creation.

Urdu is a well-spoken language throughout the world and much work has been done on the stemming of Urdu. Riaz explained the Urdu stemming challenges and introduced a rule based model with a few rules that were enforced to inspire the specifics for Urdu [11]. This showed that originating from Urdu, due to the complex nature of Urdu is quite difficult.

Kansal has proposed a rule based on stemmer [12]. He established and implemented rules for this purpose to eliminate the suffix and prefix from the inflected words of Urdu. In addition the rule based stemmer for the Urdu language was created by Gupta [13].

By applying the truncation of affixes, light weight Stemming is to find a representative type of word indexing [14]. In Urdu, for a single word form there are large numbers of variant variants. Khan raised a number of morphological questions relating to Urdu stemmer's law-based development [15].

IX. Experiments and Results

A. *Results Recorded using the Lovins Stemming AlgoritHM*

Various writers castoff the Lovin's stemming algorithm to measure corpus accuracy by using various languages such as English, Urdu, Arabic and Sindhi. Table II shows the Lovin's stemming algorithm's precision.

With 99.1% and 93.37% precision, Wahiba and Sandeep used English, Haider used 100% precision of Arabic Language. When using 20583 words and 50000 characters, Rohit and Qurat-ul-ain used urdu with 85.15% and 91.2% precision. Fig. 2 demonstrates the precision of terms by giving different languages to readers.

TABLE. II. Accuracy of Lovin's Stemming Algorithm

Title	**Author**	**Language**	**Corpus**	**Accuracy %**
A new stemmer to improve Information Retrieval	Wahiba et al..	English	30000 Words	99.1
An effective stemmer in Devanagari script	Dogra et al.,	Devangari	1670 Words	94.26
Strength and accuracy analysis of Affix Removal Stemming Algorithms	Sandep et al.,	English	29417 Words	93.37
A Rule Based Extensible stemmer for information retrieval with application to Arabic	Haidar et at.,	Arabic	-----	100
Rule based Urdu Stemmer	Rohit et al.,	Urdu	20,583 Words	85.15
Assas-Band, an Affix-Exception-List Based Urdu Stemmer	Qurat-ul-Ain et al.,	Urdu	50,000 Words	91.2
Stemmer of Sindhi Secondary words for Information Retrieval System Using Rule based stripping Approach	Mohsin	Sindhi	50,327 Words	84.85

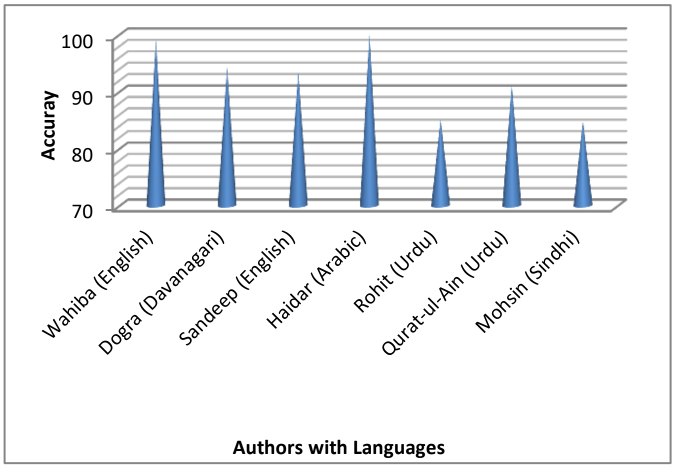


Fig. 2. Lovins Algorithm Stemming Accuracy.

B. Results Recorded using the Porters Stemming Algorithm

Various writers recycled this algorithm to measure the reliability of the corpus by using different languages. The accuracy of the Porters stemming algorithm is discussed in Table III.

Sandeep, Joshi and Wahiba used English with 73.61 percent, 98.4 percent and 99.5 percent accuracy. Widjaja used Indonesian with 96.31 percent accuracy, Guastad used Dutch with a consistency of 79.23 percent when using 45000 words. Fig. 3 demonstrates the accuracy of the concept when offering different languages to authors.

TABLE. III. PORTERS ACCURACY OF THE STEMMING ALGORITHM

Title	Authors	Language	Corpus	Accuracy%
Implementation of Porters Modified Stemming Algorithm in an Indonesian word Error Detection Plug in Application	Widjajja et al.,	Indonesian	3000 Words	96.31
A new stemmer to improve information Retrieval	Wahiba et al.,	English	30000 Words	99.5
Accurate stemming of Dutch for text classification	Gaustad et al.,	Dutch	45000 Words	79.23
Development of a stemmer for the Greek language	Georgios et al.,	Greek	880 Words	92.1
Strength and accuracy analysis of Affix removal stemming algorithms	Sandeep et al.,	English	29417 Words	73.61
Modified Porter Stemming Algorithm	Joshi et al,	English	30K Words	98.4

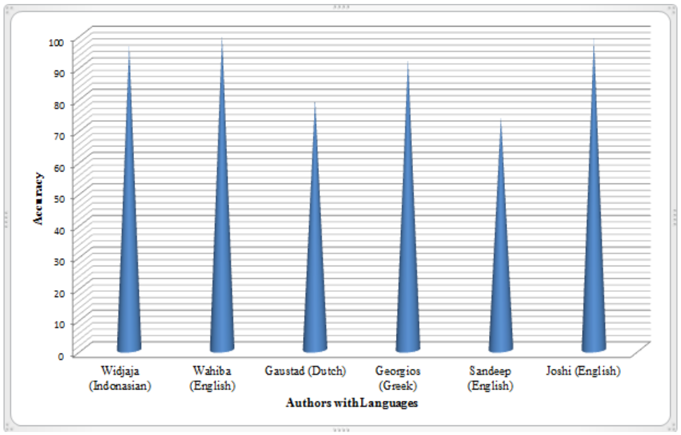


Fig. 3. Porters Precision Stemming Algorithm.

C. Results Recorded using the PAICE / HUSK Stemming Algorithm

Various authors used the Paice / Husk Stemming Algorithm to calculate corpus accuracy using various languages such as English and Portuguese. The accuracy of the Paice / Husk Stemming Algorithm is explained in Table IV.

Wahiba, Chris and Sandeep utilized English with 99.3 percent precision, 67 percent precision and 95.47 percent precision. Orengo utilized Portuguese dialect with 96 percent precision while the utilization of 30000 words. Fig. 4 appears phrase precision via providing readers exclusive languages.

D. Results recorded using N-Grams Stemming Algorithm

Various authors used the stemming algorithm N-Grams to calculate corpus accuracy using different languages such as Arabic, Malay, Marathi and English. The N-Gram stemming algorithm accuracy is explained in Table V.

TABLE. IV. PAICE / HUSK STEMMING ACCURACY ALGORITHM

Title	Authors	Language	Corpus	Accuracy%
Strength and Accuracy Analysis of Affix removal stemming algorithms	Sandeep et al.,	English	29417 Words	95.47
A new stemmer to improve information retrieval	Wahiba et al.,	English	30000 Words	99.3
Stemming algorithm for the Portuguese language	Orengo et al.,	Portuguese	2800 Words	96
An Evaluation Method for Stemming Algorithm	Chris et al.,	English	9,757 Words	67

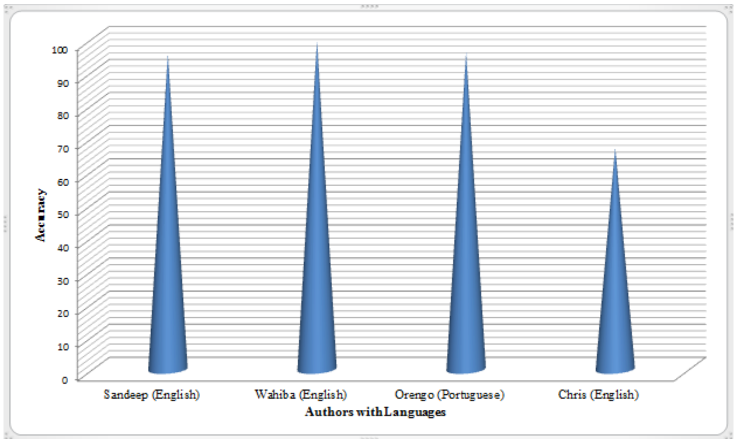


Fig. 4. Precision of PAICE / HUSK Stemming Algorithm.

TABLE. V. N-GRAMS PRECISION STEMMING ALGORITHM

Title	Authors	Language	Corpus	Accuracy %
Generation, implementation and Appraisal of an N-gram based Stemming Algorithm	Oande et al.,	English	COCA	96.7
Discovering suffixes: A case study for Marathi language	Majgao-nker et al.,	Marathi	1500 Words	82.50
Corpus-Based Arabic Stemming Using NGrams	Zitouni et al.,	Arabic	1000,000 Words	99.7
Effectiveness of stemming and Ngrams String Similarly Matching on Malay Documents	Sembok et al.,	Malay	2238 Words	98.2
Comparison of Stemming and N-Grams Matching for term Conflation in Arabic Text	Hani et al.,	Arabic	50000 Words	96

With 99.7 percent and 96 percent accuracy, Zitouni and Hani used Arabic. Pande used English language is 96.7 percent accuracy. Majgaonke recycled Marathi language when using 1500 words, which was 82.50 percent accuracy. Sambok used Malay language for 98.2 accuracy when using 2238 words. Fig. 5 demonstrates word accuracy by giving different languages to writers.

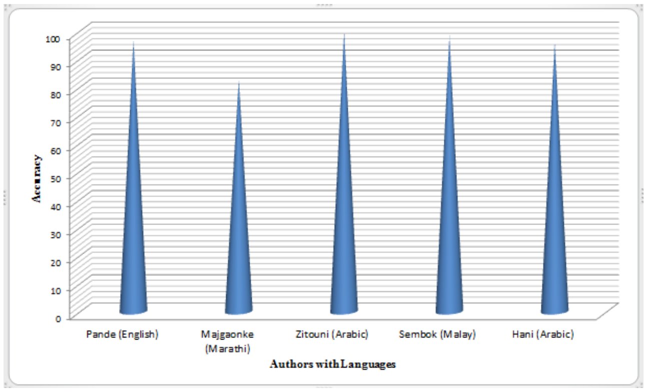


Fig. 5. N-Gram Precision of the Stemming Algorithm.

E. Results Recorded using HMM Stemming Algorithm

Various authors used the HMM Stemming Algorithm to calculate the accuracy of corpus by using different languages such as Arabic, Persian, Assamese and English. Table VI shows the HMM stemming algorithm accuracy.

TABLE. VI. HMM ALGORITHM STEMMING ACCURACY

Title	Authors	Language	Corpus	Accuracy%
An improved stemming approach using HMM for a highly inflectional language	Navanath et al.,	Assamese	2000 Words	92
PHMM: Stemming on Persian Texts using Statistical Stemmer based on hidden Markov Model	Fatemeh et al.,	Persian	500 Words	79
Hidden markov model based Arabic morphological analyzer	Alajmi et al.,	English	15 Million Words	95
A novel method for stemmer generation based on hidden markov models	Massimo et al.,	Arabic	1950 Words	90.5

Alajmi used 15 Million words to get 95 percent accuracy in English. Massimo used Arabic Language 90.5 percent were right when using the 1950 terms. Using 500 letters, the Persian language used by Fatimah was 79 percent correct. Navanath used Assamese language, 92 percent accuracy when using 2000 words. Fig. 6 displays the accuracy of the word by giving writers different languages.

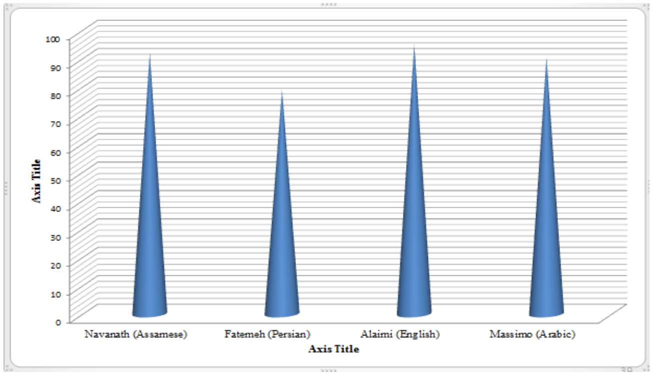


Fig. 6. Precision of HMM Stemming Algorithm.

F. Results Recorded using YASS Stemming Algorithm

Several researchers used different titles to calculate corpus accuracy when using the YASS stemming algorithm with different languages like English, Hungarian and Kebang. The precision of this stemming algorithm is shown in Table VII.

TABLE. VII. YASS ALGORITHM STEMMING ACCURACY

Title	Authors	Language	Corpus	Accuracy%
YASS: Yet Another Suffix Stripper	Prasenjit et al.,	English	262128 Words	96.5
Hungarian and Czech Stemming using YASS	Prasenjit et al.,	Hungarian	536678 Words	86.68
The First Step Towards Suffix Stripping of Missing Words using YASS	Sadiq et al.,	Kebang	30,000 Words	87

Prasenjit using English with the precision of 96.5 percent when using 262128 letters. Prasenjit also had 86.68 percent accuracy when using 536678 letters.By using 30000 letters, the accuracy of Sadiq used Kebang language was 87 percent. Fig. 7 demonstrates word accuracy by giving different languages to readers.

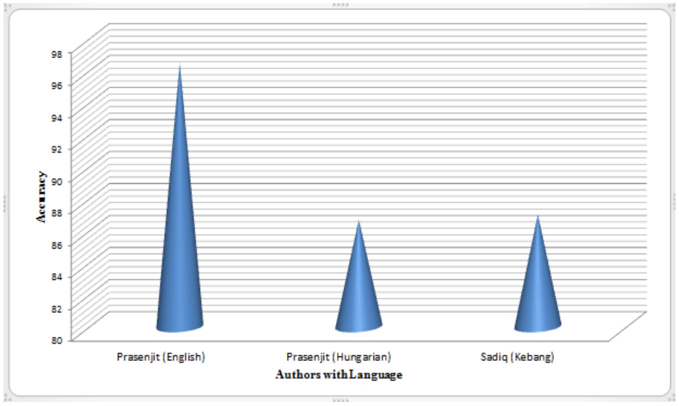


Fig. 7. YASS Precision Stemming Algorithm.

X. Discussion

The steps of Dawson Stemming Algorithm are similar as Lovins algorithm in a great extent. Hence, researchers use Lovins algorithm in its place of Dawson Algorithm.

By using the Lovins stemming algorithm, 100% accuracy is calculated by Haidar using the corpus of Arabic language. And maximum accuracy is reported with the data set of English which is 99.1%. This kind of algorithm is also used with Sindhi by Mohsin but he has not achieved good results as relate to Arabic and English since partial linguistics rules were applied. If he raises the commands then accuracy may also increase.

Wahiba and Sandeep implemented Lovins, Porters and Paice/ Husk algorithms on the corpus of English and accomplished acceptable level results.

Among the statistical stemming algorithms, most of the researchers used N-Grams Based algorithm for the task of stemming. Zitouni achieved 99.7% accuracy with Arabic and Sembok calculated 98.2% accuracy with malay. Concluded the nature of both languages are entirely opposite from each other but due to the N-gram language modeling is does not affect the performers of the stemmers.

Limited number of researchers used YASS stemming algorithm for researchers believed that this algorithm is tough in terms of implementation and require more time for execution as compare to other statistical stemming algorithms.

XI. Conclusion

A relative evaluation of statistical and truncating algorithms in specific languages with in the literature is provided. The work purposes to suggest for the dissimilar dialects a generic stemming annotation. Truncating algorithms, especially Porters and Lovins, have been observed to be more appropriate for Sindhi, Urdu and other languages based on Arabic scripts since both processes are working on the specific rules of linguistics.

XII. Future Work

Although much research has already been done in the development of stemmers, much remains to be done to advance the accuracy.

A stemmer that uses both syntactic and semantic knowledge to reduce stemming errors should be developed.

For the Sindhi stemming method, the Porters and Lovins algorithms could be used to evaluate which set of rules is more appropriate for Sindhi.

Acknowledgment

This research has been conducted in SALU Khairpur as part of Master's dissertation.

References

[1] Jivane, A. G., (2011), "Comparative Study of Stemming Algorithms", International Journal of Computer Technology Applications, Vol. 2, Number 6, Pp. 1930-193.

[2] Husain, M. S., (2012), "An Unsupervised Approach To Develop Stemmer", International Journal on Natural Language Computing, Vol. 1, Number.2, Pp. 1523.

[3] Al-Omari, A., Abuata, B., (2013), "Building and Benchmarking New Heavy/Light Arabic Stemmer", The 4th International Conference on Information and Communication Systems.

[4] Harman Donna, (1991), "How effective is suffixing?" Journal of the American Society for Information Science, Vol. 42, Pp. 7-15 7.

[5] Lovins, J. B, (1968), "Development of a Stemming Algorithm," Mechanical Translation and Computer Linguistic, Vol.11, Number 1/2, Pp. 22-31.

[6] Porter M.F, (1980), "An Algorithm for Suffix Stripping", Program, Vol. 14, Pp. 130-137.

[7] Prasenjit, M., Mandar, M., Swapan K. Parui, G., Pabitra, M., (2007), "YASS: Yet another Suffix Stripper", ACM Transactions on Information Systems, Vol. 25(4).

[8] Melucci, M., Orio, N., (2013), "A Novel Method for Stemmer Generation based on Hidden Markov Models", Proceedings of the Twelfth International Conference on Information and Knowledge Management, Pp. 131-138.

[9] Lovi7*+ns, J. B., (1968), "Development of a stemming algorithm", Mechanical Translation and Computational Linguistics, Volume 11, Pp. 22–31.

[10] Porter, M., (1980), "An Algorithm for Suffix Stripping". Program, Volume 14, Number 3, Pp. 130-137.

[11] Riaz, K., (2007), "Challenges in Urdu Stemming", A Progressive Report In BCS IRSG Symposium: Future Directions in Information Access (FDIA 2007). http://citeseerx.ist.psu.edu/viewdoc/download/ 1 0.1.1.102.3051.pdf

[12] Kansal, R., Goyal, V., Lehal, G. S., (2012), "Rule Based Urdu Stemmer", Proceedings of COLING 2012, Pp. 267–276.

[13] Gupta, V., Joshi, N., Mathur, I., (2013), "Rule Based Stemmer in Urdu", In Proceedings of 4th International Conference on Computer and Communication Technology, Pp. 129-132.

[14] Al-Sughaiyer, I., Al-Kharashi, I., (2004), "Arabic morphological analysis techniques: a comprehensive survey", Journal of the American Society for Information Science and Technology, Volume 55, Number 3, Pp. 189 – 213.

[15] Khan, S. A., Anwar, W., Bajwa, U. I., Wang, X., (2011), "Challenges in Developing a Rule based Urdu Stemmer", Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing, Pp. 46–51.

Testing different Channel Estimation Techniques in Real-Time Software Defined Radio Environment

Ponnaluru Sowjanya[1], Penke Satyanarayana[2]
Department of Electronics and Communication Engineering
Koneru Lakshmaiah Educational Foundation, Vaddeswaram, Andhra Pradesh, India

***Abstract*—In modern wireless communication to maximize spectral efficiency and to minimize the bit error rate OFDM (Orthogonal frequency-domain multiplexing) is used. OFDM is used broadly in networks using various protocols, including wireless vehicular environment IEEE 802.11p, IEEE 802.16d/e Wireless Metropolitan Area Networks, Long-Term Evolution 3GPP networks and IEEE 802.11a/g/n Wireless Local Area Networks. The main challenges involved when using OFDM for wireless communications are short channel-coherence bandwidth and the narrow coherence time, and both have a major effect on the reliability and latency of data packet communication. These properties increase the difficulty of channel equalization because the channel may change drastically over the period of a single packet. Spectral Temporal Averaging is an enhanced decision-directed channel equalization technique that improves communication performance (as far as the frame delivery ratio (FDR) and throughput) in typical channel conditions. This paper reports tests of Spectral Temporal Averaging channel equalization in an IEEE 802.11a network, compared with other channel equalization techniques in terms of the FDR in a real-time environment. Herein, a software defined Radio (SDR) platform was used for estimating the channel. This proves that the system can provide over 90% of delivery ratio at 25 db of Signal to Noise Ratio (SNR) for various digital modulation techniques. For this purpose, an experimental setup consisting of software-defined radio, Universal Software Radio Peripheral (USRP) N210 along with wide bandwidth daughter board as hardware and GNU radio is used.**

Keywords*—Channel Estimation; GNU Radio Companion (GRC); Orthogonal frequency-domain multiplexing (OFDM); software-defined radio (SDR); Spectral Temporal Averaging (STA); Universal Software Radio Peripheral (USRP)

I. Introduction

The latest developments in wireless communications like phones, protocols and applications are impressive. For large data high data rates applications and low energy usage are important for upcoming wireless technologies. [1]. When the number of users with common Mobile Bandwidths expands, traffic congestion will be a harder task, and increased flexibility in the requirements of transmission will be required if multiple access contention is to be handled. Such problems are often faced by wireless communication protocols. The new mobile technology Standard Long-Term Evolution (LTE) focuses a lot on experimental and evaluation of wireless protocols.

A widely used modulation technique by means of which the flow of symbol is split in several narrowband subcarriers which are modulated progressively, and which have very close connections for high data transmission rates is OFDM. This method eliminates the frequency-selective loss signal. OFDM generates several non-selective parallel frequency streams, thus reducing inter-symbol interference (ISI). The decoder process is utilized to gain information on the channel state Information (CSI) and to account for channel effects such as delay spread and lower Doppler distribution and this process makes the OFDM system reliable.

SDR may provide flexible, upgradeable and longer lifetime radio equipment for the military and for civilian wireless communications infrastructure as shown in Fig. 1. SDR may also provide more flexible and possibly cheaper multi-standard-terminals for end users. It is also important as a convenient base technology for the future context-sensitive, adaptive and learning radio units referred to as cognitive radios. SDR also poses many challenges, however, some of them causing SDR to evolve slower than otherwise anticipated. Transceiver development challenges include size, weight and power issues such as the required computing capacity, but also software architectural challenges such as waveform application portability. SDR has demanding implications for regulators, security organizations and business developers. In a multicarrier technology such as OFDM, it is crucial to maintain orthogonality among all the subcarriers to prevent the Inter Carrier Interference (ICI) which if not done leads to significant performance degradation. In a high mobility environment, such as aerial vehicle communication, multi carrier transmission leads to severe ICI due to Doppler shift.

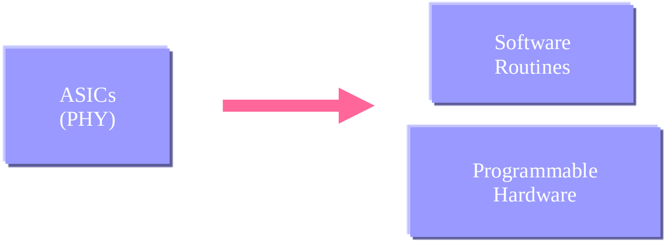


Fig. 1. Software Defined Radio.

II. Literature Review

Several authors have tried to use several channel estimations schemes to implement OFDM in real-time. with an emphasis on the Least Square (LS) channel estimation scheme and signal capture they, experimentally investigated the preamble detection, transmitter sensing and physical layer-capture effect [2]. Another team developed and evaluated a channel estimation system for pilot subcarriers with limited complexity, high precision, little pilot bandwidth and buffer-free data flow [3], and Great strength [4]. Using Field-Programmable Gate Arrays (FPGAs), that require changeable-power loading system of three levels, such experiments have been executed in real-time OFDM transceivers. To this end, a high-performance, radio-based software platform (CuSora) was developed as a modern processor for GPuS high-speed signal processing [5]. This software supports development of different protocols using an entire hardware and software framework. Signal processing components for the 802.11a and 802.16 protocols were tested in CuSora, and 2-dimensional linear interpolation was implemented for channel estimation in these tests. This simple algorithm modulates signals along both the frequency axis (FA) and the time axis (TA). Along the FA, the major technique is the evaluation of received samples and pilots within the frequency period between two adjacent pilot packets. Along the TA, the main determinant of transmission performance is channel estimation. Armour et al. discussed the potential trade-offs between efficiency and complexity when using the decision directed Least Mean Square (LMS) algorithm with the Hiperlan E and IEEE P802.1la standards [6].

SDR is an ideal solution for rapid prototyping and testing of wireless communication protocols [7][8]. The tools included in SDR platforms allow for easy reconfiguration of transceiver devices to quickly implement standards and protocols. SDR solutions can be used to test wireless sensor networks [9], satellite communications [10], and many other applications. Currently, research on implementing different systems using SDR technology is being conducted. In [11], the authors integrated their design of a joint DA-ML estimator using SDR combined with FPGA. And finally, they tested their prototype in a real-time environment. In [12], the author used USRP N200/N210 as a frontend and proved that the experimentation is possible at 60 GHz using SDR. In [13], using SDR platform, the authors designed, implemented, and evaluated a MIMO system with eight antennas. They chose LTE parameters as their system parameters. In [14], the authors detected the human activity using SDR platform. They extracted the channel state information using two USRPs as transmitter and receiver using 64-FFT point's OFDM modulation technique. In [15], the author implemented some of the communication features like pulse shaping, demodulation, and synchronization in a real=time environment; therefore, they have chosen the SDR platform and USRPs are used as SDR front end. Through these references, it can be concluded that SDR is a present trending technology.

For signal processing GRC is an open source platform. On a regular PC, this allows for the installation of a general-purpose processor. To support device networking, it is one of the best tools for SDR. This offers several projects to make a flexible software radio system utilizing various software libraries. In [16], The researcher received a 3-30MHz frequency band experimental SDR receiver with the aid of GRC / FPGA programmable hardware. In [17], Writer suggested a wide-band frame-level adaptation that can be simply implemented in various protocols. They showed their viability with GRC / SDR platform. In [18], Compatible range, instead of channel condition maintaining that energy-constrained systems scale down their sampling levels is called Sampless Wi-Fi. GRC / USRP was used to test their SDR system. In [19] and [20], the anti-jamming strategy was suggested in various methods and their approach was tested using GRC. In [21], The author has created a mechanism called TREKS to facilitate communication with Spread Spectrum without pre-shared information. They used a GRC and USRP test bed to evaluate this process. In [22], To research OFDM quality in the 802.11 standard, they built an SDR-based test bed. Consequently, GRC / USRP was chosen as their perfect platform. In [23], Two separate graphics processing unit implementation methods in a Software Defined Radio set-up were evaluated and it was found that only one proposed method was environmentally friendly. In [24], A new PHY / MAC protocol, known as Diversity-aware Wi-Fi, is developed and tested using USRP / GRC platform, and compared to existing methods. Despite this variety of features, GRC can be considered to be one of the strongest SDR apps [25].

In this paper, understanding the challenges of, dynamically changing channels, however, will have serious problems with Doppler and delay spread, which cause multipath fading channel effects. Therefore, we implemented a decision-directed spectral temporal averaging (STA) channel estimator [26] in real-time data transmissions considering the IEEE 802.11a protocol. We compared the STA scheme with LS, LMS, and comb-pilot linear interpolation schemes. We prepared a custom testbed for this study comprising GRC/USRP as their software and frontend.

We designed and tested an IEEE 802.11a transceiver with a selection of channel-estimation methods in real-time environment. The results show that the STA channel-estimation scheme achieves the best frame data rate of the techniques that we tested. Here, real-time signal constellations are also observed to find the effect of multipath propagation of signals through flat fading channels. Both medium access control (MAC) and physical (PHY) layers were implemented in the SDR platform for testing. A known stream of data was transmitted and received using various modulation schemes, and the FDR performance was analyzed for comparison.

The remaining document is as follows:

Section 2 contains information on the testing of channel estimation schemes.

Section 3 explains how we used the GRC platform and USRPs to implement this network in a real-time environment.

Section 4 introduces the channel-estimation techniques that we tested for comparison against the STA technique.

Section 5 describes the outcomes of tests and compares the outputs of the different channel estimates and Section 6 includes our conclusions.

III. Channel Estimation Techniques

Channel approximation is the analysis of a predetermined mathematical model of communication channels.The two variables that define the computational network design are short-term CSI / instant CSI and long-term CSI / statistical CSI.

The CSI provided statistical data for the long term, including statistical distribution and average channel gain. The only element used in the short-term CSI is channel impulse response. In OFDM systems the channel impulses were detected by the time / domain signal and channel frequency response, respectively, before and after DFT transmission, and they had been defined by frequency / domain channel estimation.[27]. Three methods can be used to approximate a channel: pilot-assisted, blind stream and Decision Directed Channel Estimation (DDCE).

The pilot-aided channel approach is one of the most standard way of calculating channels. A sender transfers in this system well-known pilot data used by both sender and recipient as proof. Pilot symbols are computationally somewhat complicated, but they are used in every wireless communication device. That method, though, decreases the bit rate since some symbols are used instead of information for pilots and the network space is lost. Even if the number of pilots has decreased, it is a challenge to estimate the channel precisely. As shown in Fig. 1, using blocks or combs for pilot assignment. A pilot block model is ideal for a slow fading stream, where the signal moves slowly. A comb pilot distribution is therefore suitable for the quick fading flow, since the pilots are arranged equally throughout the symbol sequence. To decide the channel response of the information symbols and consider this more vulnerable method to frequency-selective channels, interpolation between frequency-domain is required.

The structure of the block pilot is shown in Fig. 2. At the start of each subcarrier, a pilot data is there in the OFDM symbol. The time-domain interpolations are used to approximate the channel using these pilots. Since the opposite of the Doppler rate, f_{dp} in the channel provides continuity time, the pilot symbol duration will comply with the following variation:

$$S_t \leq \frac{1}{f_{dp}} \qquad (1)$$

The comb-pilot's design is also shown in Fig. 4. Frequency-domain interpolation stream pilot tones are incorporated in every OFDM symbol between subcarriers on a regular basis. Since the bandwidth of coherence of the reverse of the delay spread σ_{max}. maximum, the pilot symbol duration should satisfy the below change:

$$S_f \leq \frac{1}{\sigma_{max}} \qquad (2)$$

Estimation of the blind channel does not require pilot symbols and relies instead on intrinsic data received from symbols. Although no bandwidth of the signal is consumed by this approximation method, the computations are much more complicated and lead to higher bandwidth. For illustration, to estimate one channel coefficient, almost 100 symbols are needed. For this reason, this blind channel estimation method is seldom used in real-world wireless communication systems.

DDCE, which is our subject below, uses both observed channel approximation information symbols and pilot symbols. The estimated values are updated as diagrammed in Fig. 3. Thus, DDCE offers superior performance than pilot-aided channel estimation.

In our study, data were transmitted and received in real time to allow testing of channel-estimation techniques in a realistic context. We assumed that the channels are positioned dynamically. We tested four channel-estimation techniques: LS, LMS, comb-pilot linear interpolation, and STA. Of these, LMS and STA are decision-directed interpolation schemes.

A. Least Squares Equalizer

In modern hardware implementations, the basic LS equalizer algorithm is frequently used as a regular method [28]. For estimating the channel, in IEEE 802.11p the long training sequence is treated as block pilots. Let us say that after the start of the frame the two long-term preambles are referred to as

$$\underbrace{y_G[n_p-128],\ldots,y_G[n_p-65]}_{Long\ Preamble\ 1},\underbrace{y_G[n_p-64],\ldots,y_G[n_p-1]}_{Long\ Preamble\ 2}$$

And they're named T1 and T2. T1[n] and T2[n] are the time domain symbols for approximation of LS channels are derived from these two LPs. Instead they measure their N-point DFTs are represented below:

$$Y_1(k) = \sum_{n=0}^{N-1} T_1[n]\, e^{-\frac{2\pi jkn}{N}} \qquad (3)$$

$$Y_2(k) = \sum_{n=0}^{N-1} T_2[n]\, e^{-\frac{2\pi jkn}{N}} \qquad (4)$$

The N-points between two learning symbols are the same, i.e. X1(k)= X2(k)= X(k). The estimate of the LS for H(k) is as follows:

$$\hat{H}(k) = \frac{(Y_1(k)+Y_2(k))}{2X(k)} \qquad (5)$$

The data in the packet will be equalized after channel estimation is completed. The obtained signal's DFT is defined as Y(k). Through equalizing the obtained DFT function, the transmitted information is calculated.

$$\hat{X}(k) = \frac{Y(k)}{H(k)} \qquad (6)$$

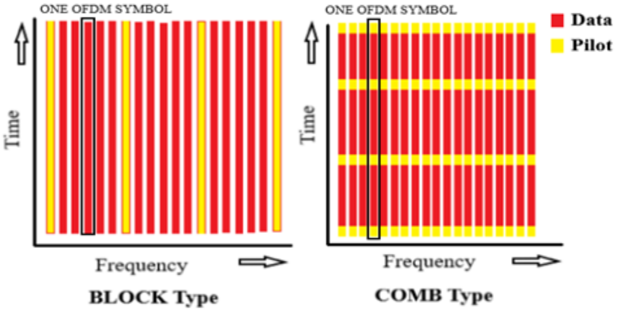


Fig. 2. Types of Pilot Assignments.

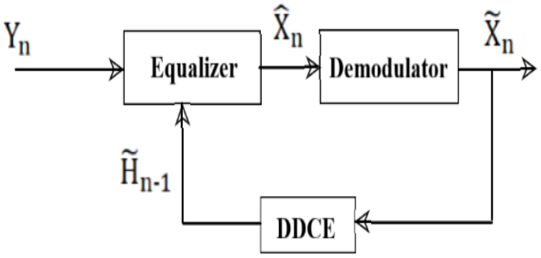


Fig. 3. Decision-Directed Channel Estimation.

A basic one-taps equalizer approach is used by each subcarrier in the frequency domain. The least square equalizer fits all the symbols in the packet. We cannot accurately represent the channel estimate H(k), if the channel drastically varies over the period of a packet. Also, equalization could falsify the received signal rather than making correction. So, a perfect and effective means of tracing the channel is needed.

B. Least Mean Squares

The LMS algorithm disables this constraint in channel tracing by acclimating the channel estimates while the signal is received. The original channel estimate is automatically adjusted to the time-variant properties of the communication channel by the LMS algorithm. The mean square error among the desired equalizer output and the actual equalizer output will be minimized using this algorithm. As the LS equalizer, beginning with a similar introductory decision after the i^th OFDM symbol update the channel utilizing the constellation point $\hat{X}_i$ onto where the obtained symbol Yi was deallocated:

$$\widehat{H}_i(k) = (1-\alpha)\widehat{H}_{i-1} + \alpha\frac{Y_i(k)}{\hat{X}_i(k)} \tag{7}$$

To average the coefficients of the time-domain channel that finds α, a low-pass filter is used. Within the equalizer filter length constraints, the signal-to-distortion ratio at its output is maximized in LMS equalizer. This approach is constrained because if the signal obtained is longer than the propagation time, the equalizer cannot decrease this distortion.

C. Comb-Pilot Interpolation

None of the LS or LMS algorithms average symbols across the frequency domain but consider each subcarrier individually instead. The received values from each pilot sub-carrier are initially obtained in a comb-pilot interpolation in the frequency domain and each symbol then demodulated. [29]. The four-element vector Y_p is used to choose these values. The protocol defines DFT values for known sent pilots in these subcarriers from the four-element matrix, X_p. At every pilot subcarrier the LS estimate is designed as

$$H_p(k) = \frac{Y_p(k)}{X_p(k)} \tag{8}$$

The above equation represents a four-element vector which denotes the regularly spaced channel estimations. The end points are attached to the vector to obtain the estimates as follows:

$$H_p^{|} = \left[m_{H_p}\, H_p^T\, m_{H_p}\right]T \tag{9}$$

where m_{H_p} is the mean of H_p. Instead of extrapolation from the subcarriers −21 and 21, this mean is used for the endpoints because the actual channel response at the edge frequencies cannot be resolved. For every OFDM symbol, this interpolation is done. In the time domain of sorting, a low-pass filter such as formula (7) can be used.

D. Spectral Temporal Averaging

The STA channel-estimation scheme is detailed below. From the training preamble as in equation (5), the initial estimate of the channel is first obtained. The first symbol in the packet goes through this primary estimation. After completion of this symbol demodulation a channel estimate is framed:

$$H_i(k) = \frac{Y_i(k)}{X_i(k)} \tag{10}$$

At symbol i, X_i, $Y_i(k)$, and H_i are the decided constellation point, demodulated subcarrier values, and the resulting estimate respectively. This estimation will then be determined on average over the symbol frequencies as a standard moving average so that the approximation at subcarrier λ is

$$H_{update}(\lambda) = \sum_{k=-\beta}^{\beta} w_k H_i(\lambda + k) \tag{11}$$

where β is an integer that determines the number of terms involved in the average and $\sum_{k=-\beta}^{\beta} w_k = 1$. From this averaging operation, absent subcarriers (subcarriers that do not contain data) are omitted. For example, subcarrier 26 and β = 3, the only used average subcarriers are 26, 25, 24, and 23, and the weights are corrected consequently. Since the information on the null subcarrier is not transmitted, the value of $H_i(0)$ is substituted with the average of subcarrier −1 ($H_i(-1)$) and subcarrier 1 ($H_i(1)$). For all 52 subcarriers, the frequency averaging is be performed, and subsequently, the channel estimate is restructured using the following equation:

$$H_{STA,t} = \left(1 - \frac{1}{\alpha}\right) H_{STA,t-1} + \frac{1}{\alpha}\left(H_{update}\right) \tag{12}$$

Where α defines the time-domain parameter of the moving average. $H_{STA,0}$ defines the initial estimate of the channel obtained during the estimation of the preamble. For equalizing the next symbol, the estimate is applied, and the whole packet is demodulated by iterating the process. In the present study, by observing we get the best performance by selecting the parameters α and β as 0.5 and 2, respectively. In LMS equalizer, the same value of α was used to facilitate easy comparison.

IV. Methodology

GRC is a technology framework open source that offers signal processing frames to radio applications. It can be used with a wide variety of hardware components compared to other SDR frameworks. GRC is not a fixed application-oriented environment, so it provides a solid foundation for the use of nearly any hardware components. The setup of the system used in this research is illustrated in Fig. 4.

For the transmitter and receiver attached to the GRC program, we created a SDR with N210 USRPs in real time (Fig. 5). Here we find a space indoors in which no signals of Wi-Fi are being sent, and the transmitter and the receiver are roughly 1 meter away. The SDR architecture comprises three sections of the baseband, Intermediate Frequency (IF) and Radio Frequency (RF). The RF signal is sent to the USRP, which includes the daughterboard, Analog to Digital converter (ADC)/Digital to Analog Converter (DAC), Field Programmable Gate Array (FPGA)s, Digital Signal Processing (DSP) and Application-Specific Integrated Circuit (ASIC)s by an intelligently designed antenna. For versatile baseband signal processing GRC modules are used [30].

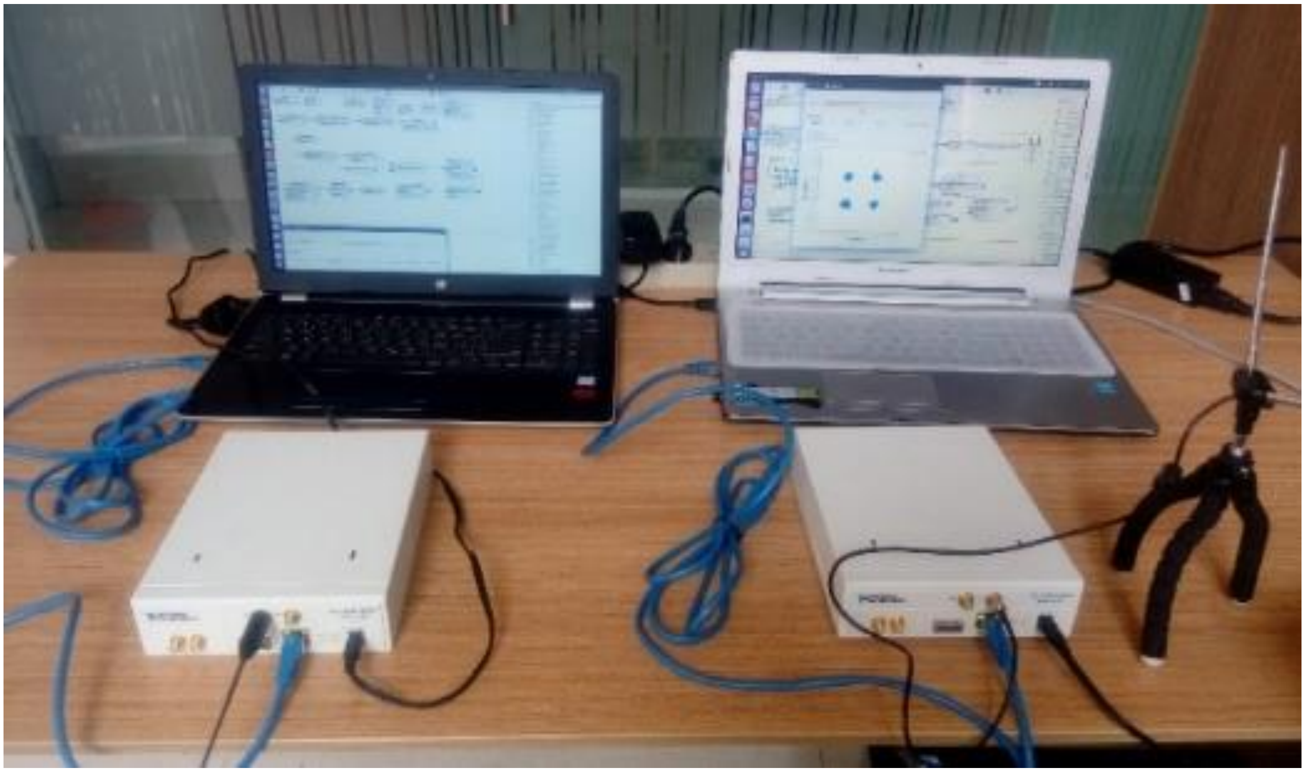


Fig. 4. Experimental Setup.

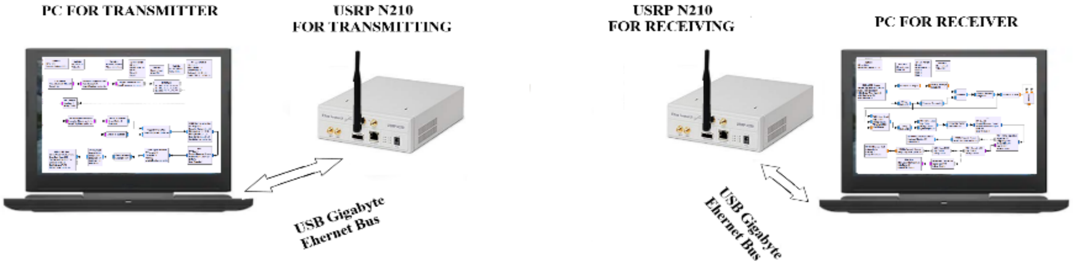


Fig. 5. Image of the Test Set-up for a Real-Time Radio System.

In order to implement the real-time radio system, two PCs were used to operate the SDR program. The used components of software and hardware are included in Table I and Table II. The PHY and MAC layers are both implemented in GRC modules, which allow us to change the layers according to specific requirements and easily analyze the results. The USRP hardware driver (UHD) is required for connecting the USRP frontends to the PCs. UHD provides common transmitting and receiving interfaces for the two USRP devices detailed in Fig. 6 and 7.

A. Transmitter

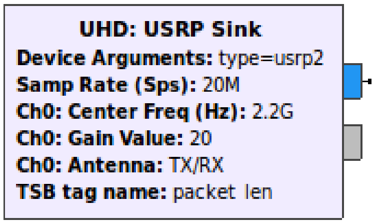
To implement an IEEE 802.11a LAN in GRC, an Out-Of-Tree (OOT) was used [29]. OOT modules are extended custom software blocks that are used for implementing application-specific functionalities. The transmitter implementation in GRC is shown in Fig. 8. Table III details the used parameters of IEEE 802.11a PHY.

TABLE. I. PC COMPONENTS USED IN THE SETUP

PC Component	Type
CPU	Intel core i7-8550U
RAM	16 GB
Operating system	16.04 LTS
Software	Ubuntu-Version 3.7
UHD Version	Version 003 011 000 000

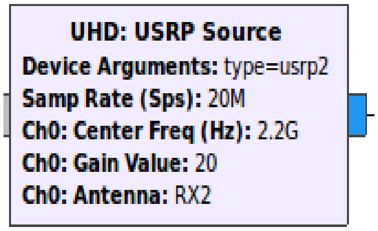
TABLE. II. HARDWARE MODULES MENTIONED IN OUR IMPLEMENTATION

Hardware Component	Type
SDR	USRP N210
Daughter Board	WBX
Transmitting Antenna	VERT 400
Receiving Antenna	Dipole Antenna
SNR	0-30 dB



UHD: USRP Sink
Device Arguments: type=usrp2
Samp Rate (Sps): 20M
Ch0: Center Freq (Hz): 2.2G
Ch0: Gain Value: 20
Ch0: Antenna: TX/RX
TSB tag name: packet len

Fig. 6. Transmitting Block in UHD.



UHD: USRP Source
Device Arguments: type=usrp2
Samp Rate (Sps): 20M
Ch0: Center Freq (Hz): 2.2G
Ch0: Gain Value: 20
Ch0: Antenna: RX2

Fig. 7. Receiving Block in UHD.

TABLE. III. PHY VARIABLES USED IN THE IMPLEMENTATION OF OFDM

Parameter	Measurement
Bandwidth	10 MHz
OFDM subcarrier	64
Subcarrier Spacing	312 KHz
OFDM Symbol time	4 µs
Guard time	1.6 µs
Comb-pilot spacing	4.4 MHz
Center frequency	2.2 GHz

B. Receiver

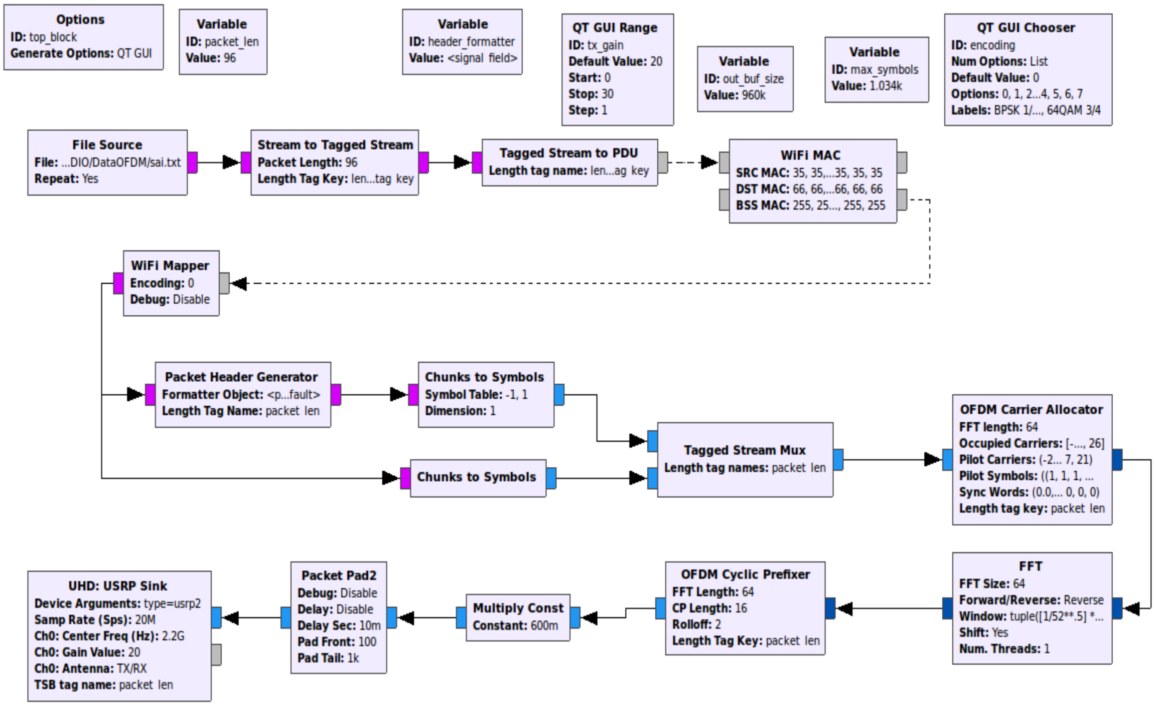
The standard IEEE 802.11 has been extended to support typical channel propagation with a latest operating model that supports instantaneous contact without setting up a previous link and An Modified Physical Layer (PHY) focused on Orthogonal Frequency-Division Multiplexing (OFDM), close to IEEE 802.11a but doubled with all timings. It switches converts the IEEE 802.11a's 20MHz signal into the new 10MHz signal for vehicular applications. The change to 10MHz, in a nutshell, renders the signal better for delay propagation but more responsive to Doppler simultaneously and channel period variations. It indicates the transition has was not clearly improved. It simply the trade-off was rebalanced and was not without any question. The receiver implementation in GRC using IEEE 802.11 OOT modules is illustrated in Fig. 9. OFDM receiver has synchronization and channel-estimation modules that help with data recovery. These modules rely on the preamble data that are appended to every frame. For this purpose, we use equations (13), (14), and (15) to detect the beginning of a frame, and the equations are related as the block diagram in Fig. 10 illustrates.

$$a[n] = \sum_{k=0}^{N_{win}+15} y_G[n+k]\overline{y_G}[n+k+16] \qquad (13)$$

$$p[n] = \sum_{k=0}^{N_{win}-1} y_G[n+k]\overline{y_G}[n+k] \qquad (14)$$

$$c(n) = \frac{|a[n]|}{p[n]} \qquad (15)$$

Equation (15) is used for detecting the beginning of the frame. The modules “WiFi Sync Short” and “WiFi Sync Long” handle frame detection, frequency offset correction, and channel estimation. While moving from “WiFi Sync Short” to “WiFi Sync Long,” only long-preamble data were transferred, and short-preamble data were detected and discarded in the “WiFi Sync Short” block. The decoded data were collected by the “Wireshark connector” module and displayed in Wireshark, which gives various information about the signal. This output is shown in Fig. 11.



Options
ID: top_block
Generate Options: QT GUI

Variable
ID: packet_len
Value: 96

Variable
ID: header_formatter
Value: <signal field>

QT GUI Range
ID: tx_gain
Default Value: 20
Start: 0
Stop: 30
Step: 1

Variable
ID: out_buf_size
Value: 960k

Variable
ID: max_symbols
Value: 1.034k

QT GUI Chooser
ID: encoding
Num Options: List
Default Value: 0
Options: 0, 1, 2...4, 5, 6, 7
Labels: BPSK 1/..., 64QAM 3/4

File Source
File: ...DIO/DataOFDM/sai.txt
Repeat: Yes

Stream to Tagged Stream
Packet Length: 96
Length Tag Key: len...tag key

Tagged Stream to PDU
Length tag name: len...ag key

WiFi MAC
SRC MAC: 35, 35,...35, 35, 35
DST MAC: 66, 66,...66, 66, 66
BSS MAC: 255, 25..., 255, 255

WiFi Mapper
Encoding: 0
Debug: Disable

Packet Header Generator
Formatter Object: <p...fault>
Length Tag Name: packet len

Chunks to Symbols
Symbol Table: -1, 1
Dimension: 1

Chunks to Symbols

Tagged Stream Mux
Length tag names: packet len

OFDM Carrier Allocator
FFT length: 64
Occupied Carriers: [-..., 26]
Pilot Carriers: (-2... 7, 21)
Pilot Symbols: ((1, 1, 1, ...
Sync Words: (0.0,... 0, 0, 0)
Length tag key: packet len

FFT
FFT Size: 64
Forward/Reverse: Reverse
Window: tuple([1/52**.5] *...
Shift: Yes
Num. Threads: 1

OFDM Cyclic Prefixer
FFT Length: 64
CP Length: 16
Rolloff: 2
Length Tag Key: packet len

Multiply Const
Constant: 600m

Packet Pad2
Debug: Disable
Delay: Disable
Delay Sec: 10m
Pad Front: 100
Pad Tail: 1k

UHD: USRP Sink
Device Arguments: type=usrp2
Samp Rate (Sps): 20M
Ch0: Center Freq (Hz): 2.2G
Ch0: Gain Value: 20
Ch0: Antenna: TX/RX
TSB tag name: packet len

Fig. 8. OFDM Transmitter using GRC.

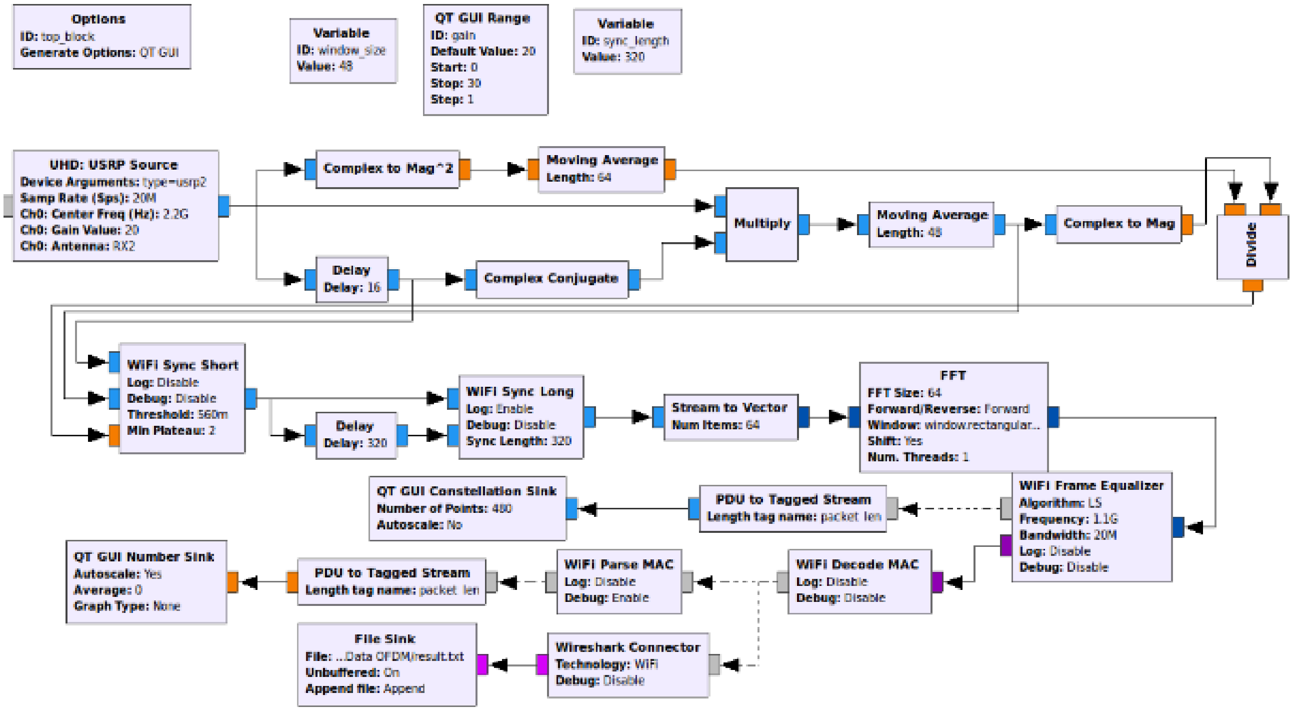


Fig. 9. OFDM Receiver using GRC.

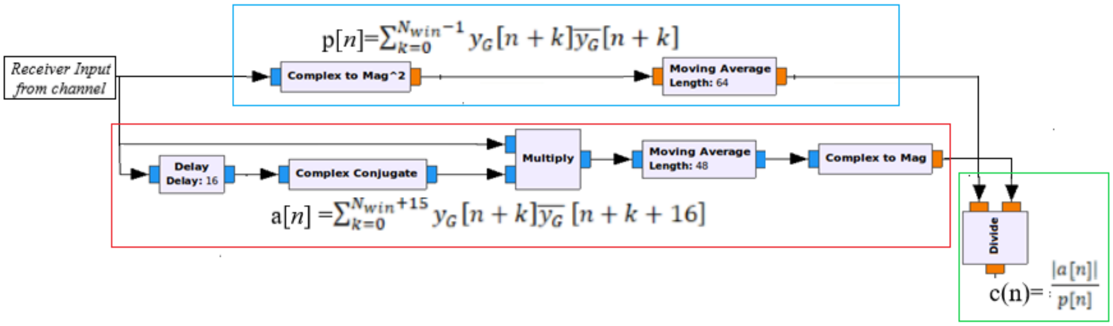
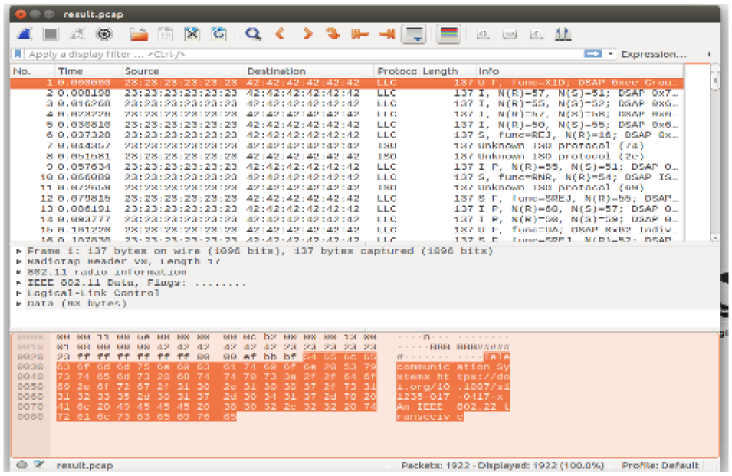

$p[n]=\sum_{k=0}^{N_{win}-1} y_G[n+k]\overline{y_G}[n+k]$
Receiver Input from channel
Complex to Mag^2
Moving Average
Length: 64
Delay
Delay: 16
Complex Conjugate
Multiply
Moving Average
Length: 48
Complex to Mag
$a[n]=\sum_{k=0}^{N_{win}+15} y_G[n+k]\overline{y_G}[n+k+16]$
Divide
$c(n)=\frac{|a[n]|}{p[n]}$

Fig. 10. Detection of Frame Starting.


result.pcap

Apply a display filter ... <Ctrl-/>

Expression...

```
No.  Time      Source             Destination        Protocol Length Info
   1 0.000000  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 U F, func=XID; DSAP 0xee Grou…
   2 0.008100  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 I, N(R)=57, N(S)=51; DSAP 0x7…
   3 0.016200  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 I, N(R)=55, N(S)=52; DSAP 0x6…
   4 0.023220  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 I, N(R)=57, N(S)=56; DSAP 0x8…
   5 0.030810  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 I, N(R)=50, N(S)=55; DSAP 0x6…
   6 0.037320  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 S, func=REJ, N(R)=16; DSAP 0x…
   7 0.044467  23:23:23:23:23:23  42:42:42:42:42:42  ISO      137 Unknown ISO protocol (74)
   8 0.051681  23:23:23:23:23:23  42:42:42:42:42:42  ISO      137 Unknown ISO protocol (2c)
   9 0.057634  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 I P, N(R)=55, N(S)=51; DSAP O…
  10 0.066000  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 S, func=RNR, N(R)=54; DSAP IS…
  11 0.072604  23:23:23:23:23:23  42:42:42:42:42:42  ISO      137 Unknown ISO protocol (60)
  12 0.079815  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 S F, func=SREJ, N(R)=55; DSAP…
  13 0.086191  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 I P, N(R)=60, N(S)=57; DSAP O…
  14 0.093777  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 I P, N(R)=58, N(S)=59; DSAP 0…
  15 0.101298  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 U F, func=UA; DSAP 0x82 Indiv…
  16 0.107830  23:23:23:23:23:23  42:42:42:42:42:42  LLC      137 S F, func=SREJ, N(R)=52; DSAP
```

```
▸ Frame 1: 137 bytes on wire (1096 bits), 137 bytes captured (1096 bits)
▸ Radiotap Header v0, Length 17
▸ 802.11 radio information
▸ IEEE 802.11 Data, Flags: ........
▸ Logical-Link Control
▸ Data (88 bytes)
```

```
0000  00 00 11 00 0a 00 00 00  00 0c b7 00 00 00 12 00   ....n... ........
0010  01 00 00 00 00 42 42 42  42 42 42 23 23 23 23 23   .....BBB BBB#####
0020  23 ff ff ff ff ff ff 00  00 af bb bf 54 65 6c 65   #....... ....Tele
0030  63 6f 6d 6d 75 6e 69 63  61 74 69 6f 6e 20 53 79   communic ation Sy
0040  73 74 65 6d 73 20 68 74  74 70 73 3a 2f 2f 64 6f   stems ht tps://do
0050  69 2e 6f 72 67 2f 31 30  2e 31 30 30 37 2f 73 31   i.org/10 .1007/s1
0060  31 32 33 35 2d 30 31 37  2d 30 34 31 37 2d 78 20   1235-017 -0417-x 
0070  41 6e 20 49 45 45 45 20  38 30 32 2e 32 32 20 74   An IEEE  802.22 t
0080  72 61 6e 73 63 65 69 76  65                        ransceiv e
```

result.pcap   Packets: 1922 · Displayed: 1922 (100.0%)   Profile: Default

Fig. 11. Wireshark Connector Output.

V. Results and Discussion

We tested four different channel-estimation techniques in a real-time implementation of the IEEE 802.11a standard [31]. This section gives constellation plots as measured from the various modulation schemes at different transmitting powers, along with the FDR results. We highlight the systems applicability for real-time environment by determining the frame delivery ratio. This gives the percentage of successfully delivered frames by transmitted frames. In this comparison test between different channel estimations we considered 100 frames per run for each transmitting power from 0dB to 30dB. After testing different system factors (i.e., α, and β, as applicable) we select the best performing scheme and the results are generated from that system.

We trust that on-road situations can be reflected by this evaluation procedure. This structural proof uses genuine experimental data that are gathered from field measurements. The accuracy of the channel models is undoubted in this process. This method permits us to compare various equalization methods. This approach is faithful, easy to use, and repeatable.

We extended our tests by comparing different modulation schemes using their constellation plots at the receiver. Here we see the pilot information as the learning signals that enable us to measure the changes in the frequency-domain channel for a given test package in each subcarrier. We can generate the 802.11a waveforms using the stored frequency-time channel response. The performance of 802.11a at each 100 frames is done using this method. Results of this study are shown below.

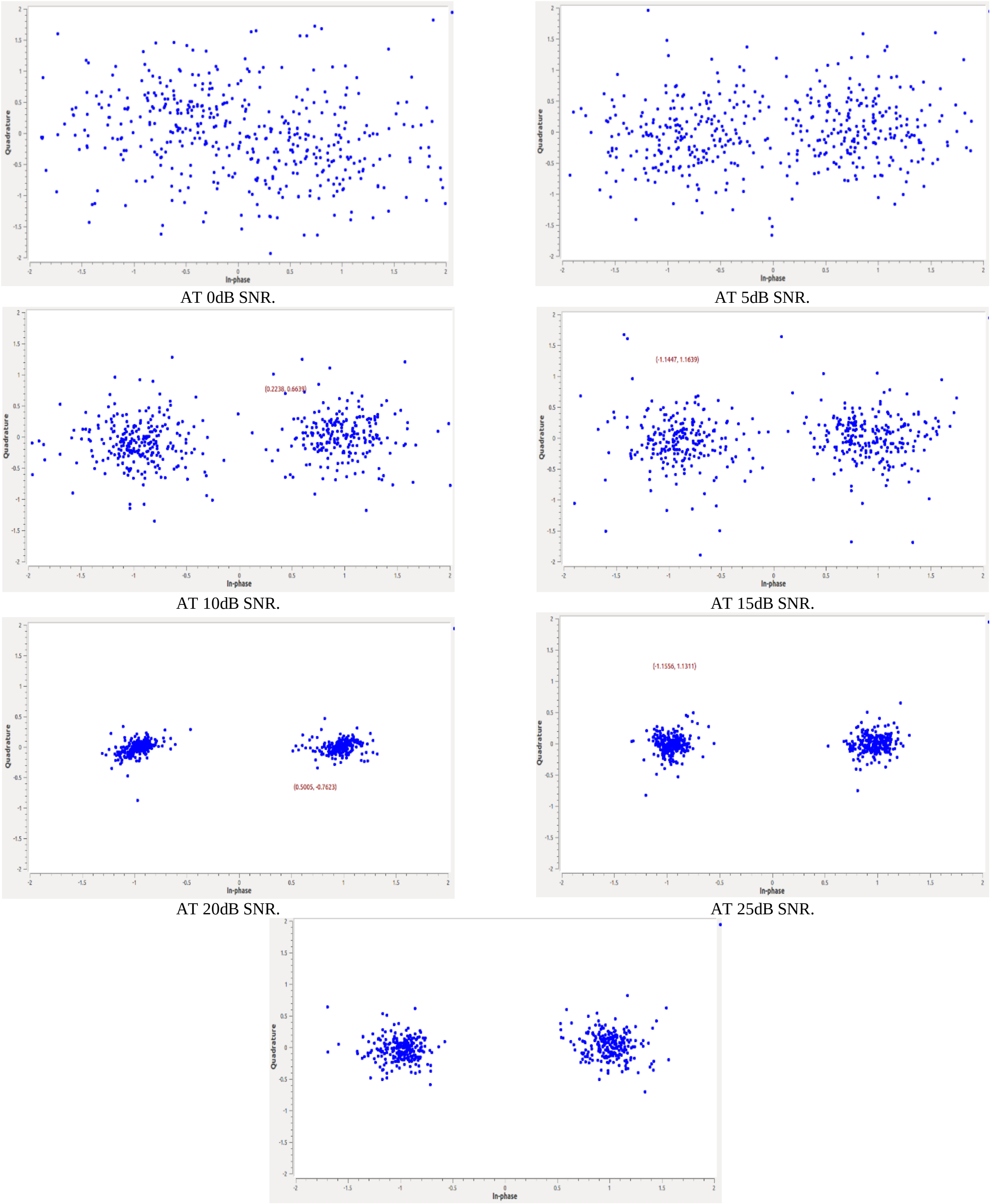
A. Comparison of Frame Delivery Ratio (FDR) for different Modulation Schemes

Simulations show that noise and interference in network simulators could be treated similarly. To provide evidence for these findings through measurements, we installed radio transceivers in our laboratory. Below, the FDR results from all four modulation schemes are plotted for several SNRs in dB. Constellation plots and FDR results for different modulations and different coding rates in the real-time testbed environment are shown in Fig. 12, Fig. 13, Fig. 14, Fig. 15, and Fig. 16 and Fig. 17 respectively. Our results show that technological considerations such as system parameters have a lesser impact on the signal than the transmission of multipaths. Under practical propagation conditions, such a learning environment allows us to determine the importance of the evaluated variables to a localization process. We demonstrate that propagation conditions vary even in LOS conditions in each SNR and that the radio signals undergo distinct propagation conditions in specific SNRs. When we found in each situation in the constellation diagrams, the received symbols form into clusters, each of which has its ideal point when the SNR is at 30dB but at 0dB, the received symbols scatter randomly over the constellation diagram. Therefore, the optimal points for the obtained signs are difficult to identify, and demodulation errors may occur frequently. These plots show that the BPSK modulation performs the best, irrespective of the encoding rate. All the results shown in this section were recorded when using the STA channel-estimation technique and the Schmidl-cox synchronization technique. Here we considered 100 frames per run for each transmitting power. Observing these results, we can conclude that for typical channel conditions we can get best results at BPSK.

B. Comparison of FDR for different Channel-Estimation Techniques

To compare the channel-estimation techniques in real-time transmissions, data were sent repeatedly in the form of frames, and the FDR was observed at the receiver using the ratio of received frames and transmitted frames. Among the four channel estimation methods we tried, these outcomes show that the STA channel estimator offers the best execution. The difference in the algorithms' performance is small in general, but under dynamic channel conditions, STA will perform the best. In this case the receiver and transmitter are in Line of Sight (LOS). But fading may happen even there is a presence of a LOS due to the reflection, scattering etc. of the transmitted signal from the ground and surrounding area objects. The receiving antenna receives a signal which depends on the frequency and bandwidth of the transmission signal propagation, and which can vary widely either in amplitude or even phase. This estimation technique is generally used in IEEE 802.11p protocols, since with vehicle-to-vehicle communications, the relative positions of the transmitter and receiver can change very quickly, which leads to Doppler spread and delay that can exacerbate fading effects. Under such channel conditions, a DDCE-based adaptive channel equalizer must be used.

In Fig. 18, one can see the difference between the constellation's plots measured with different channel-estimation techniques in our real-time environment. These four constellation plots were measured with transmissions using BPSK (1/2) modulation, transmitting power of 30 dB, and receiving power of 20dB. Fig. 19 shows FDR plots of data transmitted using the four channel-estimation techniques. These data also show that the STA channel estimator offers the best FDR at low SNR and that it reaches the highest FDR of all channel estimators that we tested.



AT 0dB SNR.

AT 5dB SNR.

AT 10dB SNR.

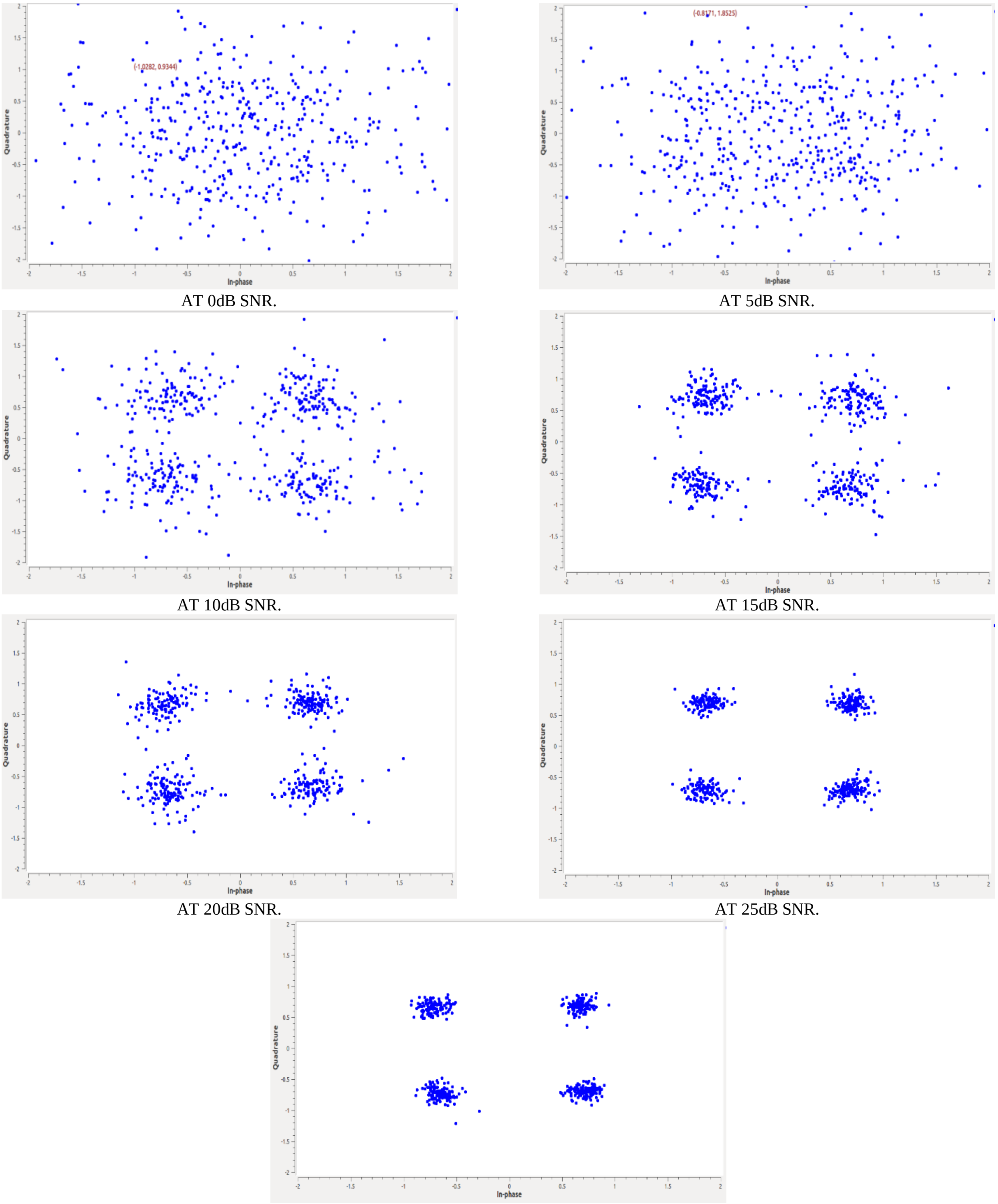
AT 15dB SNR.

AT 20dB SNR.

AT 25dB SNR.

AT 30dB SNR.

Fig. 12. Constellation Plots for BPSK of SNR from 0 to 30 dB.



AT 0dB SNR.

AT 5dB SNR.

AT 10dB SNR.

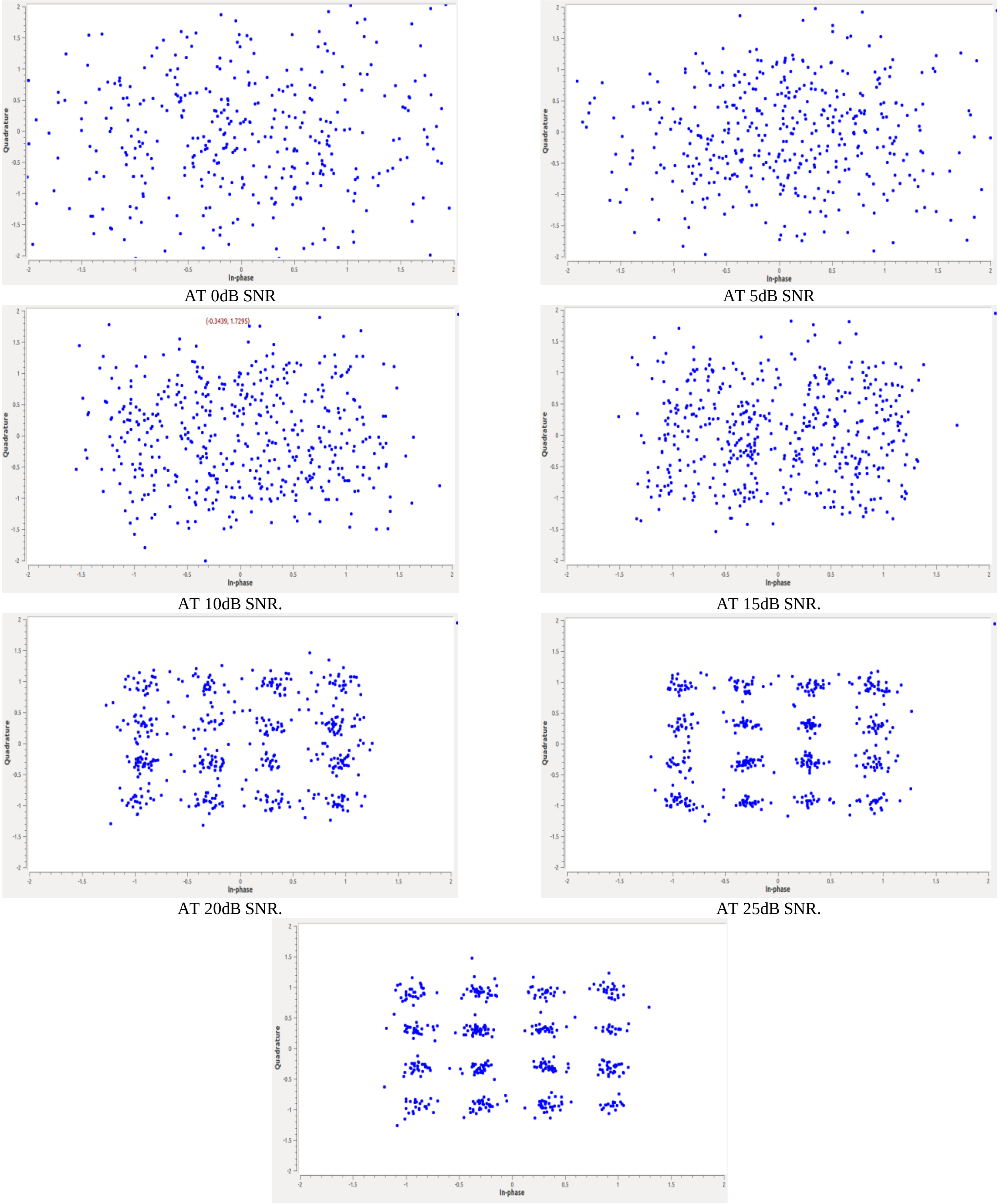
AT 15dB SNR.

AT 20dB SNR.

AT 25dB SNR.

AT 30dB SNR.

Fig. 13. Constellation Plots for QPSK of SNR from 0 to 30 dB.



AT 0dB SNR

AT 5dB SNR

AT 10dB SNR.

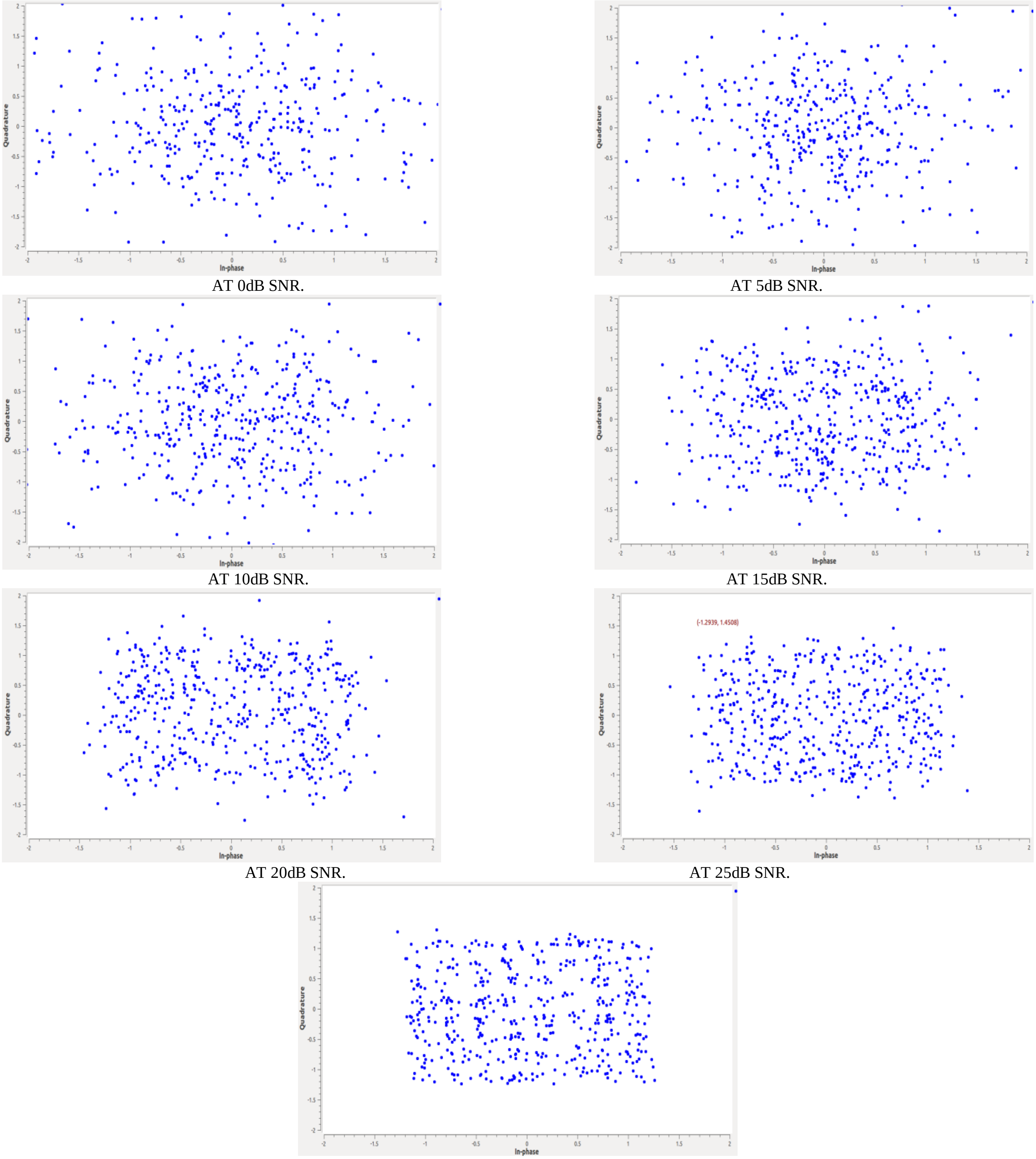
AT 15dB SNR.

AT 20dB SNR.

AT 25dB SNR.

AT 30dB SNR.

Fig. 14. Constellation Plots for 16QAM of SNR from 0 to 30 dB.



AT 0dB SNR.

AT 5dB SNR.

AT 10dB SNR.

AT 15dB SNR.

AT 20dB SNR.

AT 25dB SNR.

AT 30dB SNR.

Fig. 15. Constellation Plots for 64QAM of SNR from 0 to 30 dB .

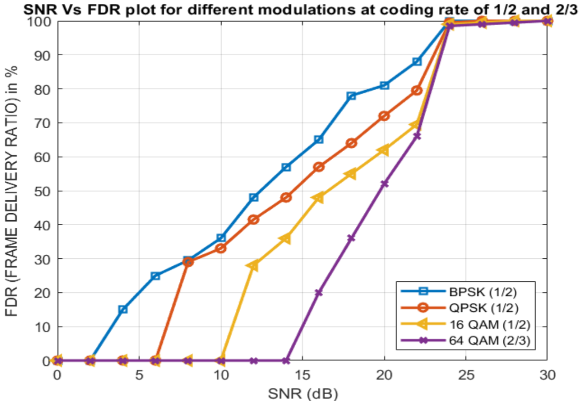


Fig. 16. SNR vs FDR for BPSK (1/2), QPSK (1/2), 16 QAM (1/2), and 64 QAM (2/3) Coding rate.

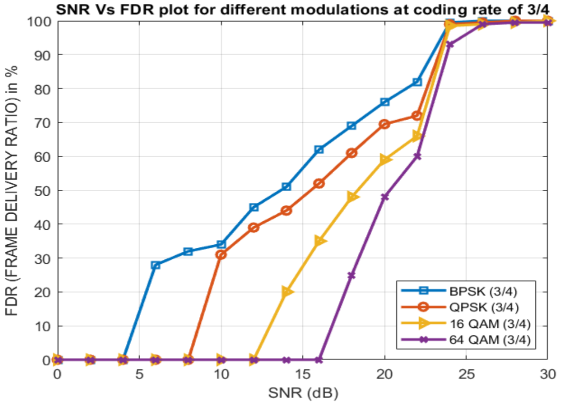
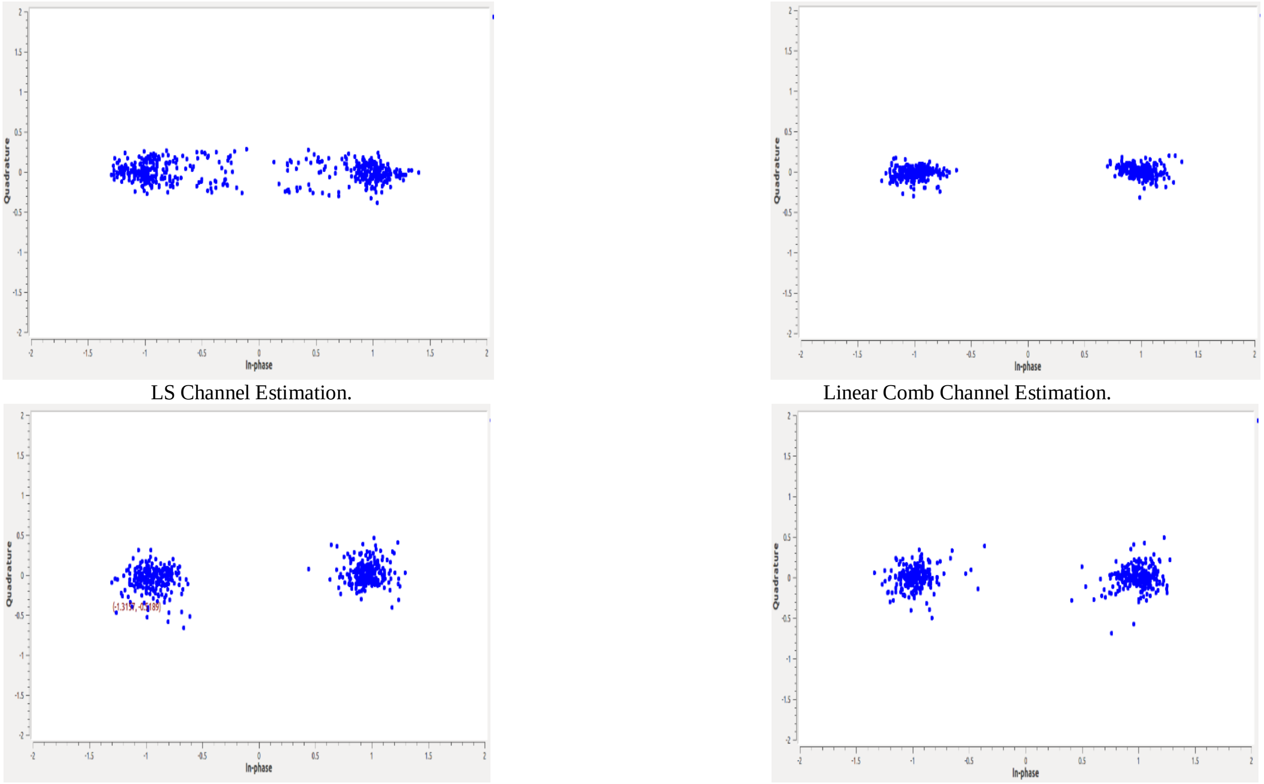


Fig. 17. SNR vs FDR for BPSK, QPSK, 16 QAM, and 64 QAM at ¾ Coding rate.



LS Channel Estimation.

Linear Comb Channel Estimation.

LMS Channel Estimation.

STA Channel Estimation.

Fig. 18. Constellation Plots for different Channel-Estimation Techniques.

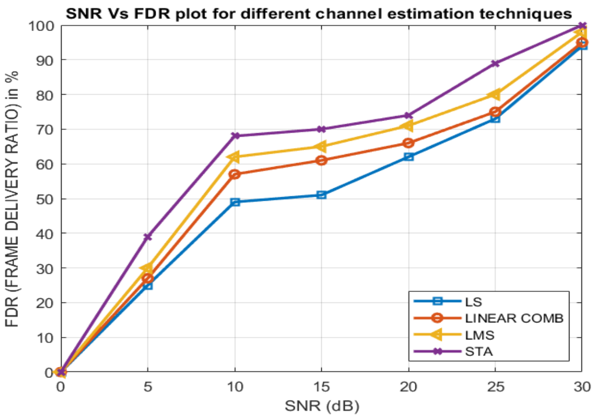


Fig. 19. SNR vs FDR for different Channel-Estimation Techniques.

VI. Conclusion

Short coherence time and narrow coherence bandwidth degrade the performance of the physical layer in typical channels. Preamble-based equalization is a traditional scheme, which cannot reimburse for these channel impacts. Data must be used to update the estimates of the channel because preamble-based standards do not sufficiently provide pilot-signal feedback. Thus, improvements in wireless communication system performance depend on channel-estimation techniques. This paper reports tests of channel-estimation techniques in a real-time environment that measured the FDR and proved that the system can provide over 90% of delivery ratio at 25 db of SNR for different digital modulation techniques using STA. Tests were performed with an implementation of the IEEE 802.11a standard protocol using the open-source GRC software to construct a novel SDR testbed that can be used with a wide variety of frontend hardware. Two N210 USRPs were used as frontend transmitter and receiver in our tests. This testbed offers a pathway for the investigation of various parameters in real time, and we performed a series of tests to validate the usefulness of the STA channel estimator in 802.11a networks. The results clearly show that STA outperforms other schemes, with the FDR clearly higher than the other four estimators that we tested.

References

[1] Albreem MAM. “5G wireless communication systems: vision and challenges”. I4C, pp. 493–97, April 2015.

[2] Lee J, Ryu J, Lee S-J, Kwon T. “Improved modeling of IEEE 802.11a PHY through fine-grained measurements”, Computer Networks: The International Journal of Computer and Telecommunications Networking, Vol. 64, No.4, pp.641–5, March 2010.

[3] Giddings RP, Jin XQ, Tang JM. “First experimental demonstration of 6Gb/s real time optical OFDM transceivers incorporating channel estimation and variable power loading”. Optics Express, Vol.17, No.22, pp. 19727-19738, 2009.

[4] Jin XQ, Giddings RP, Tang JM. “Real-time transmission of 3Gb/s 16-QAM encoded optical OFDM signals over 75km SMFs with negative power penalties”. Optical Express, Vol.17, No.17, pp. 14574-14585, 2009.

[5] Li R, Dou Y, Zhou J, Deng L, Wang S. CuSora: “Real-time software radio using multi-core graphics processing unit”. Journal of Systems Architecture, Vol. 60, No.3, pp. 280-292, March 2014.

[6] Armour S, Nix A, Bul D. “Complexity evaluation for the implementation of a Pre-FFT Equalizer in an OFDM receiver”. IEEE Transactions on Consumer Electronics Vol. 46, No.3, pp. 428-437, June 2000.

[7] Pozza M, Rao A, Flinck H, Tarkoma S. “Network-in-a-box: a survey about on-demand flexible networks”. IEEE Communications Surveys & Tutorials, Vol. 20, No.3, pp. 2407-2428, February 2018.

[8] W. H. Tuttlebee, Software Defined Radio: Enabling Technologies. John Wiley & Sons, 2003.

[9] Luo H, Wu K, Ruby R, Liang Y, Guo Z, Ni LM. “Software-defined architectures and technologies for underwater wireless sensor networks: a survey”. IEEE Communications Surveys & Tutorials, Vol. 20, No.4, pp. 2855-2888, May 2018.

[10] Cai X, Zhou M, Xia T, Fong WH, Lee W-T, Huang X. “Low-power SDR design on an FPGA for intersatellite communications”. IEEE Transactions on Very Large-Scale Integration (VLSI) System, Vol. 26, No.11, pp. 2419–2430, July 2018.

[11] Haithem Haggui, Sofiène Affes and Faouzi Bellili. “FPGA-SDR Integration and Experimental Validation of a Joint DA ML SNR and Doppler Spread Estimator for 5G Cognitive Transceivers”, IEEE Access, Vol. 7, pp. 69464-69480, May 2019.

[12] Per Zetterberg and Ramin Fardi, “Open Source SDR Frontend and Measurements for 60-GHz Wireless Experimentation”, IEEE Access, Vol. 3, pp. 445-456, May 2015.

[13] Xintong Lu, Luyao Ni, Shi Jin, Chao-Kai Wen and Wen-Jun Lu, “SDR Implementation of a Real-Time Testbed for Future Multi-Antenna Smartphone Applications”, IEEE Access, Vol. 5, pp. 19761-19772, October 2017.

[14] Muhammad Bilal Khan, Xiaodong Yang, Aifeng Ren, Mohammed Ali Mohammed Al-Hababi, Nan Zhao, Lei Guan, Dou Fan and Syed Aziz Shah, “Design of Software Defined Radios Based Platform for Activity Recognition”, IEEE Access, Vol.7, pp. 31083-31088, March 2019.

[15] Christos Politis, Sina Maleki, Juan Merlano Duncan, Jevgenij Krivochiza, Symeon Chatzinotas and Björn Ottesten, “SDR Implementation of a Testbed for Real-Time Interference Detection with Signal Cancellation”, IEEE Access, Vol.6, pp. 20807-20281, May 2018.

[16] Leonardo A. Agüero Guzmán, Elias M. Ovalle, and Rodrigo A. Reeves, “Measurement of the ionospheric reflection height of an HF wave in vertical incidence with a resolution of minutes”, IEEE Geoscience And Remote Sensing Letters, Vol. 15, No.11, pp. 1637–1641, November 2018.

[17] Wei Wang, Yinejie Chen, Zeyu Wang, Jin Zhang, Kaishun Wu, and Qian Zhang, “Wideband Spectrum Adaptation Without Coordination” IEEE Transactions on Mobile Computing, Vol. 16, No.1, pp. 243-256, January 2017.

[18] Wang W, Chen Y, Wang L, Zhang Q, “Sample less Wi-Fi: bringing low power to Wi-Fi communications”, IEEE/ACM Transactions on Networking, Vol. 25, No.3, pp. 1663-1672, June 2017.

[19] Suman Bhunia, Edward Miles, Shamik Sengupta, and Felisa Vazquez-Abad, “CR-Honeynet: A Cognitive Radio Learning and Decoy Based Sustenance Mechanism to Avoid Intelligent Jammer”, IEEE Transactions on Cognitive Communications and Networking, Vol. 4, No.3, pp. 567-581, September 2018.

[20] Song Fang, Yao Liu, and Peng Ning, “Wireless Communications under Broadband Reactive Jamming Attacks”, IEEE Transactions on Dependable and Secure Computing, Vol. 13, No.3, pp. 394-408, June 2016.

[21] Cassola A, Jin T, Noubir G, Thapa B “Efficient spread spectrum communication without pre-shared secrets”, IEEE Transactions on Mobile Computing, Vol. 12, No.8, pp. 1669-1680, 2013.

[22] Vilches T, Dujovne D, “GNURadio and 802.11: performance evaluation and limitations”, IEEE Network, Vol. 28, No.5, pp. 27-31, October 2014.

[23] Horrein P-H,·Hennebert C, Pétrot F, “Integration of GPU computing in a software radio environment”, Journal of Signal Processing Systems, Vol. 69, No.1, pp. 55-65, October 2012.

[24] Lee S, Choi J, Yoo J, Kim C-K, “Frequency diversity-aware Wi-Fi using OFDM-based bloom filters”, IEEE Transactions On Mobile Computing Vol. 14, No.3, pp. 525-537, March 2015.

[25] Ponnaluru Sowjanya, Penke Satyanarayana, "Real-Time Data Transfer Based on Software Defined Radio Technique using Gnu radio/USRP", International Journal of Engineering and Advanced Technology, Vol. 9, No. 1, pp. 279-288, October 2019.

[26] Fernandez JA, Stancil DD, Bai F, "Dynamic Channel Equalization for IEEE 802.11p Waveforms in the Vehicle-to-Vehicle Channel". In 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IEEE, 542-551 2010.

[27] Kim H. Wireless Communications Systems Design. John Wiley & Sons; 2015.

[28] Van Nee R, Prasad R. OFDM wireless multimedia communications. Boston: Artech House; 2000.

[29] Bloessl B, Segata M, Sommer C, Dressler F, "Performance assessment of IEEE 802.11p with an open-source SDR-based prototype", IEEE Transactions on Mobile Computing, Vol. 17, No.5, pp. 1162-1175, May 2018.

[30] Ozdemir MK, Arslan H. "Channel estimation for wireless OFDM systems", IEEE Communications Surveys and Tutorials, Vol. 9, No.2, pp. 18-48, June 2007.

[31] IEEE, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," IEEE, STD 802.11-2016, Dec. 2016.

AUTHOR'S PROFILE

Ponnaluru Sowjanya completed her B.Tech (Electronics and Communication Engineering) in Sri Mittaplli Institute of Technology for Women, Tummalapalem, Guntur, India; M.Tech (Embedded Systems) in Nalanda Institute of Engineering and Technology, Kantepudi, Satenepalli, India. At present, she is pursuing a Ph.D. in Electronics and Communication Engineering at Koneru Lakshmaiah Educational Foundation, Vaddeswaram, India. Her research interest includes Wireless Communication and Signal processing.

Dr. Penke Satyanarayana received a B.Tech in Electronics and Communication Engineering (ECE) from Koneru Lakshmaiah College of Engineering, Vaddeswaram, Vijayawada, India; M.Tech in Computers and Communications Systems from JNTU, Hyderabad, India; and Ph.D. in ECE from JNTU, Kakinada, India. He is working as a Professor in the Department of ECE, K L E F, Guntur, India. He has published over 20 research papers in reputed international and national journals and conferences. He is a member in Indian Society of Technical Education (ISTE). His research interests include Wireless Communication and Signal processing, Embedded Systems and VLSI.

Measures of Organizational Training in the Capability Maturity Model Integration (CMMI)

Mahmoud Khraiwesh
Faculty of Information Technology
Zarqa University, ZU
Zarqa, Jordan

***Abstract*—Training has a major impact on organizational commitment. Organizational objectives can be met by executing several training strategies and programs to enhance training. Organizational training aimed at developing employees' knowledge and skills. It shall enable employees to carry out their duties efficiently and effectively. Two goals and seven practices of the organizational training process area in the capability maturity model integration (CMMI) framework are analyzed through this study. That's done to set common measures for organizational training. CMMI is a framework for assessing and improving software systems. The researcher implemented the Goal Questions Metrics (GQM) model on two objectives and seven specific practices of the organizational training process area in CMMI. That's done to set measures.The researcher confirmed that the defined measures are the true measure for each one of the seven specific practices.**

***Keywords*—Organizational training; training; measures; CMMI; GQM**

I. Introduction

Improving human resources in any organization must be done by HR management [1].Organizational training (OT) aims to develop employees' knowledge and skills of the employee. It aims at enabling employees to carry out their functions efficiently and effectively. Itaims at enabling employees to meet the organization's business goals and satisfy the tactical training requirements [2].Measures accompanied with the first specific CMMI is a framework for assessing and improving software projects that are developed by the Software Engineering Institute (SEI) in Carnegie Mellon University in the United States. Goal Questions Metrics (GQM) paradigm was applied in the organizational training process area in CMMI. It was applied for defining the measures of specific goals and its specific practices.

Organizational training controls the training providedfor meetingthe organization's strategic business goals and the tactical training demands. It shall provide support to the teams responsible for carrying out the project.Training needs are identified by the support teams. Individual projectsare carried out to reach specific goals [2]. The main elements of the training process include written plans, a controlled training development program, and mechanisms for assessing the efficiency of the training elements. The main elements of the training process also include a team with an appropriate ability of disciplines and other areas of knowledge. Carrying out effective training requires identifying needs, having an instructional design, andproper training media and making plans. Naming thetraining requirements is based on the abilities and experiences required for performing the standard processes of the organization.Achieving success in training is manifested through the availability of chances for obtaining the knowledge and skills that are required to implement continuous and new activities.

Knowledge and skills can be organizational, contextual or technical. Contextual skills refer to the self-control, and contact capabilities and capabilities to hold interpersonal relations. Such skills are required for carrying out the work successfullyin the organization and the project. Having organizational skills shall positively affect the roles and responsibilities that are handled.Technical skills refer to the capabilities of using materials, data, equipment, tools and processes which are needed for carrying out a project or a process[2].

The organizational objectivescan be achieved byexecuting the strategies andtraining programs. That shall enhance the training process and facilitate it [3]. Training has an important impact on organizational promises [4]. Cheahet al. found that trainingcan significantly affect organizational commitment [5]. Achieving competitive advantage for organizations depends mainly on the trained human resource. Therefore, there is a need for observing the process of training employees and empowering them [1]. Job training and satisfactions shall affect the extent of having internal cooperation in the organization [6].

Measures are carried outfor carrying out processes or producing products [7]. Carrying out measures help us in understanding, controlling, and assessing processes and products [8]. It shall enable organizations to achieve the objectives sought from developing software [9] and [10].

It shall participate in managing the project [11]. It shall participate in assessing processes and products [12]. For developing a software process, there is a need to define the relevant attributes of the process. After that, there is a need to set the relevant metrics of these attributes.Measurement practices perform a significant role in managing and recognizing processes and products in computer information systems [13].

Through carrying out software measures, we assign numbers or symbols to the attributes of the selected objects in the actual life. That's done for determining the attributes by some defined rules [14]. Measures are important for project process understanding, managing, and change [15]. Measurement is carried out to assess the extent of accomplishing goals and supporting the improvement process [16] and [17].

We do notbelieve that we can define the reliable quantitative rules with having the same consistency and accuracy as physics’ rules [18]. Precise metrics –like velocity and voltage- are rare in the software measures area. Software measurements are usually subject to argument [19]. There is a need for establishing a structure for effectiveness metrics [20].

A study was conducted by the Software Engineering Institute (SEI) in the mid-1980s about the capability of software developers. The study led to the development of a new framework which is a software capability maturity model for software (CMM/SW) [21]. The (CMMI) framework was introduced through the collaboration of the Software Engineering Institute (SEI) with several software companies. CMMI is a comprehensive structure. It’s employed much for improving the processes of projects. That’s done for producingproducts of high quality [22].

To ensure organizational survival and developing an effective strategy, it’simportant to adopt a CMMI model. This model will improve and assess the quality of the software [23]. Adopting a CMMI model will enable the organization to improve the project management and quality of the product. It will enable the organization to increase its production level. It will enable the organization to setthe project schedule and budget efficiently [24] and [25].

CMMI seeks to describeseveral software development process areas (PA). After the appraisal, CMMI determines the satisfaction of a process area. CMMI defines its activities and practices in a well-structured and easy manner[26]. CMMI includes all the activities that must be done for developing the software. It influences many products and processes in the life cycle of the software[27]. CMMI develops. It becomes a well-known, accepted standard of software process enhancement. It sheds light on several details [28].

Every process area in the CMMI framework has general goals. It is accompanied with general practices and specific goals and practices. A specific goal seeks to describe the things that must be done in the process area. Specific practice is considered as a central action in order for the concerned specific goal to be met. Specific practices in CMMI are considered goals[29] and [30]. Look at Fig. 1.

Goal-Question-Metric (GQM) is a manner for specifying the metrics. Basiliand Weiss [18] developed GQM in order to identify effective metrics forsoftware development activities.

This paper identifies measures for seven specific practices that are related to the two specific goals in the organizational training process area in CMMI-SW (Staged Representation) model within level 3. That is shown through Fig. 1. The Goal-Question-Metrics method was applied to the seven specific practices which are accompanied with 2 objectives in the organizational training process. Through applying this method, the researcher defined a set of measures.

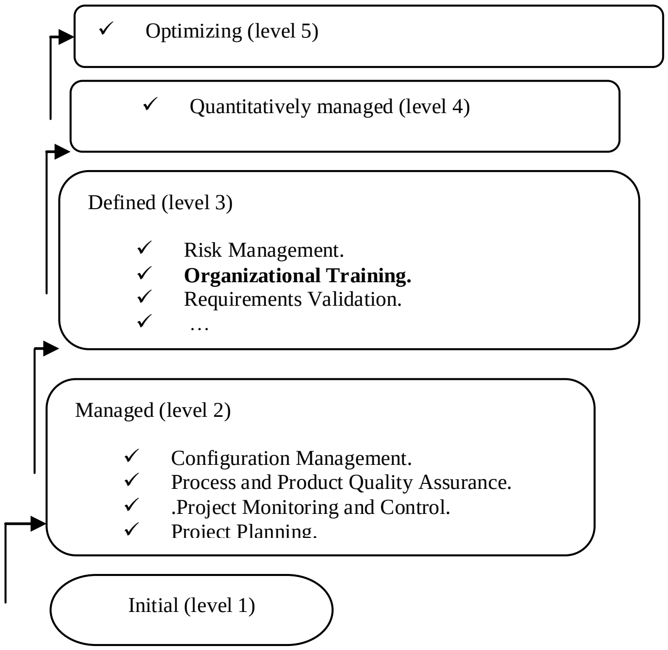

✓ Optimizing (level 5)
✓ Quantitatively managed (level 4)
Defined (level 3)
✓ Risk Management.
✓ **Organizational Training.**
✓ Requirements Validation.
✓ …
Managed (level 2)
✓ Configuration Management.
✓ Process and Product Quality Assurance.
✓ .Project Monitoring and Control.
✓ Project Planning.
Initial (level 1)

Fig. 1. The Five Maturity Levels in CMMI with Related Process Areas [2].

The researcher controlled and assessed processes and products in organizational training. That was done byusing a set consisting of defined measures.The next section presents the relevant work of the two frameworks CMMI and CMM.The fourth and third sections present a summary for the CMMI/SW and GQM respectively.The fifth section presents the implementation of the GQM to the seven specific practices that are accompanied with the two specific goals in the organizational training. That resulted in the definition of measures. The sixth section presents proof for the reliability and the validity of the defined measures. Sectionsevenpresents the study’s conclusion.

II. The Relevant Works

Various researches aimed to measure the software process that has been suggested, such as [21, 31-41]. The works that are related the most to our work are represented in the works of [21, 31, 32, 35-41]. Baumertand McWhinney[42] point out to several observations that are usefulfor measuring the practices (general characteristics) that are defined in the Capability Maturity Model for CMM/SW. The specified observations include 13 classes. These classes do not occur at all maturity levels. The work of Baumert and McWhinneydoesn’t focus on a particular process. It shed a light on CMM instead of CMMI.

Several researchers studied measures of key process areas related to CMM/SW, but not to CMMI. Such researchers include Paulk [21] and Loconsole [13]. Paulkand others named several patterns that are related to a particular KPA. Naming such patterns is a Requirements Management. Their effortsfocus on a definite process (i.e. CMM/SW)rather than focusing on (CMMI/SW). Loconsole [13] specified several measures for a key process area in CMM/SW which is Requirements Management. Her work is concerned with

(CMM/SW), rather than (CMMI/SW). Khraiwesh, [31, 35-39, 41] identifies a set of measures for processessuch as validation, process and product quality assurance, risk management, project planning, project monitoring and control, and configuration management in CMMI/SW.

The present study provides a description of a set of general measures for the organizational training process in CMMI/SW. The measures we chose are related to the seven specific practices which are accompanied by 2 goals in the organizational training process.

III. Overview ofthe CMMI-SW

There are five maturity levels in the CMMI/SW (Staged Representation): Initial, managed, defined, quantitatively managed and optimizing levels. There are several process areas in each level of the 5 levels, except for the first level [2].

A process area has specific goal/ goals with related specific practices. It also has general goals with related general practices. Regarding each specific goal, it controls several features. That's done for performing the process area. As for each practice, it represents an action that must be carried out to meet the associated specific goal [2]. A generic goal holds proper institutionalization exercises. That's because each goal is accompanied with different processes named (generic). That's seen through Fig. 2.

The researcher presented below 2goals. He presented practices that are related to each specific goal in the organizational training process:

1). Establish an organizational training capability:-

a) Establish strategic training needs.

b) 1.2Identify which the organization must meet training needs.

c) 1.3 Create an organizational training tactical plan.

d) 1.4 Establish a training capability.

2). Provide training:-

a) 2.1 Deliver training.

b) 2.2 Create training records.

c) 2.3 Assess the effectiveness of the training.

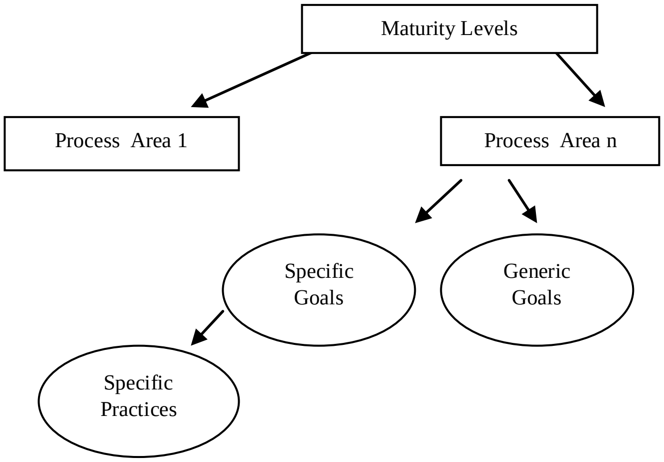


Fig. 2. Specific and Generic Goals.

IV. Overview about the GQM

Organizations use the Goal/Question/Metric (GQM) paradigm for concentrating on the measures of their objects. Any organization should identify the goals before collecting data when employing the GQM model [12]. After defining goals, the quantifiable questionsrelated to each goalmust be defined. After that, a set of measures associated with each questionmust be defined. After that, data should be collected to meet the goals.

The three steps in the GQM model are listed below:

1). Defining a set of objects that are linked to the requirements in the organization's projects [42].

2). Generating a set of quantifiable questions that are accompanied with each defined object. Various sets of instructions must be used to analyze the questions that were defined by Basili and Rombach [12].

3). Generating a set of measures which are fit to the quantifiable questions. That must be done in order for the generated measures to present information. That shall provide answers to the quantifiable questions. Many measures perhaps specified for one goal. A measure may refer to various questions.

V. Applying GQM to the CMMI-SW

The organizational training process in CMMI/SW has 2 specific goals and 7 related specific practices. In the present study, the researcher treated specific practices as being goals. These goals shall be met through the GQM model.

The organizational training process has the following seven specific practices:

1). Establish strategic training needs: Building and managing the main training requirements of the organization.

2). Identity which training needs are the responsibilities of the organization: Deciding which training requirements must be met by the organization and determining which requirements should be met through the individual project.

3). Establish an organizational training tactical plan: Building and managing the tactical plan of the organizational training needs.

4). Establish a training capability: Building and maintaining a training ability to mark the organizational training needs.

5). Deliver training: Releasing training after establish the organizational training tactical plan.

6). Create training records:Building and managing the documents of organizational training.

7). Assess training effectiveness: Assessing the impact of the training program of the organization.

As it ismentioned earlier, the seven specific practices shall be treated as goals. Following the GQM process, a set of quantifiable questions shall be produced. After that,a definition shall be provided for a set of measures that provide the required quantitative data for satisfying the produced questions. The work products and the sub-practices thatare proposed in each of the seven specific practices shall be taken into account when the measures are determined.

The tables below (i.e. Tables I-VII) describe a set of measures and questions. Every table describes a specific practice (i.e. a goal). The questions and measures are interfering. Each measure can offer information for clarifying several questions.

A. *Measures for Specific Practice 1*

Establish strategic training needs:Establish and maintain the strategic training needs of the organization.

Table I presents several measures and questions. Suchmeasures and questions are accompanied with the 1st practice inthe OT.

B. *Measures for Specific Practice 2*

Identifying which training needs must be met by the organization:Identifying which training needs must be met by the organization and identifying which training needs must be met by the individual.

Table II presents measures and questions. Thesemeasures and questionsaccompanied with the 2nd practice in OT.

C. *Measures for Specific Practice 3*

Establish organizational training tactical plan:Establish and maintain an organizational training tactical plan.

Table III presents measures and questions. Thesemeasures and questions are accompanied with the 3rd practice in the OT.

TABLE. I. MEASURES AND QUESTIONS THAT ARE ACCOMPANIED WITH THE 1ST PRACTICE IN OT

	Questions	**Measures**
Q1	Do you classify the training needed for improving the skills needed for carrying out the project activities?	• Classify the training needed for improvingskills. • # Training needed classifications. • # Needed skills. (# means the number of)
Q2	Do you identify the training needed for developing the information that's required for carrying out the project activities?	• Identifying the training needed for developing the information required to carry out the project activities. • # Training needed classifications. • # Project activities.
Q3	Do you build strategic training requirements? (Strategic training requirements write long-term goals to develop skills, normally from two to five years).	• Building strategic training requirements. • # Long term goals. • # Years.
Q4	Do you define the sources of the strategic training requirements? (Some of the sources include the standard processes of the organization, risk analysisis,the improvement plan of the organization, the strategic business plan of organizations, and skill assessments).	• Providing a definition for the sources of the strategic training requirements. • # Sources ofstrategic training requirements.
Q5	Do you analyze the strategic business plan of the organization and process improvement plan of the organization to know possible training needs?	• Analyzing the strategic business plan and process improvement plan of the organization. • # Training needed classifications. • # Skills that are needed.
Q6	Do you document the strategic training requirements of the organization?	• Documenting the strategic training requirements of the organization. • # The requirements ofstrategic training.
Q7	Do you classify the strategic training requirements of the organization? (Some examples of categories are: requirements analysis, quality engineering, testingteam building, disaster recovery, leadership,analysis and documentation,and negotiation skills).	• Classifying the strategic training requirements of the organization. • # The requirements of strategic training.
Q8	Do you define the skills and the roles needed for performing the standard processes of the organization?	• Defining the skills and roles needed to perform the standard processes of the organization. • # Skills that are needed. • # Roles that are performed.
Q9	Do you document the training needs forperforming the roles in the standard processes of the organization?	• Documenting the training needs for performing theroles in the standard processes of the organization. • # Roles that are performed.
Q10	Do you review the strategic needs of the organization and the needed training as it is required?	• Reviewing the strategic needs of the organization and the needed training as it is required. • # Therequirements of strategic training. • # Skills that are needed.

TABLE. II. MEASURES AND QUESTIONSTHAT ARE ACCOMPANIED WITH THE 2ND PRACTICE IN OT

	Questions	**Measures**
Q1	Do you decide which training is the responsibility of the organization? (Organizational training related to the training needs that are generally over projects).	• Deciding which training is the responsibility of the organization. • # Training within the responsibility of the organization.
Q2	Do you decide which training is the duty of the individual projects? (projects have the main duty of recognizing their training needs.	• Deciding which training is the duty of the individual projects. • # Training is the duty of the individual projects.
Q3	Do you investigate the training needs that are determined by the projects group? (For defining the general training needs that can be addressed organization-wide).	• Investigating the training needs that are determinedby theproject groups. • # Training needs that are determined by the project groups.
Q4	Do you consult project groups about the way of meeting their training needs? (Some examples of training offered by the project group include training in the application, training in the unique tools, and training in security and safety).	• Consulting project groups about the way of meeting their training needs
Q5	Do you record promises for providing the project groups with training?	• Recording promises for providing training to project groups. • # Recorded promises for providing training.

TABLE. III. MEASURES AND QUESTIONSTHAT ARE ACCOMPANIED WITH THE 3RD PRACTICE IN OT

	Questions	**Measures**
Q1	Do you develop a training tactical plan for the organization? (Tactical plan refers to the plan set for delivering the training that is the duty of the organization).	• Development of a training tactical plan for the organization. • # The objectives of the tactical plan.
Q2	Do you determine the training subjects?	• Determining the training subjects. • # The training subjects
Q3	Do you set schedules based on the training actions and their related issues?	• Setting schedules based on the training actions and their related issues. • # The training subjects.
Q4	Do you set methods to be used for training?	• Setting methods to be used for training. • # training methods.
Q5	Do you set quality requirements and standardsfor the training materials?	• Settingquality requirements and standards for the training materials. • # Quality standards
Q6	Do you determine the needed resources such as tools, facilities, knowledge, skills, staffing, and the environment?	• Determining the needed resources. • # The needed resources.
Q7	Do you document the extent of commitments shown bythe ones responsible for executing the plan?	• Documenting the extent of commitments shown by the ones responsible for executing the plan. • # The documented commitments.
Q8	Do you review the plan and commitments as required?	Reviewing the plan and commitments as required.

D. Measures for Specific Practice 4

Establish a training capability:Establish and maintain a training capability to address organizational training needs.

Table IV presents measures and questions. Thesemeasures and questions are accompanied with the 4th practice in the OT.

E. Measures for Specific Practice 5

Deliver training:Deliver training following the organizational training tactical plan.

Table V presents measures and questions. Thesemeasures and questions are accompanied with the 5thpractice in OT

F. Measures for Specific Practice 6

Creating training records: Creating and maintaining the records oforganizational training.

Table VI presents measures and questions. Thesemeasures and questions areaccompanied with the 6th practice in OT.

G. Measures for Specific Practice 7

Assess the training effectiveness: Assess the effectiveness of the organization’s training program.

Table VII presents measures and questions. Thesemeasures and questions areaccompanied withthe 7thpractice in OT.

TABLE. IV. MEASURES AND QUESTIONS THAT ARE ACCOMPANIED WITHTHE4TH PRACTICE IN OT

	Questions	Measures
Q1	Do you choose suitable methods for meetingthe demands of organizational training? (Some examples of training methods include facilitated videos, guided self-study, computer-aided instruction,classroom training,and chalk talks).	• Choosing the suitable methods to meet the organizational training demands • # The methods to be used for meeting organizational training demands
Q2	Do you take into account the reasons that may affect the selection of training methods, such asaudience-specific knowledge, work environment, schedule, and costs?	• Taking into account the reasons that may influence the selection of training methods. • # The methods to be used for meeting organizational training demands
Q3	Do you decide whether to produce training materials inside or outside the organization? (Some examples of measures that can be used to decide, availability of training from outside sources, availability of time to prepare for the project, availability of internal expertise, and applicability to business goals).	• Deciding whether to produce training materials inside or outside the organization. • # Measures that can be used to decide. • # Inside produced material. • # Outside produced material.
Q4	Do you develop or acquire qualified lecturers?	• Acquisition of development of qualified lecturers. • # Acquisition of qualified lecturers. • # Development of qualified lecturers.
Q5	Do you describe the training in the training curriculum of the organization? (Some examples of description of each course include the intended audience, training objectives, topics, prerequisites, lesson plans, and length of the training).	• Describing the training in the training curriculum of the organization. • # The intended audience. • # The training objectives. • # The prerequisites. • The length of the training.
Q6	Do you review the supporting artifacts and the training materials as required? (Some examples of conditions must be reviewed when training needs change (e.g. due to the new technology). That must be done after assessing the training results in need to change (e.g. assessment for the training or instructors)).	• Reviewing the supporting artifacts and training materials as needed. • # Training evaluations. • # Instructor evaluations. • # Supporting artifacts. • # Training needs change.

TABLE. V. MEASURES AND QUESTIONSTHAT AREACCOMPANIED WITH THE5TH PRACTICE IN OT

	Questions	Measures
Q1	Do you take into account the required experience when you choose somebody to be trained?	• Taking into account the required experience. • # Trainers who have the required experience to receive training. • # Trainers who don't have the required experience to receive training.
Q2	When you choose the trainees, do you take into account the abilities and skills that trainees must have to perform their roles?	• Taking into account the skills and abilities that must be possessed by the trainees to do their roles. • # The trainers who have the required abilities and skills to do their roles. • # The trainers who do not have the required abilities and skills to do their roles.
Q3	Do you take into account the need to present a competency addition to the crucial working area?	• Taking into account the need to present competency for the crucial working area. • # Crucial working area.
Q4	Do you choose people who will get the training needed for doing their tasks efficiently?	• Choosing people who will get the trainingneeded for doing their tasks. • # People chosen to get training
Q5	Do you waive people who already possess the skills and knowledge needed for doing their tasks?	• Waiving people who already possess the skills, and knowledge needed for doing their tasks. • # People who already possess skills. • # People who already possess knowledge.
Q6	Do you schedule resources combined with training (e.g. instructors, facilities)?	• Scheduling resources combined with training. • # Instructors. • # Tools.
Q7	Is the provided training consistent with the real environment conditions?	• Consistency between the provided training and real operation conditions. • # Training courses that are consistent with the real environmental conditions.
Q8	Does the training involve activities that simulate the actual work circumstance?	• Involving activities that simulate the actual work circumstance. • # Activities that simulate the actual real work circumstance.
Q9	Do you track the performance of the training based on the plan?	• Tracking the performance of training based on the plan. • # Contradictions with the plan. • # Agreements with the plan.
Q10	Do you support the training within a feasible time after the training was planned?	• Supporting the training to be in a feasible time after the training was planned. • The range of time the training will start.

TABLE. VI. MEASURES AND QUESTIONS THAT AREACCOMPANIED WITH THE 6TH PRACTICE IN OT

	Questions	Measures
Q1	Do you create and maintain documents about the organization's training?	• Creating and maintaining documents about the organization's training. • # Documents related to training
Q2	Do you save the histories of every student who completes each training course successfully or unsuccessfully?	• Saving thehistories of every student who complete each training course successfully or unsuccessfully • # The students who complete the training. • # The students who complete the training successfully. • # The students who unsuccessfully complete the training.
Q3	Do you keep documents about every employee who waived from training?	• Keeping documents about every employee who waived from training. • # Documents about the employee who waived from training.
Q4	Do you offer the training documents to the relevant people for assignment purposes?	• Offering the training documents to the relevant people for assignment purposes. • # Offeringthe training documents to students.
Q5	Do you produce a skill matrix that includes records of training?	• Producing a skill matrix that includes records of training. • Academic qualification • # Years of experience. • # Training sponsored by the organization.

TABLE. VII. MEASURES AND QUESTIONS THAT AREACCOMPANIED WITH THE 7TH PRACTICE IN OT

	Questions	Measures
Q1	Do you distribute post-training surveys to the training members to assess the training performance?	• Distributing post-training surveys to the training members to assess the training performance • # The training members that the post-training surveys were distributed to them
Q2	Do you distribute the post-training surveys to managers for assessing the training performance?	• Distributing post-training surveys to managers for assessing the training performance. • # The managers that thepost-training surveys were distributed to them
Q3	Do you use the outcomes of training evaluation to improve the materials of training?	• Using the outcomes of training evaluation to improve the materials of training. • # Improved course materials.
Q4	Do you ask the training participants to take an exam?	• Making examinations for the training participants. • # The training participants who were Examined.
Q5	Do you assess the projects (completed or in-progress or) to identify whether the staff information is sufficient or not?	• Assessing the projects. • #Completed evaluated projects. • # In-progress evaluated projects.
Q6	Do you collect student evaluations about how the activities of the training met their requirements?	• Collecting student evaluations about how the activities of the training met their requirements. • # Students with satisfying requirements. • # Students with unsatisfied requirements.

VI. VALIDITY AND RELIABILITY OF THE DEFINED MEASURES

To proof the validity and reliability of the set of the defined measures linked to the Organizational training process (OT), the researcher developed a questionnaire. The results obtained through the questionnaire were used to prove that the defined measures are really suitable for measuring the seven goals (specific practices). The researcher calculated the Cronbach alpha coefficient value through the SPSS.

Cronbach alpha coefficient valuescalculated formeasuring the inner coherence and consistency. It provides an answer to the following question: (Do all of the defined measures for each specific practice meet the same goal?) The values of the Cronbach alpha coefficient are within the range of 0 and 1. The closer the value to 1, the higher the inner consistency between the items shall be [20]. If the value is less than 0.5, the internal consistency between the items shall be considered low and unaccepted [20].

The researcher checked the validity of the questionnaire by passing it to academics working at the software engineering department at Zarqa University. The questionnaire was also passed to professionals (designers and programmers) who work atZarqa University. The researcher distributed the questionnaire in six software development institutions in Jordan. Three hundred questionnaire forms were retrieved. The questionnaire forms are filled by analysts, designers, and programmers. Each questionnaire consists of seven parts. Each part addresses a goal (specific practice) of the organizational training process (OT).

Asit's displayed through appendix A, every part includes a combination of statements (measures) that are linked to each goal (specific practice). Five multiple answers are provided, which are: strongly agree, agree, neither agree nor disagree, disagree, strongly disagree. The participant must choose one answer from these answers.

After calculating theCronbach Alpha coefficient values, the researcher obtained values that are within the range of 1-0.5 for the seven parts. These values are: 0.758, 0.668, 0.680, 0.659, 0.730, 0.698, and 0.767. That means that the statements (measures) are consistent with one another, reliable and considered valid for measuring the seven specific practices. These practices are:

1). Establishing strategic training needs.

2). Identifying which training needsis the responsibility of the organization,

3). Establishing the organizational training tactical plan,

4). Establishing a training capability,

5). Delivering training,

6). Establishing training records,

7). Assessing the training effectiveness.

VII. Conclusion

Through this paper, the researcher defined common measures for one important process in (CMMI-SW). Thisprocess is represented in organizational training. The researcher applied the Goal Question Metrics (GQM) model to seven relevant goals (specific practices) of organizational training. That was done for defining the measures.

The researcher confirmed that the defined measures are true measures for each one of the seven goals (specific practices). That's confirmed through using a questionnaire for proving the validity and the reliability of the set of measures that the researcher defined for the organizational training process (OT). The researcher calculated the Cronbach alpha coefficient value through the SPSS.

In this study, through using the combination of the defined measures, organizations shall have a valid method for checking the tasks related to the organizational training process. The researcher will have a mature organizational training process, provided that the defined measures in this paper get implemented. The set of the defined measures can be implemented for managing and assessing the project's products and processes in the OT.

In the future, other process areas in the capability maturity model integration (CMMI) will be analyzed and measured.

References

[1] F. Rabbanikhah, A. M. Jaghagh, R. M. Gholizadeh, S. Sabouri, and S. Alirezaei, "Analyzing effective factors in efficiency of organizational trainings (A Case Study: Employees of Ministry of Health and Medical Education)," International Journal of Humanities and Cultural Studies (IJHCS) ISSN 2356-5926, pp. 2136-2154, 2016.

[2] C. P. Team, Capability Maturity Model® Integration for Development Version 1.3 (Software Engineering Institute). 2010.

[3] A. M. Saks and L. A. Burke-Smalley, "Is transfer of training related to firm performance?," International Journal of Training and Development, vol. 18, no. 2, pp. 104-115, 2014.

[4] H. N. Ismail, "Training and organizational commitment: exploring the moderating role of goal orientation in the Lebanese context," Human Resource Development International, vol. 19, no. 2, pp. 152-177, 2016.

[5] C. S. Cheah, V. S. W. Chong, S. F. Yeo, and K. W. Pee, "An empirical study on factors affecting organizational commitment among generation X," Procedia-Social and Behavioral Sciences, vol. 219, pp. 167-174, 2016.

[6] C. J. Chang, "The Factors on Elderly Employment Project Outcome: Appropriation of work, Job Training Satisfaction, Intra-organizational Cooperation," International Journal of Social Science and Humanity, vol. 6, no. 1, p. 14, 2016.

[7] C. Ebert, "Software Measurement for Better Project and Process Quality," UPGRADE (the European Journal for the Informatics Professional), vol. 10, no. 5, 2009.

[8] O. Gómez, H. Oktaba, M. Piattini, and F. García, "A systematic review measurement in software engineering: State-of-the-art in measures," in International Conference on Software and Data Technologies, 2006: Springer, pp. 165-176.

[9] C. Jones, "Implementing a Successful Measurement," IT Metrics and Benchmarking: Part II, vol. 16, no. 11, p. 12, 2003.

[10] L. O. Ejiogu, "Five principles for the formal validation of models of software metrics," ACM SIGPLAN Notices, vol. 28, no. 8, pp. 67-76, 1993.

[11] B. Kitchenham, D. R. Jeffery, and C. Connaughton, "Misleading metrics and unsound analyses," IEEE software, vol. 24, no. 2, pp. 73-78, 2007.

[12] V. R. Basili and H. D. Rombach, "The TAME project: Towards improvement-oriented software environments," IEEE Transactions on software engineering, vol. 14, no. 6, pp. 758-773, 1988.

[13] B. Kitchenham, S. L. Pfleeger, and N. Fenton, "Towards a framework for software measurement validation," IEEE Transactions on software Engineering, vol. 21, no. 12, pp. 929-944, 1995.

[14] N. E. Fenton, R. W. Whitty, and Y. Iizuka, Software Quality Assurance and Measurement: A Worldwide Perspective. Itp-Media, 1995.

[15] N. Fenton and J. Bieman, Software metrics: a rigorous and practical approach. CRC press, 2014.

[16] R. E. Park, W. B. Goethert, and W. A. Florac, "Goal-Driven Software Measurement. A Guidebook," Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst, 1996.

[17] V. Mahnic and N. Zabkar, "Measurement repository for Scrum-based software development process," in Conference on computer Engineering and Application (CEA, 08) Acapulco, Mexico, 2008.

[18] L. C. Briand, S. Morasca, and V. R. Basili, "An operational process for goal-driven definition of measures," IEEE Transactions on Software Engineering, vol. 28, no. 12, pp. 1106-1125, 2002.

[19] R. S. Pressman, Software engineering: a practitioner's approach. Palgrave Macmillan, 2005.

[20] D. George, "Mallery. 2003. SPSS for Windows step by step: A simple Guide and reference 11.0 Update," ed: Allyn and Bacon. Boston.

[21] M. C. Paulk, C. V. Weber, S. M. Garcia, M. B. Chrissis, and M. Bush, "Key practices of the capability maturity model for software, version 1.1," Pittsburgh, PA: Software Engineering Institute (SEI), 1993.

[22] A. Pyster, "What beyond CMMI is needed to help assure program and project success?," in Software Process Workshop, 2005: Springer, pp. 75-82.

[23] P. Monteiro, R. J. Machado, R. Kazman, C. Simões, and P. Ribeiro, "RUP Alignment and Coverage Analysis of CMMI ML2 Process Areas for the Context of Software Projects Execution," in International Conference on Software Quality, 2014: Springer, pp. 214-228.

[24] Y. Lee and J. Chen, "Experience in introducing CMM to a telecommunication research organization," Journal of software engineering studies, vol. 1, no. 1, pp. 8-16, 2006.

[25] H.-C. Young, T. Fang, and C. Hu, "A successful practice of applying software tools to CMMI process improvement," Journal of Software Engineering Studies, vol. 1, no. 2, pp. 78-95, 2006.

[26] Z. D. Kelemen, R. Kusters, J. Trienekens, and K. Balla, "Towards complexity analysis of software process improvement frameworks," Budapest, Technical Report TR201301, 2013.

[27] W. Xiong and Y. Cao, "Comments on Software Process Improvement Methodologies Using QFD," Applied Mathematics & Information Sciences, vol. 7, no. 3, p. 1137, 2013.

[28] W. E. Wong and T. Ma, Emerging technologies for information systems, computing, and management. Springer, 2013.

[29] C.-S. Wu and D. B. Simmons, "Software Project Planning Associate (SPPA): a knowledge-based approach for dynamic software project planning and tracking," in Proceedings 24th Annual International Computer Software and Applications Conference. COMPSAC2000, 2000: IEEE, pp. 305-310.

[30] G. Xu, H. Hu, P. Yu, J. Lv, P. Qu, and M. Zhu, "Supporting flexibility of the CMMI process framework with a multi-layered process model," in 2013 10th Web Information System and Application Conference, 2013: IEEE, pp. 409-414.

[31] M. Khraiwesh, "Requirements Validation Measures in CMMI," World of Computer Science and Information Technology Journal (WCSIT), vol. 1, no. 2, pp. 26-33, 2011.

[32] J. H. Baumert and M. S. McWhinney, "Software measures and the capability maturity model," CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 1992.

[33] T. F. Hammer, L. L. Huffman, L. H. Rosenberg, W. Wilson, and L. E. Hyatt, "Doing requirements right the first time," CROSSTALK The Journal of Defense Software Engineering, pp. 20-25, 1998.

[34] D. Janakiram and M. Rajasree, "Request: Requirements-driven quality estimator," ACM SIGSOFT Software engineering notes, vol. 30, no. 1, p. 4, 2005.

[35] M. Khraiwesh, "Risk management measures in CMMI," International Journal of Software Engineering & Applications, vol. 3, no. 1, p. 149, 2012.

[36] M. Khraiwesh, "Project Planning Measures in CMMI," International Journal of Software Engineering & Applications, vol. 4, no. 2, p. 103, 2013.

[37] M. Khraiwesh, "Process and product quality assurance measures in CMMI," International Journal of Computer Science and Engineering Survey, vol. 5, no. 3, p. 1, 2014.

[38] M. Khraiwesh, "Integrated project management measures in CMMI," International Journal of Computer Science and & Information Technology (IJCSIT), vol. 7, no. 5, 2015.

[39] M. Khraiwesh, "Configuration Management Measures in CMMI," International Journal of Applied Engineering Research, vol. 12, no. 18, pp. 7546-7557, 2017.

[40] A. Loconsole, "Measuring the requirements management key process area," in Proceedings of ESCOM-European Software Control and Metrics Conference, London, UK, 2001, pp. 67-76.

[41] M. Khraiwesh, "Project Monitoring and Control in CMMI, Project Monitoring and Control Measures in CMMI," International Journal of Computer Science & Information Technology (IJCSIT) Vol, vol. 5, 2013.

[42] V. Caldiera and H. Rombach, "Goal Question Metric Paradigm, Encyclopedia of Software Engineering," ed: Wiley: Hoboken, NJ, USA, 1994.

APPENDIX A

A. *Questionnaire And AnalysisQuestionnaire:*

The table below presents a questionnaire that is constructed for theorganizational training process. Organizational training aimed at developing employees' knowledge and skills. It shall enable employees to carry out their duties efficiently and effectively.

The organizational training process has the following goals:

1). Establish strategic training needs.
2). Identify which training needs must be met by the organization.
3). Create an organizational training tactical plan.
4). Establish a training capability.
5). Deliver training.
6). Create training records.
7). Assess the effectiveness of the training.

To estimate the fulfillment of the defined specific practices (goals), we will determine a few sentences linked to each specific practice. The collected information from these sentences will help us in the realization of the above seven goals.

Please, fill the form by writing down (√) in a suitable position. Responding to the related question: do you see that the following statements possess an impact on the realization of the related specific practice (goal)?

Goal1: Establish strategic training needs.

(Do you see that these sentences can be used to check the achievement of goal1: Establish strategictrainingneeds?)

Statement Serial	Statements	Strongly agree	Agree	Neither agree nor disagree	disagree	Strongly disagree
1	Classifying the training needed for improving the skills for carrying out the project activities.					
2	Identifying the training needed for developingthe information required for carrying out the project activities.					
3	Buildingstrategic training requirements.					
4	Defining the sources of strategic training requirements.					
5	Analyzing the strategic business plan of the organization and process improvement plan of the organization to know possible training needs.					

Adaptive Sequential Constructive Crossover Operator in a Genetic Algorithm for Solving the Traveling Salesman Problem

Zakir Hussain Ahmed
Department of Mathematics and Statistics, College of Science
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Kingdom of Saudi Arabia

***Abstract*—Genetic algorithms are widely used metaheuristic algorithms to solve combinatorial optimization problems that are constructed on the survival of the fittest theory. They obtain near optimal solution in a reasonable computational time, but do not guarantee the optimality of the solution. They start with random initial population of chromosomes, and operate three different operators, namely, selection, crossover and mutation, to produce new and hopefully better populations in consecutive generations. Out of the three operators, crossover operator is the most important operator. There are many existing crossover operators in the literature. In this paper, we propose a new crossover operator, named adaptive sequential constructive crossover, to solve the benchmark travelling salesman problem. We then compare the efficiency of the proposed crossover operator with some existing crossover operators like greedy crossover, sequential constructive crossover, partially mapped crossover operators, etc., under same genetic settings, for solving the problem on some benchmark TSPLIB instances. The experimental study shows the effectiveness of our proposed crossover operator for the problem and it is found to be the best crossover operator.**

***Keywords*—Genetic algorithm; adaptive sequential constructive crossover; traveling salesman problem; NP-hard**

I. Introduction

The usual travelling salesman problem (TSP) is very famous combinatorial optimization problem that finds a least cost Hamiltonian cycle in a network. The RAND Corporation introduced the TSP in 1948. The Corporation's reputation helped to make the TSP well-known and popular problem. The TSP also became popular at that time due to the new subject of linear programming and attempts to solve combinatorial optimization problems. It can be stated as.

A network with 'n' nodes, with 'node 1' as 'depot' and a travel cost (or distance, or travel time etc.,) matrix C= [c_{ij}] of order n associated with ordered pairs (i, j) of nodes is given. The problem is to find a least cost Hamiltonian cycle. Based on the structure of the cost matrix, the TSPs are classified into two types as symmetric and asymmetric. If $c_{ij} = c_{ji}$, $\forall$ i, j, the TSP is symmetric, otherwise, it is asymmetric. For asymmetric TSP with n nodes, there are $(n-1)!$ possible solutions with at least one of them provide the minimum cost. For symmetric TSP, there are $\frac{(n-1)!}{2}$ possible solutions along with same valued reverse cyclic permutations. If there are only 10 nodes, then there are 362,880 and 181,440 tours for asymmetric TSP and symmetric TSP, respectively. The number of possible solutions in both types is very large for any size, n; so, a complete search is very difficult, if it is not impossible. That means, the problem is very difficult to solve. The TSP has been researched by several researchers for mainly three reasons. First, it can model many real-life problems. Second, it is NP-Hard [1]. Third, NP-Hard problems are so difficult that no one has found any efficient algorithm to solve them for large sized problem instances. Also, NP-hard problems are equivalent to each other; so, if one can develop efficient algorithm for solving one of them, then one could develop efficient algorithm for others.

The TSP has application in several situations such as automatic drilling of printed circuit boards and threading of scan cells in a testable very-large-scale-integrated (VLSI) circuit, automatic drilling of printed circuit boards and circuits, computer wiring, X-ray crystallography, movement of people [2].

Several exact and heuristic/metaheuristic algorithms have been reported for solving the TSP. Branch and bound [3], branch and cut [4], and lexisearch algorithm [5] are some exact algorithms. These algorithms provide the exact optimal solution to the problem, but as the problem size increases computational time increases exponentially. As reported by Deng et al. [6] only small sized TSP instances can be solved to exact optimality. Since some practically large problem instances must be solved, hence it is important to obtain heuristically optimal solution by ensuring the quality of the solution in reasonable time, rather to obtain exact optimal solution in hell of time. Heuristic/metaheuristics algorithms give near optimal solution in a reasonable computational time, but do not guarantee the optimality of the solution. Example of metaheuristic algorithms are ant colony optimization [7], genetic algorithm [8], simulated annealing [9], state transition algorithm [10], tabu search [11], artificial neural network [12], artificial bee colony [13], black hole [14], and particle swarm optimization [7]. Out of these metaheuristic algorithms, genetic algorithm (GA) is one of the best and widely used algorithm to solve the TSP as well as other combinatorial optimization problems in computer science and operations research.

GAs are proposed by John Holland in 1970s which are based on imitating Darwin's theory of 'the survival of the fittest' in natural biology [8]. To solve a real-world problem using GAs, two most important conditions are to be fulfilled: (i) a chromosome representing a solution, and (ii) an objective/fitness function can be defined. Any simple GA begins with random initial population, called gene pool of chromosomes, and operates three different operators to produce new, usually better, populations in consecutive iterations/generations. Selection is the 1st operator in which chromosomes are duplicated to next generation probabilistically. Crossover is the 2nd operator in which couples of chromosomes are selected randomly and mated to produce new and better chromosomes. Mutation is the 3rd operator which alters occasionally a chromosome position value. Crossover along with selection operator is the main leading procedure in GAs. Mutation expands search space and defends from loss of any genetic substance due to selection and crossover operators.

Though GA is one of the best algorithms, however, its performance depends on initial population, selection, crossover and mutation operators, and some parameters such as population size, crossover probability, mutation probability and stopping condition (Goldberg, 1989). Among different operators, crossover plays very important role in GAs, and accordingly many crossover operators have been developed and reported in the literature for solving the TSP [15]. This paper aims to propose a modified version of sequential constructive crossover (SCX) [16] named adaptive sequential constructive crossover (ASCX) and then compare with eight crossover operators including SCX to assess suitability for the TSP.

This paper is organized as follows: Section II discusses GAs using some existing crossover operators and our proposed crossover operator, named adaptive crossover sequential constructive crossover operator, for the TSP, while design of variant GAs is discussed in Section III. Section IV describes computational experiences for sixteen variant GAs using eight crossover operators with two possibilities of mutation operator and discussions. Finally, Section V presents concluding remarks and future works.

II. Genetic Algorithms for the TSP

For applying GA to any optimization problem, one must find a way for representing solutions as legal chromosomes such that crossovers of legal chromosomes result in legal chromosomes. The methods for representing solutions differ by problem and, contain a certain art. There are many representation methods for solving the TSP using GAs. Some of them are binary, adjacency, ordinal, matrix and path representations. We consider only the path representation that simply lists the node labels such that no node can appear twice in the same chromosome. For example, let {1, 2, 3, 4, 5, 6, 7, 8, 9} be the node labels in a 9-node instance, then a tour {1→ 9→6→ 2→7 → 4→3→8→ 5 →1} may be represented as (1, 9, 6, 2, 7, 4, 3, 8, 5). The objective function is the sum of the costs of all edges in the tour.

A. Initial Population and Selection Operator

In GAs, after generating the random population of chromosomes, selection operator is applied. In selection operator, chromosomes are copied into mating pool with a probability related to their fitness value. By transferring highly fit chromosomes to next generation mating pool, selection mimics the Darwinian theory of survival-of-the-fittest in the natural biology. In natural biology, fitness is determined by an individual's capability to survive predators, epidemic, and other difficulties to maturity and following selection. In this stage no new chromosome is created. The commonly used selection operator is the proportionate selection operator, where an individual is selected for the mating pool according to a probability related to its fitness value. We have considered the stochastic remainder selection process [17] for our GAs.

B. Existing Crossover Operators

Since the crossover operator plays a vital role in GA, so many crossover operators have been proposed for the TSP. However, the traditional crossover operators such as one-point, two-point, and uniform crossover operators are not suitable for the TSP. Two kinds of crossover operators have been developed for the TSP – distance-based and blind crossover operators [18]. We consider some of them from both kinds and compare our proposed crossover operator with them.

1). Partially mapped crossover operator: Goldberg and Lingle [19] developed the partially mapped crossover (PMX) that used two crossover points. It defines an interchange mapping in the section between these points. PMX was the first crossover for the GA to solve the TSP. Consider, for example, the two parent chromosomes P_1: (1, 2, 3, 4, 6, 9, 5, 7, 8) and P_2: (1, 3, 5, 7, 8, 9, 4, 2, 6). We shall consider same pair of chromosomes for illustrating all the crossover operators considered here. Also, we fix headquarters (first gene) as 'node 1'. Suppose the randomly selected cut points are between 3rd and 4th genes and between 7th and 8th genes as follows (these cut points are marked with "|"):

P_1: (1, 2, 3 | 4, 6, 9, 5 | 7, 8) and

P_2: (1, 3, 5 | 7, 8, 9, 4 | 2, 6)

We always fix first gene as 'node 1'. The mapping sections are between the cut points. In this example, the mapping systems are 4↔7, 6↔8, 9↔9, and 5↔4. Now these mapping sections are copied with each other to build offsprings as follows:

O_1: (1, *, * | 7, 8, 9, 4 | *, *),

O_2: (1, *, * | 4, 6, 9, 5 | *, *)

Then we can add more genes from the original parents which do not result any conflict as follows:

O_1: (1, 2, 3 | 7, 8, 9, 4 | *, *),

O_2: (1, 3, * | 4, 6, 9, 5 | 2, *)

The first * in the first offspring should be 7 that comes from first parent, but it is already present in this offspring, so we check mapping 4↔ 7, but 4 is also present in this offspring, again check mapping 5↔4, so 5 is added. Similarly, the second * in first offspring should be 8 that comes from first parent, but it is present in this offspring, so we check mapping 6↔8 and hence, we add 6 at second *. Thus, the first offspring becomes

O_1: (1, 2, 3 | 7, 8, 9, 4 | 5, 6),

Similarly, we build the second offspring as:

O_2: (1, 3, 7 | 4, 6, 9, 5 | 2, 8)

2). Ordered crossover operator: Davis [20] developed the ordered crossover (OX) that builds offspring by choosing a subsequence of a tour from one parent and preserving the relative order of nodes from the other parent. Consider the same example parent chromosomes with randomly chosen two cut points marked by "|":

P_1: (1, 2, 3 | 4, 6, 9, 5 | 7, 8) and

P_2: (1, 3, 5 | 7, 8, 9, 4 | 2, 6)

We always fix first gene as 'node 1'. At first, the offsprings are built by copying the genes between the cuts with similar way into the offsprings that lead the offsprings as:

O_1: (1, *, * | 4, 6, 9, 5 | *, *),

O_2: (1, *, * | 7, 8, 9, 4 | *, *)

Then beginning from the second cut point of one parent, the genes from the other parent are copied in the same order except the existing genes. The sequence of the genes in the second parent from the second cut point is "2 →6→3→5→7→8 →9 →4." After omitting the genes 4, 6, 9 and 5 that are already present in the first offspring, the sequence becomes "2→3→7→8", which is placed in the first offspring beginning from the second cut point:

O_1: (1, 7, 8 | 4, 6, 9, 5 | 2, 3).

Similarly, we build the second offspring as:

O_2: (1, 6, 5 | 7, 8, 9, 4 | 2, 3)

3). Alternating edges crossover operator: Grefenstette et al. [21] proposed alternating edges crossover (AEX) operator that assumes a chromosome as a directed cycle of arcs. Only one offspring is built by selecting alternative arcs from both parents, with some additional random selections in case of infeasibility. Consider the same example parent chromosomes P_1: (1, 2, 3, 4, 6, 9, 5, 7, 8) and P_2: (1, 3, 5, 7, 8, 9, 4, 2, 6).

First, the arc (1, 2) is first selected from the first parent and copied to the offspring. Then the arcs (2, 6) from second parent, (6, 9) from first parent and (9, 4) from second parent are selected and copied to the offspring. Then, arc (4, 6) is selected from first parent, however, this arc produces a cycle and a new arc leaving the node 4 to a node not yet visited is selected randomly. Suppose the arc (4, 3) is chosen. Then, the arcs (3, 5) from second parent, (5, 7) from first parent and (7, 8) from second parents are selected. This way the offspring is built as follows:

O: (1, 2, 6, 9, 4, 3, 5, 7, 8)

All arcs in the offspring are inherited from the parents, apart from the arc (4, 3).

4). Cycle crossover operator: Oliver et al. [22] developed cycle crossover (CX) that builds an offspring where every node and its corresponding position originated from one of the parents. Consider the same example parent chromosomes P_1: (1, 2, 3, 4, 6, 9, 5, 7, 8) and P_2: (1, 3, 5, 7, 8, 9, 4, 2, 6).

As we fix first gene as node 1, for the next position, we select randomly between 2 and 3. Suppose we select node 2, then the offspring becomes:

O_1: (1, 2, *, *, *, *, *, *, *)

Every gene in the offspring is taken from one of its parents with the same position, so the next gene to be considered must be bit 3, as this gene from the second parent is just below the selected gene 2. In the first parent this gene is at 3rd position; thus, the offspring becomes:

O_1: (1, 2, 3, *, *, *, *, *, *)

Next gene will be 5 of second parent as it is just below the current gene 3, which is present at 7th position in first parent. Thus, the offspring becomes:

O_1: (1, 2, 3, *, *, *, 5, *, *)

Next gene will be 4 of second parent as it is just below the current gene 5, which is present at 4th position in first parent. Thus, the offspring becomes:

O_1: (1, 2, 3, 4, *, *, 5, *, *)

Next gene will be 7 of second parent as it is just below the current gene 4, which is present at 8th position in first parent. Thus, the offspring becomes:

O_1: (1, 2, 3, 4, *, *, 5, 7, *)

Next, we have node 2, which is already present in the offspring; thus, we have completed a cycle and hence, we fill the remaining blank positions with the genes of those positions which are present in second parent. This way the offspring is built as follows:

O_1: (1, 2, 3, 4, 8, 9, 5, 7, 6)

Similarly, we build the second offspring as (same as P_2):

O_2: (1, 6, 5, 7, 8, 9, 4, 2, 3)

5). Greedy crossover operator: Grefenstette et al. [21] also proposed greedy crossover (GX) for the TSP that selects a starting node randomly. Then in each step, four neighbor nodes of currently selected node in both parents are considered, and the cheapest one (not present in the offspring) is selected. If the cheapest node or all four neighbour nodes are present in the offspring, then any node from the remaining is selected randomly. This operator creates only one offspring from two parents. Let us illustrate the GX through the 9-node example given as cost matrix in Table I and the same parent chromosomes considered above.

TABLE. I. THE COST MATRIX

Node	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	999	7	15	9	10	6	8	9	10
2	11	999	8	7	11	3	6	4	3
3	15	5	999	16	12	5	8	13	4
4	2	5	11	999	9	13	14	4	2
5	8	6	3	5	999	6	7	10	9
6	6	13	8	11	5	999	5	4	5
7	5	15	3	7	12	6	999	8	9
8	9	3	9	14	3	11	8	999	10
9	11	16	3	9	10	7	9	10	999

As we fixed first gene as ‘node 1’, the offspring is initiated as (1). The nodes 2 and 3 are neighbours of 1 with their costs 7 and 15 respectively. The node 2 is cheaper, so, it is copied into the incomplete offspring: (1, 2).

Next, the nodes 3, 1, 6 and 4 are neighbours of 2 with their costs 8, 11, 3 and 7 respectively. The node 6 is the cheapest, so, it is copied into the incomplete offspring: (1, 2, 6).

Next, the nodes 9, 4, 1 and 2 are neighbours of 6 with their costs 5, 11, 6 and 13 respectively. The node 9 is the cheapest so, it is copied into the incomplete offspring: (1, 2, 6, 9).

Next, the nodes 5, 6, 4 and 8 are neighbours of 9 with their costs 10, 7, 9 and 10 respectively. The node 6 is the cheapest, but it is already present in the offspring, so, node 3 is selected randomly and it is copied into the incomplete offspring: (1, 2, 6, 9, 3).

Next, the nodes 4, 2, 5 and 1 are neighbours of 3 with their costs 16, 5, 12 and 15 respectively. The node 2 is the cheapest, but it is already present in the offspring, so, node 4 is selected randomly and it is copied into the incomplete offspring: (1, 2, 6, 9, 3, 4).

Continuing in this way, we have the complete offspring: (1, 2, 6, 9, 3, 4, 5, 7, 8) with cost 67.

6). Sequential constructive crossover operator: Ahmed [16] proposed the sequential constructive crossover (SCX) operator which is modified in [23] that constructs an offspring using better arcs based on their cost present in the parents' structure. Furthermore, it also uses the better arcs that are present neither in the parents' structure. SCX sequentially searches both parent chromosomes and considers the first legitimate node (i.e. unvisited node) that appeared after the present node and in case, if no legitimate node is found in either of the parent chromosomes, it sequentially searches from the beginning of the chromosome and then compares their associated cost to decide the next node of the child chromosome. The SCX is compared with edge recombination crossover (ERX) and generalized N-point crossover (GNX) on symmetric and asymmetric TSPLIB instances. As reported, SCX is better than ERX and GNX. Khan [24] presented a comparative study among eight different crossover operators, namely, Two-Point Crossover, PMX, CX, Shuffle Crossover, ERX, Uniform Order-based Crossover, Sub-tour Exchange Crossover and SCX, and found that SCX outperformed other operators in achieving good quality solution for the TSP. Further, SCX is successfully applied to many other combinatorial optimization problems ([25]-[31], [32]). Let us recall the algorithm for the SCX [23].

Step 1: Start from 'node 1’ (i.e., current node p =1).

Step 2: Sequentially search both parent chromosomes and consider the first ‘legitimate node' (the node that is not yet visited) appeared after 'node p’ in each parent. If no 'legitimate node' after 'node p’ is present in any of the parents, search sequentially from the starting of the parent and consider the first 'legitimate node', and go to Step 3.

Step 3: Suppose the 'node α' and the 'node β' are found in 1st and 2nd parent respectively, then for selecting the next node go to Step 4.

Step 4: If $c_{p\alpha} < c_{p\beta}$, then select 'node α', otherwise, 'node β' as the next node and concatenate it to the partially constructed offspring chromosome. If the offspring is a complete chromosome, then stop, otherwise, rename the present node as 'node p' and go to Step 2.

Let us illustrate the SCX through the same example given above. Select 'node 1' as the 1st gene. The legitimate nodes after node 1 in P_1 and P_2 are 2 and 3 respectively with c_{12}=7 and c_{13}=15. Since c_{12}<c_{13}, we accept node 2. So, the partially constructed chromosome will be (1, 2).

The legitimate nodes after node in P_1 and P_2 are nodes 3 and 6 respectively with c_{23}=8 and c_{26}=3. Since c_{26}<c_{23}, we accept node 6. So, the partially constructed chromosome will be (1, 2, 6).

The legitimate node after node 6 in P_1 is 9 with c_{69}=5, but none in P_2. So, for P_2, we sequentially search from the beginning of the chromosome and find the first legitimate node 3 with c_{63}=8. Since c_{69}<c_{63}, we accept node 9. So, the partially constructed chromosome will be (1, 2, 6, 9).

The legitimate nodes after node 9 in P_1 and P_2 are 5 and 4 respectively with c_{95}=10 and c_{94}=9. Since c_{94}<c_{95}, we accept node 4. So, the partially constructed chromosome will be (1, 2, 6, 9, 4).

The legitimate node after node 4 in P_1 is 5 with c_{45}=9, but none in P_2. So, for P_2, we sequentially search from the beginning of the chromosome and find the first legitimate node 3 with c_{43}=11. Since c_{45}<c_{43}, we accept node 5. So, the partially constructed chromosome will be (1, 2, 6, 9, 4, 5).

Continuing this way, we obtain the complete offspring chromosome: (1, 2, 6, 9, 4, 5, 7, 8, 3) with cost 72.

7). Bidirectional circular sequential constructive crossover operator: The bidirectional circular sequential constructive crossover (BCSCX) was proposed by Kang et al. [33] to modify SCX that searches for next neighbor in both left and right directions in both parents. Thus, four neighbor genes are considered. Also, during searching for the next neighbor gene, if it reaches to the end or to the beginning of the genes list in any of the parents, it will wrap around. Let us illustrate the BCSCX through the same example given above.

Select 'node 1' as the 1st gene. The legitimate nodes after node 1 in both directions in P_1 are 2 and 8 (after wrapping around), and in P_2 are 3 and 6 (after wrapping around), with their costs 7, 9, 15 and 6 respectively. We accept node 6 as it is cheapest. So, the partially constructed offspring chromosome will be (1, 6).

The legitimate nodes after node 6 in both directions in P_1 are 9 and 4, and in P_2 are 3 (after wrapping around) and 2, with their costs 5, 11, 8 and 13 respectively. We accept node 9 as it is cheapest. So, the partially constructed chromosome will be (1, 6, 9).

The legitimate nodes after node 9 in both directions in P_1 are 5 and 4, and in P_2 are 4 and 8, with their costs 10, 9, 9 and 10 respectively. We accept node 4 and the partially constructed chromosome will be (1, 6, 9, 4).

The legitimate nodes after node 4 in both directions in P_1 are 5 and 3, and in P_2 are 2 and 8, with their costs 9, 11, 5 and 4 respectively. We accept node 8 and the partially constructed chromosome will be (1, 6, 9, 4, 8).

Continuing this way, we obtain the complete offspring chromosome: (1, 6, 9, 4, 8, 2, 7, 3, 5) with cost 56.

Among the above discussed crossover operators GX, SCX and BCSCX are called distance-based crossover operators because they care about the distance between nodes. On the other hand, crossover operators like PMX, OX, AEX, CX, GNX and ERX are called blind crossover operators because they only concern about to satisfy the constraints of the problem and do not use any information associated with the problem [18]. We propose to compare our proposed ASCX against both kind of crossover operators.

C. *Proposed Crossover Operator: Adaptive Sequential Constructive Crossover Operator*

We are going to propose a modification of the SCX operator, named adaptive SCX (ASCX). In BCSCX, four neighbor genes are considered. We propose to construct offspring either in forward direction from the first gene or in backward direction from the last gene or in mixed direction adaptively depending on the cost of the next node. Hence, we consider a total of eight neighbour nodes of a current node, four nodes for each of the two genes (nodes). Since there are n genes in a chromosome, we select 'node 1' as the first and $(n+1)^{th}$ (it is not shown in the chromosome) genes. Let us define the algorithm for the ASCX as follows.

Step 1: Start from the first gene, 'node 1' (i.e., current node p =1 in position i=1) in forward direction and from the $(n+1)^{th}$ gene, ‘node 1’ (it is not shown in the chromosome), (i.e., current node q =1 in position j=n+1) in backward direction.

Step 2: Sequentially search both parent chromosomes in right direction and consider the first ‘legitimate node' (the node that is not yet visited) appeared after 'node p’ in each parent. If no 'legitimate node' after 'node p’ is present in any of the parents, search sequentially from the starting of the parent (wrap around) and consider the first 'legitimate node'. Suppose the 'node α' and the 'node β' are found in 1st and 2nd parent respectively. Go to Step 3.

Step 3: Sequentially search both parent chromosomes in left direction and consider the first ‘legitimate node' appeared after 'node p’ in each parent. If no 'legitimate node' after 'node p’ is present in any of the parents, search sequentially from the end of the parent (wrap around) and consider the first 'legitimate node'. Suppose the 'node γ’ and the 'node δ' are found in 1st and 2nd parent respectively. Now, suppose among four nodes, 'node u' is the cheapest with cost s=min. $\{c_{p\alpha}, c_{p\beta}, c_{p\gamma}, c_{p\delta}\}$. Go to Step 4.

Step 4: Sequentially search both parent chromosomes in left direction and consider the first ‘legitimate node' appeared after 'node q’ in each parent. If no 'legitimate node' after 'node q’ is present in any of the parents, search sequentially from the end of the parent (wrap around) and consider the first 'legitimate node'. Suppose the 'node w' and the 'node x' are found in 1st and 2nd parent respectively. Go to Step 5.

Step 5: Sequentially search both parent chromosomes in right direction and consider the first ‘legitimate node' appeared after 'node q’ in each parent. If no 'legitimate node' after 'node q’ is present in any of the parents, search sequentially from the beginning of the parent (wrap around) and consider the first 'legitimate node'. Suppose the 'node y’ and the 'node z' are found in 1st and 2nd parent respectively. Now, suppose among four nodes, 'node v' is the cheapest with cost t=min. $\{c_{wq}, c_{xq}, c_{yq}, c_{zq}\}$. Now, for selecting the next node as well as adding it in a position in the offspring chromosome go to Step 6.

Step 6: If $s \le t$, then add 'node u' in position ‘i' in the partially constructed offspring chromosome and set p=u, i=i+1. Otherwise, add 'node v' in position ‘j' in the partially constructed offspring chromosome and set q=v, j=j-1. Now, If the offspring is a complete chromosome, then stop, otherwise, go to Step 2.

Let us illustrate the ASCX through the same example parent chromosomes given above. Since there are 9 genes in the parent chromosomes, we select 'node 1' as the first and 10th gene (it is not shown in the chromosome). The legitimate nodes after first gene, node 1, in both directions in P_1 are 2 and 8 (after wrapping around), and in P_2 are 3 and 6 (after wrapping around), with their costs 7, 9, 15 and 6 respectively. Among them node 6 with cost 6 is the cheapest. On the other hand, the legitimate nodes before 10th gene, node 1 (though it is not shown in the chromosome), in both directions in P_1 are 8 and 2 (after wrapping around), and in P_2 are 6 and 3 (after wrapping around), with their costs 9, 7, 6 and 15 respectively. Among them node 6 with cost 6 is the cheapest. Since both cheapest nodes are same 6, we add it as the 2nd gene in the offspring, and hence the partially constructed offspring chromosome will be (1, 6, *, *, *, *, *, *, *).

The legitimate nodes after 2nd gene, node 6, in both directions in P_1 are 9 and 4, and in P_2 are 3 (after wrapping around) and 2, with their costs 5, 11, 8 and 13 respectively. Among them node 9 with cost 5 is the cheapest. On the other hand, the legitimate nodes before 10th gene, node 1, in both directions in P_1 are 8 and 2 (after wrapping around), and in P_2

are 2 and 3 (after wrapping around), with their costs 9, 11, 11 and 15 respectively. Among them node 8 with cost 9 is the cheapest. Since between two cheapest nodes, node 9 is cheaper, we add it as the 3rd gene in the offspring, and hence the partially constructed offspring chromosome will be (1, 6, 9, *, *, *, *, *, *).

The legitimate nodes after 3rd gene, node 9, in both directions in P_1 are 5 and 4, and in P_2 are 4 and 8, with their costs 10, 9, 9 and 10 respectively. Among them node 4 with cost 9 is the cheapest. On the other hand, the legitimate nodes before 10th gene, node 1, in both directions in P_1 are 8 and 2 (after wrapping around), and in P_2 are 2 and 3 (after wrapping around), with their costs 9, 11, 11 and 15 respectively. Among them node 8 with cost 9 is the cheapest. Since both cheapest nodes have same costs, we add node 4 as the 4th gene in the offspring, and hence the partially constructed offspring chromosome will be (1, 6, 9, 4, *, *, *, *, *).

The legitimate nodes after 4th gene, node 4, in both directions in P_1 are 5 and 3, and in P_2 are 2 and 8, with their costs 9, 11, 5 and 4 respectively. Among them node 8 with cost 4 is the cheapest. On the other hand, the legitimate nodes before 10th gene, node 1, in both directions in P_1 are 8 and 2 (after wrapping around), and in P_2 are 2 and 3 (after wrapping around), with their costs 9, 11, 11 and 15 respectively. Among them node 8 with cost 9 is the cheapest. Since between two cheapest nodes, node 8 is cheaper, we add it as the 5th gene in the offspring, and hence the partially constructed offspring chromosome will be (1, 6, 9, 4, 8, *, *, *, *).

The legitimate nodes after 5th gene, node 8, in both directions in P_1 are 2 (after wrapping around) and 7, and in P_2 are 2 and 7, with their costs 3, 8, 3 and 8 respectively. Among them node 2 with cost 3 is the cheapest. On the other hand, the legitimate nodes before 10th gene, node 1, in both directions in P_1 are 7 and 2 (after wrapping around), and in P_2 are 2 and 3 (after wrapping around), with their costs 5, 15, 5 and 15 respectively. Among them node 7 with cost 5 is the cheapest. Since between two cheapest nodes, node 2 is cheaper, we add it as the 6th gene in the offspring, and hence the partially constructed offspring chromosome will be (1, 6, 9, 4, 8, 2, *, *, *).

The legitimate nodes after 6th gene, node 2, in both directions in P_1 are 3 and 7 (after wrapping around), and in P_2 are 3 (after wrapping around) and 7, with their costs 8, 6, 8 and 6 respectively. Among them node 7 with cost 6 is the cheapest. On the other hand, the legitimate nodes before 10th gene, node 1, in both directions in P_1 are 7 and 3 (after wrapping around), and in P_2 are 7 and 3 (after wrapping around), with their costs 5, 15, 5 and 15 respectively. Among them node 7 with cost 5 is the cheapest. Since between two cheapest nodes, node 7 is cheaper, we add it as the 9th gene in the offspring, and hence the partially constructed offspring chromosome will be (1, 6, 9, 4, 8, 2, *, *, 7).

The legitimate nodes after 6th gene, node 2, in both directions in P_1 are 3 and 5 (after wrapping around), and in P_2 are 3 (after wrapping around) and 5, with their costs 8, 11, 8 and 11 respectively. Among them node 3 with cost 8 is the cheapest. On the other hand, the legitimate nodes before 9th gene, node 7, in both directions in P_1 are 5 and 3 (after wrapping around), and in P_2 are 5 and 3 (after wrapping around), with their costs 7, 8, 7 and 8 respectively. Among them node 5 with cost 7 is the cheapest. Since between two cheapest nodes, node 5 is cheaper, we add it as the 8th gene in the offspring, and hence the partially constructed offspring chromosome will be (1, 6, 9, 4, 8, 2, *, 5, 7).

Continuing this way, we obtain the complete offspring chromosome: (1, 6, 9, 4, 8, 2, 3, 5, 7) with cost 59.

D. Mutation Operator

After applying crossover operator, mutation operator is applied. The mutation operator randomly selects a position in the chromosome and changes the corresponding allele (value of a gene), thereby modifying information. The need for mutation comes from the fact that as the less fit members of successive generations are discarded; some aspects of genetic material could be lost forever. By performing occasional random changes in the chromosomes, GAs ensure that new parts of the search space are reached, which selection and crossover alone couldn't fully guarantee. In doing so, mutation ensures that no important features are prematurely lost, thus maintaining the mating pool diversity. For the TSP, the classical mutation operator does not work. For this investigation, we have considered the reciprocal exchange mutation that selects two nodes randomly and swaps them.

III. Design of Our Genetic Algorithms

A simple GA may be summarized as follows:

Step 1: Create initial random population of chromosomes of size Ps and set generation = 0.

Step 2: Evaluate the population.

Step 3: Set generation = generation + 1 and select good chromosomes by selection procedure.

Step 4: Perform crossover with crossover probability P_c.

Step 5: Perform bit-wise mutation with mutation probability P_m.

Step 6: Replace old population with new one.

Step 7: Repeat Steps 2 to 6 until the terminating criterion is satisfied.

As suggested in [18] if the performance of the distance-based crossover is compared with blind crossovers, the comparison is not going to be as fair as it should be. So, we consider both types of crossover operators. There are eight possible selections for crossover operator, which are: PMX, OX, AEX, CX, GX, SCX, BCSCX and ASCX respectively. Within one selection, a single crossover operator is executed.

However, we apply two possibilities of selecting mutation–presence or absence of mutation. There are eight possible selections for crossover operator along with two possibilities of mutation, thus providing altogether sixteen variants of GAs. The goal of such separate execution is to measure effectiveness of specific operator and to find their comparative ranking. Note that each variant GA is purely simple or non-hybrid, which is built of GA procedures and operators, and it does not combine elements of any other heuristic or metaheuristic algorithm. However, GA search process is guided by some parameters, namely, population size that determines number of chromosomes in a population, crossover probability that states the probability of performing crossover between two parent chromosomes, mutation probability that specifies the probability of performing bit-wise mutation, and termination condition that specifies condition to stop the GA search.

IV. Computational Experiences and Discussions

In order to compare the efficiency of the different crossover operators, variant GAs using different crossovers have been encoded in Visual C++ on a Laptop with i3-3217U CPU@1.80 GHz and 4 GB RAM under MS Windows 7, and run for twelve benchmark TSPLIB instances [34]. In these twelve problem instances, the ftv33, ftv38, ft53, kro124p, ftv170, rbg323, rbg358, rbg403 and rbg443 are asymmetric, and gr21, fri26 and dantzig42 are symmetric TSPs. Initial population of chromosomes is generated randomly. The following common parameters are selected for all algorithms: population size is 50, crossover probability is 1.0 (i.e., 100%), mutation probability is 0.09 (i.e., 9%), and maximum of 1,000 generations is the terminating condition. Though GA is structured, yet randomized, so, its repeated execution on the same input data with the same number of procedures usually gives slightly different results. To compensate this randomization effect, the experiments have been repeated 50 times for each instance. The results of experiments by the sixteen GA variants are summarized in Tables II and IV. All tables are organized in the same way: a row corresponds to a problem instance (its best known solution is reported within brackets) and a column to a GA variant considered by a certain selection of crossover operator. Thus, a table entry presents the summary of results of the corresponding instance by the corresponding GA variant. The result is described by its best solution cost, average solution cost, average percentage of excess to the best known solution, standard deviation (S.D.) of costs, and average convergence time (in second). The best result for a chosen instance over all variants is marked by bold face. The percentage of excess above the best known solution, reported in TSPLIB website, is given by the

$$Excess\ (\%) = \frac{Sol.Obtained\ \ - Best\ Known\ Sol.}{Best\ Known\ Sol} \ x\ 100$$

Fig. 1 and Fig. 2 present results for the instance ftv170 (considering only 30 generations). Fig. 1 refers to the GA variants without mutation, and Fig. 2 to the variants with mutation, respectively. In both figures, each graph corresponds to a crossover operator, and it shows how the current solution improves depending on the number of generations. Only the three best performing crossover operators, namely, SCX, BCSCX and ASCX, are reported.

In the figures, the labels on the left margin denote the solution cost, while the labels on the right margin refer to percentage of excess to the best known solution (Excess (%)). All crossover operators have some randomized factors that make them more effective when trying to add an allele. The more randomized these operators are, the more possibilities of progress should have. Fig. 1 shows that SCX has some variations, but it is not the best. Though BCSCX and ASCX have less variations and are competing each other, still ASCX provides us best results. But it has limited variation range and gets stuck in local minimums very quickly. From Fig. 2, it is observed that mutation always improves performance by helping crossovers to escape from local minima.

Table II reports results by the eight GA variants where mutation is not applied. With respect to the average cost, it is very clear from Table II that distance-based crossovers are far better than blind crossovers. Among the crossovers, GX, SCX and BCSCX obtain lowest average cost with lowest S.D for the instances danzig42, gr21 and fr26 respectively. The crossovers SCX and BCSCX are competing. The proposed crossover ASCX obtains lowest average costs with lower S.D. for remaining nine instances, namely, ftv33, ftv38, ft53, kro124p, ftv170, rbg323, rbg358, rbg403 and rbg443. So, among all the crossovers ASCX is found to be the best. Based on best solution costs also ASCX is found to be the best. The results are depicted in Fig. 3, which also shows the effectiveness of our proposed crossover operator ASCX.

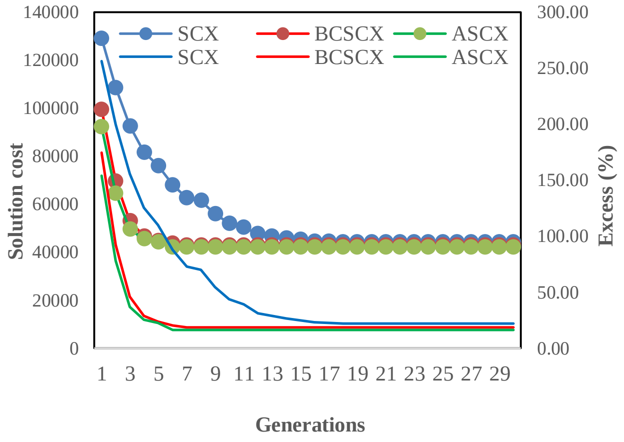


Fig. 1. Performance of Three Crossover Operators without Mutation for the Instance ftv170.

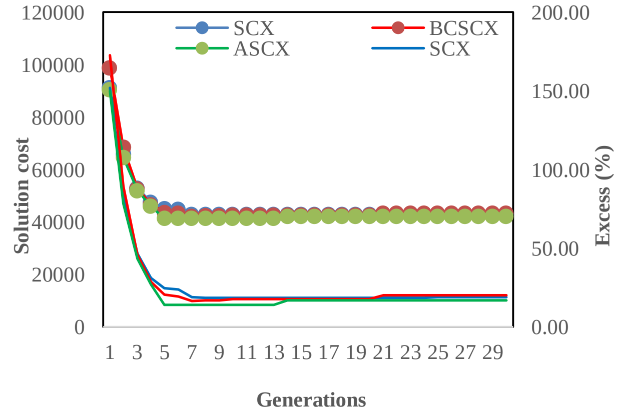


Fig. 2. Performance of Three Crossover Operators with Mutation for the Instance ftv170.

TABLE. II. SUMMARY OF THE RESULTS BY THE VARIANT GAS WITHOUT MUTATION FOR TSPLIB INSTANCES

Instance	**Results**	**PMX**	**OX**	**AEX**	**CX**	**GX**	**SCX**	**BCSCX**	**ASCX**
gr21	Best Sol	3393	2927	3887	5112	3821	2707	2707	2707
(2707)	Avg. Sol	4289.74	3806.40	4462.80	5767.94	4282.04	**2907.20**	2924.26	2916.04
	AvgExc(%)	58.47	40.61	64.86	113.07	58.18	7.40	8.03	7.72
	S.D.	447.10	485.88	339.14	284.30	219.95	112.17	100.03	67.24
	Avg. Time	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
fri26	Best Sol	1193	1014	1100	1799	993	941	937	937
(937)	Avg. Sol	1520.24	1364.70	1305.86	2060.34	1071.26	981.38	**957.70**	959.74
	AvgExc(%)	62.25	45.65	39.37	119.89	14.33	4.74	2.21	2.43
	S.D.	158.59	149.79	91.51	101.83	41.12	32.33	14.51	14.16
	Avg. Time	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ftv33	Best Sol	2236	1722	1843	3269	1604	1380	1420	1378
(1286)	Avg. Sol	2695.08	2352.22	2282.60	3539.30	1770.02	1489.20	1487.62	**1412.68**
	AvgExc(%)	109.57	82.91	77.50	175.22	37.64	15.80	15.68	9.85
	S.D.	231.60	277.64	187.21	122.80	114.42	37.26	36.08	44.42
	Avg. Time	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.01
ftv38	Best Sol	2730	2288	2244	3806	1860	1635	1619	1629
(1530)	Avg. Sol	3281.28	2853.34	2699.46	4267.20	2098.12	1772.80	1720.80	**1707.50**
	AvgExc(%)	114.46	86.49	76.44	178.90	37.13	15.87	12.47	11.60
	S.D.	267.82	324.38	221.03	164.85	119.59	47.61	36.42	32.79
	Avg. Time	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.01
dantzig42	Best Sol	1430	1159	937	2167	699	750	736	699
(699)	Avg. Sol	1786.76	1570.24	1153.74	2425.48	**711.72**	814.34	812.16	746.94
	AvgExc(%)	155.62	124.64	65.06	246.99	1.82	16.50	16.19	6.86
	S.D.	168.28	165.08	91.39	111.80	22.30	32.43	29.54	27.99
	Avg. Time	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.01
ft53	Best Sol	13539	11923	13058	20977	11504	8188	8082	7970
(6905)	Avg. Sol	17059.50	14595.36	14691.82	22342.38	12736.38	8626.44	8611.80	**8472.16**
	AvgExc(%)	147.06	111.37	112.77	223.57	84.45	24.93	24.72	22.70
	S.D.	1541.62	1746.14	856.81	620.64	602.92	239.55	228.45	272.27
	Avg. Time	0.00	0.02	0.01	0.00	0.01	0.03	0.02	0.27
kro124p	Best Sol	109400	84177	122824	148310	97683	41392	41396	40308
(36230)	Avg. Sol	130592.18	108875.06	143543.54	165422.36	107546.14	43789.36	42625.90	**42156.86**
	AvgExc(%)	260.45	200.51	296.20	356.59	196.84	20.86	17.65	16.36
	S.D.	9738.64	10308.85	9821.05	4375.44	4150.34	682.39	568.96	598.17
	Avg. Time	0.01	0.08	0.03	0.01	0.02	0.04	0.09	0.06
ftv170	Best Sol	17932	13271	9501	22545	5245	3696	3255	3258
(2755)	Avg. Sol	19522.98	16231.42	10765.90	23785.48	6158.00	3719.26	3611.54	**3473.52**
	AvgExc(%)	608.64	489.16	290.78	763.36	123.52	35.00	31.09	26.08
	S.D.	1048.53	1629.45	620.84	427.26	319.37	179.77	137.69	112.91
	Avg. Time	0.03	0.25	0.11	0.03	0.07	0.13	0.19	1.44
rbg323	Best Sol	4675	3616	5050	5645	2651	1731	1657	1620
(1326)	Avg. Sol	5014.40	4292.42	5259.02	5797.06	2985.10	1840.80	1747.90	**1689.16**
	AvgExc(%)	278.16	223.71	296.61	337.18	125.12	38.82	31.82	27.39
	S.D.	212.01	366.03	214.99	62.72	141.85	62.96	54.36	27.95
	Avg. Time	0.08	0.91	0.37	0.12	0.50	0.83	2.04	10.33
rbg358	Best Sol	5014	4081	5225	6307	2705	1678	1586	1393
(1163)	Avg. Sol	5562.44	4641.02	5600.40	6481.82	3010.70	1740.04	1713.26	**1453.60**
	AvgExc(%)	378.28	299.06	381.55	457.34	158.87	49.62	47.31	24.99
	S.D.	237.13	350.20	260.00	73.70	167.65	78.91	78.86	26.85
	Avg. Time	0.09	1.33	0.46	0.14	0.65	0.96	2.34	6.18
rbg403	Best Sol	5972	4931	6253	7031	4080	3483	3229	2928
(2465)	Avg. Sol	6346.12	5428.20	6360.18	7215.74	4310.88	3510.74	3403.54	**3012.70**
	AvgExc(%)	157.45	120.21	158.02	192.73	74.88	42.42	38.07	22.22
	S.D.	255.46	346.60	234.33	77.70	102.96	88.64	96.20	36.90
	Avg. Time	0.11	1.95	0.55	0.25	0.80	1.28	3.22	4.84
rbg443	Best Sol	6574	5538	6622	7615	4533	3731	3699	3333
(2720)	Avg. Sol	6933.96	6030.42	7076.30	7816.20	4730.06	3904.62	3881.90	**3404.44**
	AvgExc(%)	154.93	121.71	160.16	187.36	73.90	43.55	42.72	25.16
	S.D.	232.79	417.55	243.32	87.11	104.65	85.01	82.04	38.80
	Avg. Time	0.13	2.31	0.71	0.27	1.09	1.52	4.76	5.16

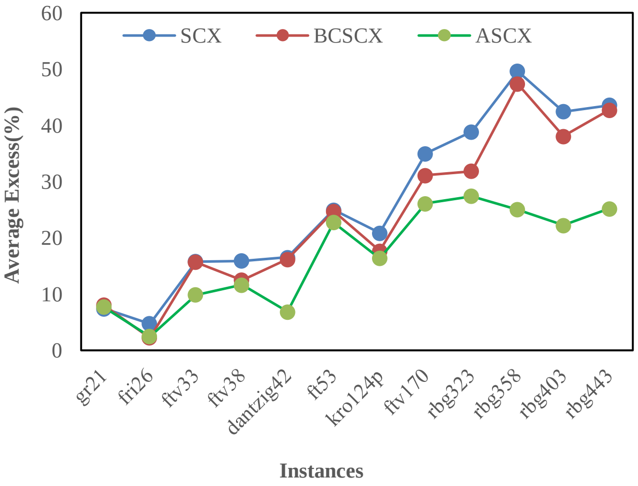


Fig. 3. Average Excess (%) by GA Variants without Mutation.

Among the blind crossovers, OX and AEX are competing. OX obtains lower average solutions for seven instances, namely, gr21, ft53, kro124p, rbg323, rbg358, rbg403 and rbg443, whereas AEX obtains lower average costs for five instances, namely, fri26, ftv33, ftv38, dantzig42 and ftv170. From this observation, one can tell that OX is better than AEX, and PMX and CX show very bad performances.

In order to decide if ASCX-based GA average (without mutation) is significantly different than the averages obtained by other GA variants, we performed Student's t-test. It is to be noted that we performed 50 runs for every problem instance considered here. We used the following t-test for the case of two big independent samples [35]:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{SD_1^2}{n_1 - 1} + \frac{SD_2^2}{n_2 - 1}}}$$

where,

$\bar{X}_1 - average\ of\ first\ sample,$

$SD_1 - standard\ deviation\ of\ first\ sample,$

$\bar{X}_2 - average\ of\ second\ sample,$

$SD_2 - standard\ deviation\ of\ second\ sample,$

$n_1 - first\ sample\ size,$

$n_2 - second\ sample\ size,$

The values of $\bar{X}_2$ and SD_2 are obtained by the ASCX-based GA, while of $\bar{X}_1$ and SD_1 values are obtained by other GA variants. The calculated values of the t statistic are reported in the Table III.

The t values can be positive or negative. The positive value indicates that the ASCX obtained better solution than the competitive GA variant. In the negative case, the competitive algorithm obtained better solution. We used confidence interval at the 95% confidence level ($t_{0.05}$ = 1.96). When t-value is greater than 1.96, the difference between the two values is significant. In this situation, the ASCX solution is better, when t has positive value. Negative t value means that the competitive GA variant has better solution. The case when t-value is less than 1.96, it corresponds to the situation that the difference between the observed values is not significant. The table also reports the information about the GA variants that obtained significantly better solutions.

In the case of three instances there is no statistically significant difference between ASCX and BCSCX. On nine instances ASCX is better than BCSCX. There is no significant difference between ASCX and SCX on one instance only. ASCX performed better than SCX on eleven instances. Next, ASCX performed better than GX on eleven instances, GX performed better than ASCX on one instance only. Finally, when comparing ASCX against blind crossovers, PMX, OX, AEX and CX, we found that ASCX performed better on all twelve instances, but it is not reported.

Table III also reports calculated values of the t statistic of blind crossovers against OX. In the case of one instance, ftv33, there is no statistically significant difference between OX and AEX. On seven instances OX is better than AEX. OX performed better than PMX and CX on all twelve instances.

TABLE. III. THE CALCULATED VALUES OF THE T STATISTIC (VARIANT GAS WITHOUT MUTATION) AND THE INFORMATION ABOUT VARIANT GAS THAT OBTAINED SIGNIFICANTLY BETTER SOLUTIONS

Instance	t-values against OX			t-values against ASCX		
	PMX	AEX	CX	GX	SCX	BCSCX
gr21	5.12	7.75	24.39	41.57	-0.47	0.47
Better	OX	OX	OX	ASCX	-----	----
fri26	4.99	-2.34	26.88	17.94	4.29	-0.70
Better	OX	AEX	OX	ASCX	ASCX	----
ftv33	6.63	-1.45	27.37	20.37	9.23	9.17
Better	OX	---	OX	ASCX	ASCX	ASCX
ftv38	7.12	-2.74	27.20	22.05	7.91	1.90
Better	OX	AEX	OX	ASCX	ASCX	----
dantzig42	6.43	-15.45	30.03	-6.89	11.01	11.22
Better	OX	AEX	OX	GX	ASCX	ASCX
ft53	7.41	0.35	29.26	45.12	2.98	2.75
Better	OX	OX	OX	ASCX	ASCX	ASCX
kro124p	10.72	17.04	35.35	109.16	12.59	3.98
Better	OX	OX	OX	ASCX	ASCX	ASCX
ftv170	11.89	-21.94	31.39	55.47	8.10	5.43
Better	OX	AEX	OX	ASCX	ASCX	ASCX
rbg323	11.95	15.94	28.36	62.75	15.41	6.73
Better	OX	OX	OX	ASCX	ASCX	ASCX
rbg358	15.25	15.40	36.01	64.20	24.06	21.82
Better	OX	OX	OX	ASCX	ASCX	ASCX
rbg403	14.92	15.60	35.23	83.09	36.31	26.55
Better	OX	OX	OX	ASCX	ASCX	ASCX
rbg443	13.23	15.15	29.31	83.14	37.47	36.83
Better	OX	OX	OX	ASCX	ASCX	ASCX

TABLE. IV. SUMMARY OF THE RESULTS BY THE VARIANT GAS WITH MUTATION FOR TSPLIB INSTANCES

Instance	Results	PMX	OX	AEX	CX	GX	SCX	BCSCX	ASCX
gr21	Best Sol	2707	2707	2835	2707	3614	2707	2707	2707
(2707)	Avg. Sol	3122.58	2827.12	3056.16	3201.74	3988.84	2885.36	2900.82	**2826.20**
	AvgExc(%)	15.35	4.44	12.90	18.28	47.35	6.59	7.16	4.40
	S.D.	219.5	90.36	156.58	272.06	169.99	104.16	86.26	61.29
	Avg. Time	0.02	0.04	0.06	0.06	0.05	0.03	0.05	0.04
fri26	Best Sol	953	1165	956	999	955	937	937	937
(937)	Avg. Sol	1103.02	1234.3	993.24	1150.16	1012.08	979.04	**953.44**	954.04
	AvgExc(%)	17.72	31.73	6.00	22.75	8.01	4.49	1.75	1.82
	S.D.	69.85	36.02	25.56	62.95	5.51	24.54	9.37	11.27
	Avg. Time	0.04	0.05	0.07	0.1	0.12	0.04	0.02	0.13
ftv33	Best Sol	1577	1671	1436	1660	1510	1371	1405	1371
(1286)	Avg. Sol	1759.54	2209.82	1604.42	1922.98	1679.9	1474.48	1478.94	**1386.72**
	AvgExc(%)	36.82	71.84	24.76	49.53	30.63	14.66	15.00	7.83
	S.D.	109.81	86.85	62.93	125.13	52.31	49.02	42.04	2.85
	Avg. Time	0.06	0.07	0.13	0.17	0.07	0.04	0.00	0.18
ftv38	Best Sol	1723	2273	1794	2217	1746	1630	1619	1599
(1530)	Avg. Sol	2077.04	2458.48	1971.24	2465.84	1911.48	1705.68	1712.56	**1648.64**
	AvgExc(%)	35.75	60.68	28.84	61.17	24.93	11.48	11.93	7.75
	S.D.	112.98	92.71	84.32	113.34	43.37	30.78	20.85	21.88
	Avg. Time	0.07	0.09	0.16	0.20	0.13	0.10	0.07	0.12
dantzig42	Best Sol	845	1069	827	1170	699	723	725	699
(699)	Avg. Sol	989.80	1108.88	915.18	1297.50	704.18	808.72	810.08	**699.72**
	AvgExc(%)	41.60	58.64	30.93	85.62	0.74	15.70	15.89	0.10
	S.D.	88.68	62.68	31.93	66.57	3.64	30.50	28.83	24.41
	Avg. Time	0.08	0.10	0.18	0.25	0.04	0.06	0.01	0.11
ft53	Best Sol	10027	10597	10299	12629	10109	7678	7848	7631
(6905)	Avg. Sol	11796.86	13902.48	12273.08	13854.44	11144.14	8494.82	8524.50	**8127.34**
	AvgExc(%)	70.85	101.34	77.74	100.64	61.39	23.02	23.45	17.70
	S.D.	701.64	421.58	288.69	588.81	455.60	246.93	183.91	156.51
	Avg. Time	0.09	0.13	0.27	0.40	0.25	0.26	0.31	0.38
kro124p	Best Sol	106539	79811	109251	110833	81824	41331	41668	40246
(36230)	Avg. Sol	117138.20	100806.48	116768.26	120254.10	89253.80	43674.54	42544.46	**41471.58**
	AvgExc(%)	223.32	178.24	222.30	231.92	146.35	20.55	17.43	14.47
	S.D.	3153.95	2264.52	2428.83	2740.37	2557.47	638.43	566.01	432.72
	Avg. Time	0.18	0.43	0.66	1.22	0.57	1.25	0.13	2.22
ftv170	Best Sol	17088	13158	9482	18962	4667	3285	3257	3232
(2755)	Avg. Sol	18689.18	15389.62	10588.86	19630.42	4817.32	3523.74	3608.40	**3393.00**
	AvgExc(%)	578.37	458.61	284.35	612.54	74.86	27.90	30.98	23.16
	S.D.	305.95	282.54	231.40	219.10	71.52	113.55	92.49	95.42
	Avg. Time	0.27	1.15	1.83	3.41	4.17	3.77	3.23	0.93
rbg323	Best Sol	4583	3558	4809	5024	2102	1658	1660	1611
(1326)	Avg. Sol	5006.74	4263.20	5075.84	5150.86	2192.08	1718.76	1725.52	**1618.80**
	AvgExc(%)	277.58	221.51	282.79	288.45	65.32	29.62	30.13	22.08
	S.D.	50.16	42.92	34.15	39.96	31.70	22.19	20.63	17.70
	Avg. Time	0.90	3.05	5.59	10.75	15.51	12.37	24.79	23.53
rbg358	Best Sol	4988	3951	5034	5624	2054	1524	1582	1327
(1163)	Avg. Sol	5428.92	4583.32	4650.54	5740.92	2203.12	1699.20	1711.30	**1387.92**
	AvgExc(%)	366.80	294.09	299.87	393.63	89.43	46.10	47.15	19.34
	S.D.	53.53	41.32	42.26	45.63	52.02	29.22	23.23	24.05
	Avg. Time	1.01	3.56	6.91	12.16	17.28	16.71	30.14	30.77
rbg403	Best Sol	5809	4848	6079	6375	3760	3314	3229	2922
(2465)	Avg. Sol	6219.88	5363.96	6273.98	6543.88	3828.34	3401.18	3479.62	**2983.38**
	AvgExc(%)	152.33	117.60	154.52	165.47	55.31	37.98	41.16	21.03
	S.D.	49.76	37.89	34.63	46.26	39.16	26.48	24.06	21.43
	Avg. Time	1.14	4.48	9.25	17.81	21.43	19.98	33.89	38.93
rbg443	Best Sol	6401	5494	6411	7053	3705	3705	3710	3252
(2720)	Avg. Sol	6893.26	5935.72	6895.30	7123.44	3742.88	3882.52	3872.66	**3321.58**
	AvgExc(%)	153.43	118.23	153.50	161.89	37.61	42.74	42.38	22.12
	S.D.	43.17	44.32	36.77	34.14	22.11	26.78	25.53	20.95
	Avg. Time	1.37	6.35	10.41	18.65	37.87	26.82	42.64	50.82

Table IV reports results by the eight GA variants where mutation is applied. With respect to the average cost, it is once again very clear from the Table IV that distance-based crossovers are superior to blind crossovers. Among the distance-based crossovers, GX is the worst, however, it obtains best cost for danzig42 only. Though SCX could not obtain lowest average cost, but it obtains best costs for the instances gr21 and fr26 at least once in 50 runs. The crossovers SCX and BCSCX are competing. SCX obtains lower average cost for nine instances, whereas BCSCX obtains lower average cost for three instances only. However, BCSCX obtains lowest average solution for the instance fri26 and ASCX obtains lowest average solutions with lower S.D. for eleven instances, namely, gr21, ftv33, ftv38, dantzig42, ft53, kro124p, ftv170, rbg323, rbg358, rbg403 and rbg443. So, among all the crossovers ASCX is found to be the best. Based on best costs also ASCX is found to be the best. The results are depicted in Fig. 4, which also shows the effectiveness of our proposed crossover operator ASCX. So, whether mutation is used or not, the best performance is accomplished by ASCX. However, based on convergence time blind crossovers found to be better than distance-based crossovers, and PMX is the best one.

Among the blind crossovers, CX show very bad performances that obtains lower average solution for no any instance; PMX obtains lower average costs for the instance ft53 only; OX obtains lower average solutions for six instances, namely, gr21, kro124p, rbg323, rbg358, rbg403 and rbg443; whereas AEX obtains lower average solutions for five instances, namely, fri26, ftv33, ftv38, dantzig42 and ftv170. From this observation one can say that OX is the best and CX is the worst. However, CX obtains best solution at least once in 50 runs for gr21.

Here also, in order to decide if ASCX based GA (with mutation) average is significantly different than the averages obtained by other GA variants, we perform Student's t-test and the calculated values of the t statistic are reported in the Table V.

In the case of one instance there is no statistically significant difference between ASCX and BCSCX, and ASCX and GX. On eleven instances ASCX is better than BCSCX and GX. ASCX performed better than SCX on all twelve instances. While comparing SCX against BCSCX (of course, not reported in any table here), we found that on most of the instances there is no statistically significant difference between them, we can treat them statistically equivalent. Finally, when comparing ASCX against blind crossovers, PMX, OX, AEX and CX, we found that ASCX performed better on all twelve instances, except for one instance, gr21, there is no statistically significant difference between ASCX and OX, but it is not reported.

Table V also reports calculated values of the t statistic of blind crossovers against OX. On two instances there is no statistically significant difference between OX and CX. On two instances CX is better than OX, whereas on eight instances OX is better than CX. On five instances PMX is better than OX, whereas on seven instances OX is better than PMX. On six instances OX and AEX are better than each other. So, among the blind crossovers OX is the best and CX is the worst.

Based on the above study it very clear that the proposed crossover operator ASCX is the best, BCSCX and SCX are equivalent and the second-best, and CX is the worst. Among blind crossovers OX is the best. About the performance of blind crossovers same observation is made in [36]. However, in terms of convergent time, PMX is found to be the best.

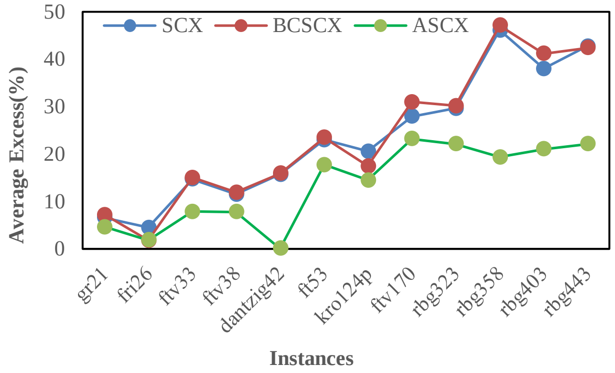


Fig. 4. Average Excess (%) by GA Variants with Mutation.

TABLE. V. THE CALCULATED VALUES OF THE T STATISTIC (CROSSOVERS WITH MUTATION) AND THE INFORMATION ABOUT VARIANT GAS THAT OBTAINED SIGNIFICANTLY BETTER SOLUTIONS

Instance	**t-values against OX**			**t-values against ASCX**		
	PMX	**AEX**	**CX**	**GX**	**SCX**	**BCSCX**
gr21	8.71	8.87	9.15	45.04	3.43	4.94
Better	OX	OX	OX	ASCX	ASCX	ASCX
fri26	-11.69	-38.21	-8.12	32.39	6.48	-0.29
Better	PMX	AEX	CX	ASCX	ASCX	----
ftv33	-22.51	-39.51	-13.18	39.17	12.51	15.32
Better	PMX	AEX	CX	ASCX	ASCX	ASCX
ftv38	-18.27	-27.22	0.35	37.88	10.57	14.80
Better	PMX	AEX	----	ASCX	ASCX	ASCX
dantzig42	-7.68	-19.28	14.44	1.26	19.53	20.45
Better	PMX	AEX	OX	----	ASCX	ASCX
ft53	-18.01	-22.32	-0.46	43.84	8.80	11.51
Better	PMX	AEX	----	ASCX	ASCX	ASCX
kro124p	29.44	33.65	38.29	128.95	19.99	10.54
Better	OX	OX	OX	ASCX	ASCX	ASCX
ftv170	55.46	-92.02	83.03	83.61	6.17	11.35
Better	OX	AEX	OX	ASCX	ASCX	ASCX
rbg323	78.84	103.71	105.96	110.53	24.65	27.48
Better	OX	OX	OX	ASCX	ASCX	ASCX
rbg358	87.53	7.96	131.63	99.57	57.58	67.70
Better	OX	OX	OX	ASCX	ASCX	ASCX
rbg403	95.80	124.10	138.13	132.50	85.85	107.81
Better	OX	OX	OX	ASCX	ASCX	ASCX
rbg443	108.34	116.64	148.61	96.82	115.48	116.81
Better	OX	OX	OX	ASCX	ASCX	ASCX

TABLE. VI. COMPARATIVE STUDY AMONG CX2 [37], HX AND OUR PROPOSED ASCX

Instance	Results	ASCX	CX2	HX
gr21	Best Sol	2707	2995	2707
	Worst Sol	2903	3576	2940
	Avg. Sol	**2790**	3245	2791
fri26	Best Sol	937	1099	937
	Worst Sol	970	1278	1014
	Avg. Sol	**953**	1128	967
ftv33	Best Sol	1341	1811	1397
	Worst Sol	1480	2322	1941
	Avg. Sol	**1393**	2083	1638
ftv38	Best Sol	1598	2252	1755
	Worst Sol	1703	2718	2389
	Avg. Sol	**1653**	2560	2068
dantzig42	Best Sol	699	699	699
	Worst Sol	795	920	985
	Avg. Sol	**703**	802	833
	Best Sol	7803	10987	11234
ft53	Worst Sol	8692	13055	13288
	Avg. Sol	**8192**	12243	12534
kro124p	Best Sol	40607	92450	103988
	Worst Sol	41543	121513	121239
	Avg. Sol	**41473**	101229	110447
ftv170	Best Sol	3244	6421	11215
	Worst Sol	3422	8416	13221
	Avg. Sol	**3384**	7019	10878
rbg323	Best Sol	1501	4212	5175
	Worst Sol	1589	5342	5341
	Avg. Sol	**1557**	4654	5072
rbg358	Best Sol	1339	5404	5560
	Worst Sol	1408	6004	5889
	Avg. Sol	**1394**	5622	5629
	Best Sol	2867	6257	6259
rbg403	Worst Sol	3023	6671	6885
	Avg. Sol	**2965**	6455	6632
	Best Sol	3268	6854	7218
rbg443	Worst Sol	3432	7388	7523
	Avg. Sol	**3356**	6981	7318

Further, we considered results reported in [37] for comparing with our proposed crossover ASCX. Recently Weise et al. [38] made a comparative study among eleven crossover operators for the TSP and found that heuristic crossover (HX) [39] is the best performing operator. The HX applies a greedy heuristic to create an offspring from two parents. So, we implemented the GA using HX and run on the above twelve problem instances. It is to be noted that the same common parameters' values selected for GAs in [37] are used for ASCX and HX. The parameters are as follows: population size, maximum generation, crossover, and mutation probabilities are 150, 500, 0.80, and 0.10, respectively, for less than 100 size instances, whereas population size and maximum generation are 200 and 1000, respectively for more than 100 size instances. Also, the experiments are performed 30 times (30 runs) for each instance. Table VI reports best, worst and average solution costs in 30 runs by ASCX, CX2 and HX.

The crossovers ASCX and HX hit best known solutions for the instances gr21 and fri26, whereas ASCX, CX2 and HX hit best known solution for the instance dantzig42 at least once in 30 runs. In terms of best solution cost, except for the instance dantzig42, ASCX is found better than CX2, and except for gr21, fri26 and dantzig42, ASCX is better than HX. However, in terms of worst and average solution costs ASCX is found to be the best among three crossover operators for all instances. From this study it is very clear that our proposed crossover ASCX is far better than CX2 and HX.

V. CONCLUSION AND FUTURE WORKS

Several crossover operators have been proposed and reported for the TSP by using GAs. We have proposed a new crossover operator, named ASCX, for the TSP. This proposed operator upgrades the SCX and improved the quality of offspring. It is easy to execute and always generates a valid offspring. We focused on some blind crossover operators, namely, PMX, OX, AEX and CX, and distance-based crossover operators, namely, GX, SCX and BCSCX along with ASCX. Firstly, we applied these operators on a pair of chromosomes in manual experiment and found that ASCX performed very good. Then for a significant performance, twelve benchmark instances from the TSPLIB (traveling salesman problem library) have been considered. We developed sixteen variant GAs using crossovers with/without mutation and carried out comparative study of the GAs on the instances. In terms of solution quality, our comparative study showed that distance-based crossovers are far superior than the blind crossovers, and our proposed crossover operator ASCX is the best, BCSCX and SCX are the second-best, and CX is the worst. However, among blind crossovers OX is found to be the best. This observation is verified by Student's t-test at 95% confidence level. Further, we carried out a comparative study among CX2, HX and ASCX, and found that our proposed crossover ASCX is the best. Thus, our proposed operator may be good operator to find more better and accurate results, researchers may use it for other related combinatorial optimization problems.

In this present study, we considered the original version of some crossover operators. Our objective was only to compare the quality of the solutions obtained by the crossover operators, neither to improve the solution quality by any of the operators nor to design the most competitive algorithm for the TSP. So, we neither used any local search technique to enhance the solution quality nor developed parallel version of algorithms to find exact solution. Consequently, we have limited ourselves to simple and pure GA process. Also, we set highest crossover probability to display exact nature of crossover operators. Mutation might be used with lowest probability just not to get stuck in local minima quickly. However, one can incorporate good local search procedure to the hybridize the algorithm, and thus, to solve problem instances exactly, which is under our investigation. Finally, the advantage and helpfulness of the ASCX can be tested on other combinatorial optimization problems.

ACKNOWLEDGMENT

The author thanks the honourable anonymous reviewers for their constructive comments and suggestions which helped the author to improve this paper.

REFERENCES

[1] S. Arora, "Polynomial time approximation schemes for Euclidean traveling salesman and other geometric problems," Journal of ACM, vol. 45, no. 5, pp. 753–782, 1998.

[2] C.P. Ravikumar, "Solving large-scale travelling salesperson problems on parallel machines," Microprocessors and Microsystems, vol. 16, no. 3, pp. 149-158, 1992.

[3] J.D.E. Little, K.G. Murthy, D.W. Sweeny, and C.Karel, "An algorithm for the travelling salesman problem," Operations Research, vol. 11, pp. 972-989, 1963.

[4] M. Padberg, and G. Rinaldi, "Optimization of a 532-node symmetric traveling salesman problem by branch and cut," Operations Research Letter, vol. 6, no. 1, pp. 1–7, 1987.

[5] S.N.N. Pandit, and K. Srinivas, "A lexisearch algorithm for the traveling salesman problem," Proc. IEEE Int. Joint Conf. Neural Networks, vol. 3, pp. 2521–2527, 1991.

[6] Y. Deng, Y. Liu, and D. Zhou, "An improved genetic algorithm with initial population strategy for symmetric TSP," Mathematical Problems in Engineering, vol. 2015, Article ID212794, 6 pages, 2015.

[7] M. Mahi, Ö.K. Baykan, and H. Kodaz, "A new hybrid method based on particle swarm optimization, ant colony optimization and 3-opt algorithms for traveling salesman problem," Applied Soft Computing, vol. 30, pp. 484–490, 2015.

[8] D.E. Goldberg, "Genetic algorithms in search, optimization, and machine learning," Addison-Wesley, New York, 1989.

[9] S.-M. Chen, and C.-Y. Chien, "Solving the traveling salesman problem based on the genetic simulated annealing ant colony system with particle swarm optimization techniques," Expert Systems with Applications, vol. 38, no. 12, pp. 14439–14450,) 2011.

[10] X. Zhou, D. Gao, C. Yang, and H. Gui, "Discrete state transition algorithm for unconstrained integer optimization problems," Neurocomputing, vol. 173, pp. 864–874, 2016.

[11] J. Knox, "Tabu search performance on the symmetric traveling salesman problem," Computer and Operations Research, vol. 21, no. 8, pp. 867–876, 1994.

[12] J.-Y. Potvin, "State-of-the-art survey—the traveling salesman problem: a neural network perspective," ORSA Journal of Computing, vol. 5, no. 4, pp. 328–348, 1993.

[13] M.S. Kıran, H. Iscan, and M. Gündüz, "The analysis of discrete artificial bee colony algorithm with neighborhood operator on traveling salesman problem," Neural Computing and Applications, vol. 23, no. 1, pp. 9–21, 2013.

[14] A. Hatamlou, "Solving travelling salesman problem using black hole algorithm," Soft Computing, vol. 22, no. 24, pp. 8167–8175, 2018.

[15] Z.H. Ahmed, "Algorithms for the quadratic assignment problem," LAP LAMBERT Academic Publishing, Mauritius, 2019, 104 pages.

[16] Z.H. Ahmed, "Genetic algorithm for the traveling salesman problem using sequential constructive crossover operator," International Journal of Biometrics & Bioinformatics, vol. 3, pp. 96-105, 2010.

[17] K. Deb, "Optimization for engineering design: algorithms and examples," Prentice Hall of India Pvt. Ltd., New Delhi, India, 1995.

[18] E. Osaba, R. Carballedo, F. Diaz, E. Onieva, A.D. Masegosa, and A.Perallos, "Good practice proposal for the implementation, presentation, and comparison of metaheuristics for solving routing problems," Neurocomputing, vol. 271, no. 3, pp. 2-8, 2018.

[19] D.E. Goldberg, and R. Lingle, "Alleles, loci and the travelling salesman problem," In J.J. Grefenstette (ed.) Proceedings of the 1st International Conference on Genetic Algorithms and Their Applications. Lawrence Erlbaum Associates, Hilladale, NJ, 1985.

[20] L. Davis, "Job-shop scheduling with genetic algorithms," Proceedings of an International Conference on Genetic Algorithms and Their Applications, pp. 136-140, 1985.

[21] J. Grefenstette, R. Gopal, B. Rosmaita, and D. Gucht, "Genetic algorithms for the traveling salesman problem," In Proceedings of the First International Conference on Genetic Algorithms and Their Applications, (J. J. Grefenstette, Ed.), Lawrence Erlbaum Associates, Mahwah NJ, pp. 160–168, 1985.

[22] I.M. Oliver, D. J. Smith and J.R.C. Holland, "A study of permutation crossover operators on the travelling salesman problem," In J.J. Grefenstette (ed.). Genetic Algorithms and Their Applications: Proceedings of the 2nd International Conference on Genetic Algorithms. Lawrence Erlbaum Associates, Hilladale, NJ, 1987.

[23] Z.H. Ahmed, "Improved genetic algorithms for the traveling salesman problem," International Journal of Process Management and Benchmarking, vol. 4, no. 1, pp. 109-124, 2014.

[24] I. H. Khan, "Assessing different crossover operators for travelling salesman problem," IJISA International Journal of Intelligent Systems and Applications, vol. 7, no. 11, pp. 19-25, 2015.

[25] Z.H. Ahmed, "A hybrid genetic algorithm for the bottleneck traveling salesman problem," ACM Transactions on Embedded Computing Systems, vol. 12, Art. No. 9, 2013.

[26] Z.H. Ahmed, "An experimental study of a hybrid genetic algorithm for the maximum travelling salesman problem," Mathematical Sciences, vol. 7, pp. 1-7, 2013.

[27] Z.H. Ahmed, "The ordered clustered travelling salesman problem: A hybrid genetic algorithm," The Scientific World Journal, vol. 2014, Art ID 258207, 13 pages, 2014.

[28] Z.H. Ahmed, "A simple genetic algorithm using sequential constructive crossover for the quadratic assignment problem," Journal of Scientific and Industrial Research, vol. 73, pp. 763-766, 2014.

[29] Z.H. Ahmed, "Experimental analysis of crossover and mutation operators for the quadratic assignment problem," Annals of Operations Research, vol. 247, pp. 833-851, 2016.

[30] Z.H. Ahmed, "The minimum latency problem: a hybrid genetic algorithm," IJCSNS International Journal of Computer Science and Network Security, vol. 18, no. 11, pp. 153-158, 2018.

[31] Z.H. Ahmed, "Performance analysis of hybrid genetic algorithms for the generalized assignment problem," IJCSNS International Journal of Computer Science and Network Security, vol. 19, no. 9, pp. 216-222, 2019.

[32] M.A. Al-Omeer, and Z.H. Ahmed, "Comparative study of crossover operators for the MTSP," 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, pp. 1-6, 3-4 April 2019.

[33] S. Kang, S.-S. Kim, J.-H. Won, and Y.-M. Kang, "Bidirectional constructive crossover for evolutionary approach to travelling salesman problem," 2015 5th IEEE International Conference on IT Convergence and Security (ICITCS), pp. 1-4, 2015.

[34] G. Reinelt, TSPLIB, http://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/

[35] M. Nikolić, and D. Teodorović, "Empirical study of the bee colony optimization (BCO) algorithm," Expert Systems with Applications, vol. 40, pp. 4609–4620, 2013.

[36] P. Larranaga, C. M. H. Kuijpers, R. H. Murga, I. Inza, and S. Dizdarevic, "Genetic algorithms for the travelling salesman problem: a review of representations and operators," Artificial Intelligence Review, vol. 13, pp. 129–170, 1999.

[37] A. Hussain, Y. S. Muhammad, M. N. Sajid, I. Hussain, A. M. Shoukry, and S. Gani, "Genetic algorithm for traveling salesman problem with modified cycle crossover operator," Computational Intelligence and Neuroscience, vol. 2017, Article ID 7430125, 7 pages, 2017.

[38] T. Weise, Y. Jiang, Q. Qi, and W. Liu, "A branch-and-bound-based crossover operator for the traveling salesman problem," IJCINI International Journal of Cognitive Informatics and Natural Intelligence, vol. 13, no. 3, pp. 1-18, 2019.

[39] J. J. Grefenstette, "Incorporating problem specific knowledge into genetic algorithms," In L. Davis (Ed.), Genetic algorithms and simulated annealing, London, UK: Pitman / Pearson, pp. 42–60, 1987.

Performance Evaluation of Deep Autoencoder Network for Speech Emotion Recognition

Maria AndleebSiddiqui[1]
Computer Science and Information Technology
N.E.D University of Engineering & Technology
Karachi, Pakistan

Syed Abbas Ali[3]
Computer & Information Systems Engineering
N.E.D University of Engineering and Technology
Karachi, Pakistan

Wajahat Hussain[2]
Deputy Manager
Karachi Shipyard and Engineering Works Ltd
Karachi, Pakistan

Danish-ur-Rehman[4]
Electronics Engineering
N.E.D University of Engineering and Technology
Karachi, Pakistan

***Abstract*—The learning methods with multiple levels of representation is called deep learning methods. The composition of simple but now linear modules results in deep-learning model. Deep-learning in near future will have many more success, because it requires very little engineering in hands and it can easily take ample amount of data for computation. In this paper the deep learning network is used to recognize speech emotions. The deep Autoencoder is constructed to learn the speech emotions (Angry, Happy, Neutral, and Sad) of Normal and Autistic Children. Experimental results evident that the categorical classification accuracy of speech is 46.5% and 33.3% for Normal and Autistic children speech respectively. Whereas, Auto encoder shows a very low classification accuracy of 26.1% for only happy emotion and no classification accuracy for Angry, Neutral and Sad emotions.**

Keywords*—*Auto-encoder; emotions; DNN; classification accuracy; autism

I. Introduction

The composition of simple but nonlinear modules results in deep-learning model. The composition is started by representation at one level, which is usually raw input, and then this raw input is transformed into a representation of higher layers, which are slightly more abstract levels [1]. The records of machine learning techniques in many domains of science, especially in image recognition [2-5] and speech recognition [6-8] has been beaten up by deep learning modules. Deep learning has revolutionized the speech signal processing. Excellent results have been achieved using the deep learning networks [9-11]. An Autoencoder is constructed in this paper by stacking two layers; First layer is used to classify the category of speech that is normal or autistic and the second layer is used to classify the emotion of that category that is Angry, Happy, Neutral and Sad. Auto-encoder is a stack of building block. It contains multiple layers of representation [12]. Auto-encoder is also called auto-associator or Diabolo network. It is used to learn a representation of a set of data typically for dimensionality reduction [13]. Comprehensive review was presented in [14] about popular deep learning algorithms for speech emotion recognition. Experimental results shown that the performance of EMO-DB using Log Mel spectrograms on CNN+LSTM is highest that is 78.10%. To improve the Chinese speech emotion recognition, a novel speech emotion recognition algorithm based on stack autoencoder, denoise autoencoder and sparse autoencoder is proposed [15]. The experimental results revealed that the proposed algorithm with stack autoencoder performs 14.3% higher than SVM. The architecture of this algorithm can be studied in Fig. 1.

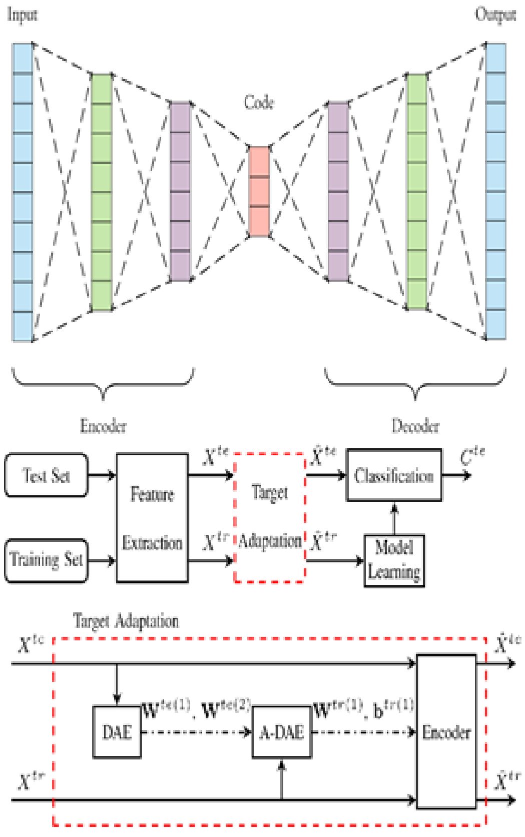


Fig. 1. Architecture of Auto-Encoder.

Here Xte are test samples, Xtr are training samples, W is weight, and b is bias

The representation of weight and bias in neural network is shown in Fig. 2. Architecture Neuron Model is an elementary neuron with "P" inputs as shown. Each input is weighted with an appropriate "w." The sum of the weighted inputs and the bias forms the input to the transfer function "f." Neurons can use any differentiable transfer function "f" to generate their output.

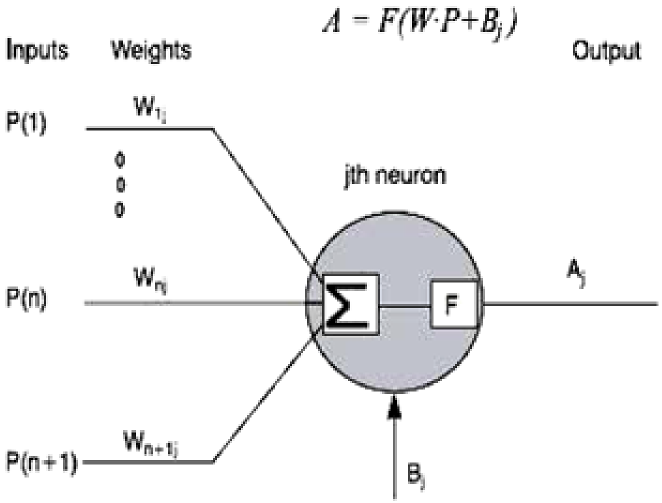


Fig. 2. Architecture of Auto-Encoder.

The basic Algorithm for auto encoder is as follows:

Step 1. Load the training speech data into memory

Step 2. Get the number of columns and rows in each sample

Step 3. Turn the training samples into vectors and put them in a matrix. As the training samples are saved into a matrix, training the network is ready to begin.

Rest of paper is organized into following three sections: Section II presents the methodology of experimental framework. Results of experimental framework with four different emotions are discussed in Section III. Conclusion is drawn in Section IV.

II. Methodology

A. Speech Data Set

The data evaluated in this study were collected from 94 normal and 94 autistic children of age group 10-13 years of both genders. Some Urdu language sentences with four different emotions (Angry, Happy, Neutral and Sad) are chosen for this study. The sentence which is suitable to utter and contain maximum phonetic information is used to implement the speech emotion corpus. The Emotion Corpus consists of 24 samples of each Angry, Happy and Neutral emotions and 22 samples of Sad Emotions. The following ITU recommendations have been used for corpus recording with specifications: SNR>= 45dB and bit rate 24120 bps. Windows 10 built in sound recorder and microphone has been used for recording the speaker's utterances with 48 kHz sampling rate and sensitivity of 56dB ± 25dB. The description of the available speech samples of both normal and autistic children for each emotion is shown in Table I.

B. Training and Configuration of Auto Encoder

The Flow diagram of auto encoder configuration, testing and training phases is presented in Fig. 3.

The training and testing of Autoencoder is discussed in subsequent sections.

1) Create and configure first auto encoder: Sparse auto encoder is trained by using speech training data set without labels. As auto encoder can replicate its input and output, so the size of the input and output will be same. The compressed representation of the input is learned by auto encoder when the size of the input is greater than the number of neurons in hidden layer. By modifying some of the settings of the feed forward neural network, the auto encoder is created.

Step 1. The size of the hidden layer, which is to be trained, is set. It is usually less than the input size.

Step 2. The number of training functions and training epochs are changed to create the network.

TABLE. I. Speech Emotion Data Set

Category	Emotions			
	Angry	*Happy*	*Neutral*	*Sad*
Normal	24	24	24	22
Autistic	24	24	24	22

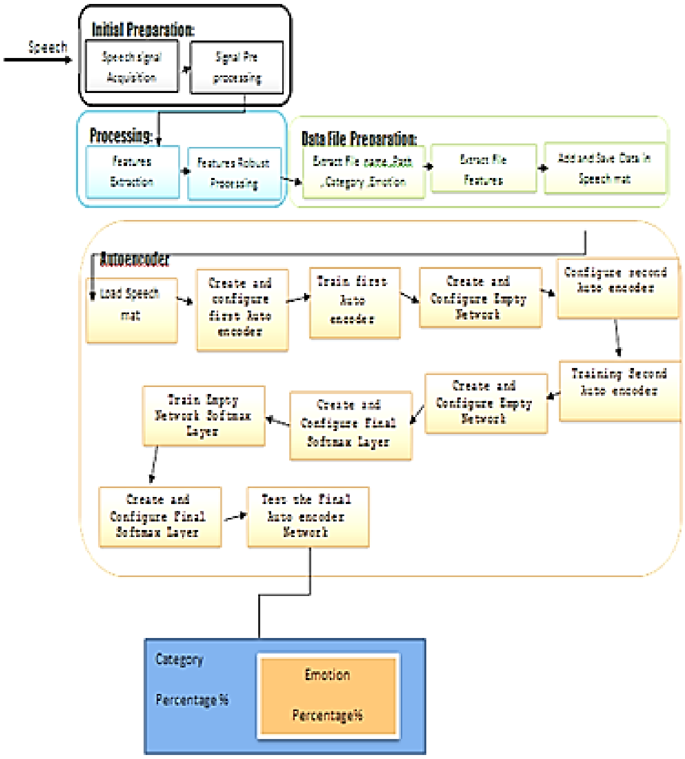

Speech
Initial Preparation:
Speech signal Acquisition
Signal Pre processing
Processing:
Features Extraction
Features Robust Processing
Data File Preparation:
Extract File name, Date, Category, Emotion
Extract File Features
Add and Save Data in Speech mat
Autoencoder
Load Speech mat
Create and configure first Auto encoder
Train first Auto encoder
Create and Configure Empty Network
Configure second Auto encoder
Training Second Auto encoder
Create and Configure Empty Network
Create and Configure Final Softmax Layer
Train Empty Network Softmax Layer
Create and Configure Final Softmax Layer
Test the Final Auto encoder Network
Category
Percentage %
Emotion
Percentage%

Fig. 3. Auto-Encoder Configuration Training and Testing Phases.

Step 3. Process function is not used at the input and output.

Step 4. The transfer function of logistic sigmoid is set for both layers.

Step 5. All the dataset is used for training.

Step 6. The first layer comprises of the sparse representation, which is learned by adding the regularizes as it encouraged the learning power of auto-encoder by using the following parameters:

a) |L2WeightRegularization|: It should be small enough to control the weighing of an L2 regularizers for the weights of the network (not the biases).

b) |sparsityRegularization|: It is used to prevent the large fraction of neurons in hidden layers from activating in response of input.

c) |sparsity|: This function is used to control the fraction of neurons.It is activated in response to the input layer in the 1st layer. Its range is between 0 and 1.

2) Train First Autoencoder

Step 1. Train the auto-encoder with the input data that should be identical to the target data.

Step 2. The auto-encoder diagram is viewed to show the size of hidden layer, input layer, output layer, as well as the transfer function for the two layers.

Step 3. From the first auto encoder the result can be visualized. Visualization helps to get the insight into the feature that can be gained. The hidden layer neurons have the weights vector associated with it in the input layer. Create and Configure empty network.

3) Create and configure empty network: Curls and Stroke patterns from the digit samples are represented by auto-encoder that are seen by the features learned. The compressed version of the input is the 100 dimensional outputs from the hidden layer of auto-encoder. Now the next auto-encoder is trained on the speech training dataset from which the set of the vectors are extracted. For training the next auto-encoder, the first version of the auto-encoder is created with the removed final layer. Removal of first layer is done by manually configuring the settings and creating an empty network object. The biases and weights can be copied from the trained auto-encoder.

Step 1. The empty network is created

Step 2. The number of inputs and outputs are set.

Step 3. The First and only layer is connected to the first input and also to the output.

Step 4. The connection for the bias term to the first layer is added.

Step 5. The size of the input and first layer is set.

Step 6. The first layer uses the logistic sigmoid transfer function.

Step 7. The first layer of trained auto encoder is used to copy the weights and biases.

Step 8. The empty network is seen by the |view| function which is equivalent to the first auto encoder with the first layer removed.

Step 9. To train the second auto-encoder, the features are now generated. This is achieved by evaluating the truncated auto-encoder on the speech training data set.

4) Configure second auto encoder: The second auto-encoder is trained in the similar way as the first auto-encoder. The main difference is that for training the second auto-encoder the speech training data set are the features obtained by hidden layer of the previous auto-encoder. The feed forward neural network is created once again and the settings are modified.

Step 1. The network is created. The number of training function, the size of the hidden layers and the training epochs are changed to conduct the experiment.

Step 2. The process function is not used at the input and output.

Step 3. The transfer function of the logistic sigmoid is set to both the layers

Step 4. All of the data is used for training.

Step 5. After creating the network, performance function is set to |msesparse|, the values of the performance function are set. The sparsity and the mean squared error with L2 weight are used to regularize the performance.

Step 6. The parameters are altered to conduct the experiment.

5) Training second autoencoder: The features generated from the previous auto-encoder are used to train the second auto-encoder.

Step 1. The second auto-encoder is trained.

Step 2. The |view| command is called once again to view the diagram of the autoencoder. The first and Second auto-encoder are similar but the size of the layers is different.

6) Create and configure empty network: As before, a version of the second auto encoder is created with the final layer removed.

Step 1. The number of inputs and layers are set.

Step 2. The first and the only layer are connected to the first input and also connect to the output.

Step 3. A connection for bias term to the first layer is added.

Step 4. The size of the input and first layer is set.

Step 5. The first layer uses the logistic sigmoid function.

Step 6. The first layer of the trained auto-encoder copies the weights and biases.

Step 7. The diagram of the network can be seen by the |view| function. With the last layer removed, it is equivalent to the second auto encoder.

Step 8. The second truncated auto encoder passes the previous set which is used to extract the second set of features.

7) Create and configure final softmax layer: The original vectors in the speech training dataset had 784 dimensions. After passing them through the first auto encoder, this was reduced to 100 dimensions. After using the second auto encoder, this was reduced again to 50 dimensions. The 50 dimensional vectors are classified into different classes to carry out the training of the final layer. A softmax layer is created for the training of the final softmax layer. The output of hidden layer from the second auto encoder is used for its training. As the softmax layer only consists of one layer, it is created manually.

Step 1. Creation of the empty network.

Step 2. The number of inputs and layers are set.

Step 3. The first and the only layer is connected to the first input and also connected to the output.

Step 4. A connection for the bias term to the first layer is added.

Step 5. The size of the input and first layer is set.

Step 6. All of the data is used for training.

Step 7. The cross-entropy performance function is used.

Step 8. The number of training functions and training epochs is changed to conduct the experiment.

8) Train Empty Network Softmax Layer

Step 1. The training of the softmax layer is carried out. Supervised learning is used to train the softmax layer unlike the auto-encoders.

Step 2. |view| command is called to view the diagram of the softmax layer.

Step 3. A multilayer neural network is formed.

Step 4. In isolation, the training of the three separate components of the deep neural network is carried out. To view these three components are useful at these points. They are the networks |autoencHid1|, |autoencHid2|, and |finalSoftmax|.

9) Create and configure final softmax layer: Multilayer neural network is formed to join together these layers. The neural network is created manually, the weights and biases from the auto encoder and softmax layers are copied and the settings are configured after the creation of the network.

Step 1. An empty network is created.

Step 2. One input and three layers are specified.

Step 3. The 1st layer is connected to the input.

Step 4. The 2nd layer is connected to the 1st layer

Step 5. The 3rd layer is connected to the 2nd layer.

Step 6. The output is connected to the 3rd layer.

Step 7. A connection for the bias term to the first layer is added.

Step 8. The size of the input is set.

Step 9. Same as the layer in autoencHid1, the size of the first layer is set.

Step 10. Same as the layer in autoencHid2, the size of the second auto encoder is set.

Step 11. Same as the layer in final softmax layer, the size of the third layer is set.

Step 12. Same as in 1st auto encoders, the transfer function for the first layer is set.

Step 13. Same as in 2nd auto encoder, the transfer function for the second layer is set.

Step 14. Same as in Softmax layer, the transfer function for the third layer is set.

Step 15. Use all of the data for training

Step 16. Copy the weights and biases from the three networks that have already been trained

Step 17. Use the cross-entropy performance function

Step 18. The experiment can be conducted by changing the number of training epochs.

Step 19. |view| command can be used to see the diagram of the multi-layer network.

10)Test the final autoencoder network: The test set is used to compute the results with the full deep neural network. Now, the test samples have to be reshaped, as was done for the training set.

Step 1. The test samples feature set is loaded.

Step 2. Confusion matrix is used to visualized the results. Overall accuracy can be calculated by the numbers in the bottom right hand square of the matrix.

Step 3. The Deep Neural Network is tuned finely. Back propagation application on the whole multi-layer network is used to improve the results for the deep neural network. This process is called fine tuning. Finally the supervised training is used to tune this network by retaining it on the speech training data set.

III. Results

The experimental results in this section are based on Confusion Matrix of categorical classification as shown in Fig. 4.

Categories: Normal and Autistic

Total Samples: 188

Target Samples:

Normal 94 Autistic 94

Output Sample Classes:

Normal:175 Autistic:13

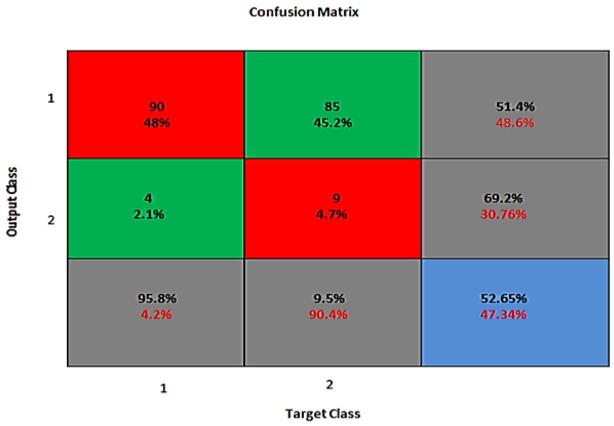


Fig. 4. Confusion Matrix of Categorical Classification.

Here 1= Normal , 2 = Autistic Results

Green are accurate hits per class, Red are error /miss per class.

Our results are 23 Normal hits, missed as Autistic and 20 Autistic hits, missed as Normal.

The Error Histogram is shown in Fig. 5.

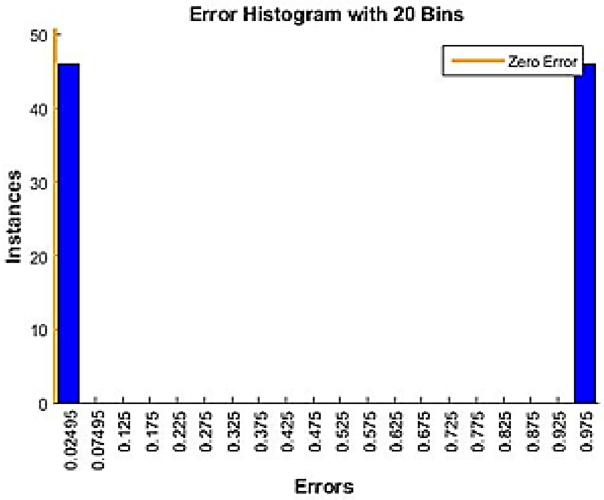


Fig. 5. Error Histogram of Categorical Classification.

Bins are sample for views. The algorithm has 45.7% error due to less number of dataset.

The emotional classification is presented in Fig. 6.

Categories: Normal and Autistic

Total Samples: 94

Target Samples:

Angry :24 Happy :24 Neutral: 24 Sad :22

Output Sample Classes:

Angry : 0 Happy :94 Neutral: 0 Sad: 0

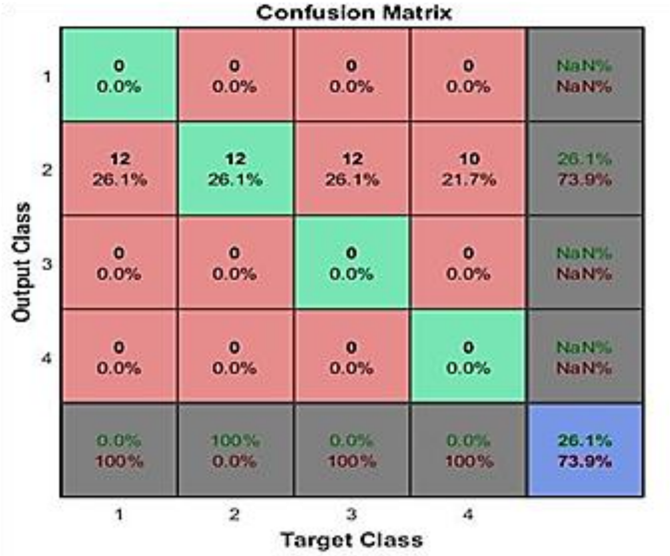


Fig. 6. Confusion Matrix Emotional Classification.

Here 1= Angry, 2 = Happy, 3 = Neutral, 4 =Sad

Green are accurate hits per class, Red are error /miss per class. The result is all Angry, Neutral and Sad are missed as Happy and all Happy are hits. The error histogram is shown in Fig. 7.

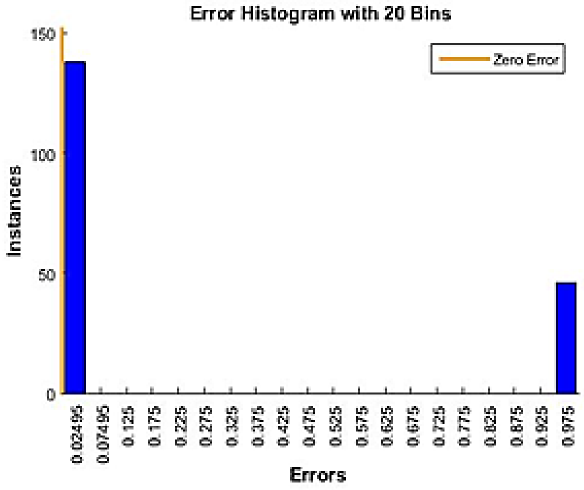


Fig. 7. Error Histogram of Emotional Classification.

IV. Conclusion

This paper evaluated the performance of deep auto encoder for four different emotions of normal and autism children. Experimental frame work were comprised on total 94 speech emotions sample in four different emotions and make used of confusion matrix to demonstrate results in term of classification accuracy. Experimental framework of categorical classification produced overall accuracy of 52.65% and the overall emotional classification of speech produces 26.1% accuracy which shows very low percentage of classification accuracy of emotions. Authors are focusing on improving classification accuracy by increasing the emotions corpus of the children.

REFERENCES

[1] Y.L.Cun , Y. Bengio, G.Hinton, “Review on Deep Learning”, Nature, vol. 521,pp.436-444.2015.

[2] A. Krizhevsky, I. Sutskever, G. Hinton, “ImageNet classification with deep convolutional neural networks”,, In Proc. Advances in Neural Information Processing Systems, vol. 25, pp. 1090–1098, 2012.

[3] C. Farabet, C.Couprie, L.Najman,Y.LeCun, "Learning Hierarchical Features for Scene Labeling," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.35, No.8,pp.1915-1929.2013.

[4] J. Tompson., A.Jain, Y. LeCun, C. Bregler, “Joint training of a convolutional network and a graphical model for human pose estimation”, In Proc. Advances in Neural Information Processing Systems , vol. 27,pp. 1799–1807.2014.

[5] C.Szegedy et al.,“Going deeper with convolutions”, Computer Vision Foundation,pp.1-9.2015.

[6] T. Mikolov ,A. Deoras, D. Povey,L. Burget, J.Cernock, “Strategies for training large scale neural network language models”, In Proc. Automatic Speech Recognition and Understanding,pp.196–201,2011.

[7] G. Hinton, et al., “Deep neural networks for acoustic modeling in speech recognition”. IEEE Signal Processing Magazine, vol. 29,pp.82–97.2012.

[8] T.Sainath, A.R.Mohamed ,B. Kingsbury, B.Ramabhadran, “Deep convolutional neural networks for LVCSR”, In Proc. Acoustics, Speech and Signal Processing , pp.8614–8618, 2013.

[9] A. Graves, N. Jaitley, A.R.Mohamed. “Hybrid Speech Recognition with Deep Bidirsctional LSTM”, in proc. of IEEE workshop on Automatic Speech Recognition and Understanding, pp 273-278, 2013.

[10] K. Han, D. Yu, I.Tashev,“Speech Emotion Recognition using deep neural network and extreme learning machine”, in Proc. of the international speech communication and association conference, pp.223-227, 2014.

[11] H. Lee, L.Yan , P. Dham, A.Y.Ng, “Unsupervised feature learning for audio classification using convolution deep belief networks, Neural Information Processing System, pp.1096-1104.2009.

[12] S.S. Mousavi, M.Schukat, E.Howlay, “Deep Reinforcement Learning: An overview”, In Proc. of SAI Intelligent System Conference (Intellisys), pp. 426-440.2018.

[13] Y.Bengio, “Learning deep architectures for Artificial Intelligence”, Foundations and trends in Machine Learning, vol.2, No.1, pp.1–127. 2009.

[14] S.K.Pandey, H.S Shekhawat, S.R.M Prasanna, "Deep LEarning Techniques for speech emotion recognition: A Review", IEEE Proc. on International Conference on Radioelectronika, Czech Republic, pp:1-5, July 2019.

[15] P.Wei, Y.Zhao, "A novel speech emotion recognition algorithm based on wavelet kernel sparse classifier in stacked deep auto-encoder model", Springer: Pers Ubiquit Comput 23, pp. 521–529, 2019.

Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*

Suryakanthi Tangirala
Faculty of Business, University of Botswana
Gaborone, Botswana

***Abstract*—Decision tree is a supervised machine learning algorithm suitable for solving classification and regression problems. Decision trees are recursively built by applying split conditions at each node that divides the training records into subsets with output variable of same class. The process starts from the root node of the decision tree and progresses by applying split conditions at each non-leaf node resulting into homogenous subsets. However, achieving pure homogenous subsets is not possible. Therefore, the goal at each node is to identify an attribute and a split condition on that attribute that minimizes the mixing of class labels, thus resulting into nearly pure subsets. Several splitting indices were proposed to evaluate the goodness of the split, common ones being GINI index and Information gain. The aim of this study is to conduct an empirical comparison of GINI index and information gain. Classification models are built using decision tree classifier algorithm by applying GINI index and Information gain individually. The classification accuracy of the models is estimated using different metrics such as Confusion matrix, Overall accuracy, Per-class accuracy, Recall and Precision. The results of the study show that, regardless of whether the dataset is balanced or imbalanced, the classification models built by applying the two different splitting indices GINI index and information gain give same accuracy. In other words, choice of splitting indices has no impact on performance of the decision tree classifier algorithm.**

Keywords—Supervised learning; classification; decision tree; information gain; GINI index

I. INTRODUCTION

Machine learning problems can be broadly classified into two categories viz. supervised learning and unsupervised learning as shown in Fig. 1. With supervised learning techniques, the training data is labeled. It means each observation in the data set has both descriptive variables (i.e., independent variables or decision variables) and a labeled outcome variable. Labels can be either categories or continuous values [1]. With supervised learning, a labeled data set is used to train the model in making predictions. A learning model maps the input variables to the output variable, with the aim of accurately predicting the output for future input variables.

Unlike supervised learning, with unsupervised learning the data is not labeled. This means that the training data has descriptive variables only and no outcome variable. The model has to determine the patterns and interesting structures in the data that are not known beforehand [2].

Classification is a supervised learning problem, where the objective is to analyse the training data and develop a model that can predict the future behavior, here the training dataset is labeled. Decision tree algorithm is commonly used for classification tasks. Decision trees classify data into finite number of classes based on the values of input variables. It is most appropriate for categorical data [3].

Decision tree is a simple flowchart that selects class labels of an output variable using the values of one or more input variables. The classification process starts at the root node of the decision tree and recursively progresses until it reaches the leaf node with class labels. At each node a split condition is applied to decide whether the input value should continue towards left or right sub tree until it reaches the leaf nodes [4]. The split condition applied at each node should result in homogenous subsets. Homogenous subsets have records with same class label. However, it is impossible to achieve pure homogenous subsets with real time data. Some kind of mixing will always be there. Therefore, while building the decision tree, the goal at each node is to select split conditions that best divide the dataset into homogenous subsets. The "goodness of split criterion" was introduced, which is derived from the notion of impurity [5]. Impurity is measured mathematically for each split condition and split condition with lowest impurity value is chosen.

To measure the impurity value of a split condition several indices are proposed viz., GINI index, Information gain, gain ratio and misclassification rate. This paper empirically examines the effect of GINI index and Information gain on classification task. The classification accuracy is measured to check the suitability of the models in making good predictions.

Rest of the paper is organised as follows: Section II introduces the theoretical notions of Information gain and GINI index. Section III is literature review. Sections IV and V gives the details of data and experimental procedure to compare Information gain and GINI index on balanced and imbalanced data set along with results obtained, and Section VI summarizes the results of the study.

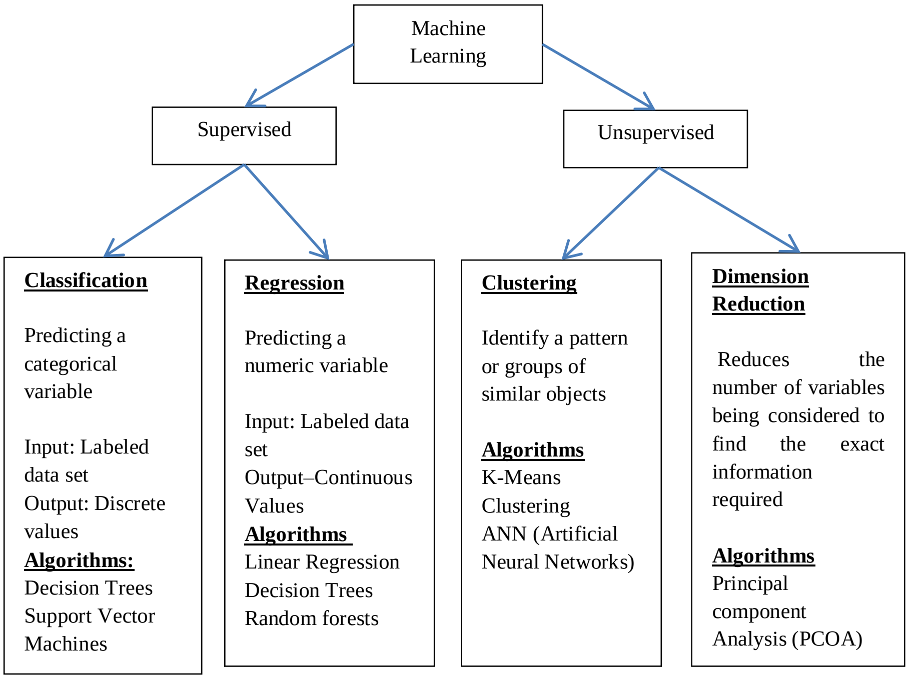


Fig. 1. Broad Classification of Machine Learning Techniques.

II. Theoritical Notation

This section briefly discusses theoretical notions of Information gain and GINI index. Raileanu and Stoffel [6] presented theoretical comparison of GINI index and Information gain.

Let L be a learning sample, L= $\{(x_1, c_1), (x_2, c_2) \dots (x_i, c_j)\}$; Where $x_1, x_2 \dots x_i$ is a measurement vector and $c_1, c_2 \dots c_j$ are class labels. x_i can be viewed as a vector of input variables, and split conditions are based on one of these variables. If p_i is probability that an arbitrary tuple belongs to class c_i, p_i can be measured as

$$p_i = \frac{C_i}{L}$$

A. Entropy

Information gain is based on Entropy. Entropy measures the extent of impurity or randomness in a dataset [7]. If the observations of subsets of a dataset are homogenous, then there is no impurity or randomness in the dataset. If all the observations of subsets belong to one class, the entropy of that dataset would be 0. Entropy is defined as the sum of the probability of each label times the log probability of that same label.

$$Entropy(L) = \langle C_1|L\rangle \log_2\langle C_1|L\rangle + \langle C_2|L\rangle \log_2\langle C_2|L\rangle + \dots + \langle C_j|L\rangle \log_2\langle C_j|L\rangle$$

$$Entropy(L) = p_1 \log_2 p_1 + p_2 \log_2 p_2 + \cdots + p_j \log_2 p_j$$

$$Entropy(L) = -\sum_{i=1}^{j} p_i \log_2(p_i)$$

For a dataset with one class label, p_i will be 1 and $\log_2(p_i)$ is 0. Hence the Entropy of homogenous data set is zero [8]. If the entropy is higher the uncertainty/impurity/mixing is higher [9].

B. Information Gain

Information gain is based on Entropy. Information gain is the difference between Entropy of a class and conditional entropy of the class and the selected feature. It measures the usefulness of a feature f in classification [10] i.e., the difference in Entropy from before to after the split of set L on a feature f. In other words, it measures the reduction of uncertainty after splitting the set on a feature. If information gain value increases, it means the feature f is more useful for classification. The feature with highest information gain is the best feature to be selected for split. Assuming that there are V different values for a feature f, $|L^v|$ represents the subset of L with f=v, Information gain after splitting L on a feature f is measured as [8].

$$IG\ (L, f) = Entropy\ (L)\ - \textstyle\sum_{v=1}^{V} \frac{|L^V|}{|L|}\ (Entropy\ (L^V))$$

C. GINI Index

GINI index determines the purity of a specific class after splitting along a particular attribute. The best split increases the purity of the sets resulting from the split. If L is a dataset with j different class labels, GINI is defined [3] as

$$\mathrm{GINI(L)} = 1 - \sum_{\mathrm{i=1}}^{\mathrm{j}} \mathrm{p_i}^2$$

Where pi is relative frequency if class i in L. If the dataset is split on attribute A into two subsets L1 and L2 with sizes N1 and N2 respectively, GINI is calculated as

$$\mathrm{GINI}_A(L) = \frac{N_1}{N}\mathrm{GINI}(L_1) + \frac{N_2}{N}\mathrm{GINI}(L_2)$$

Reduction in impurity is calculated as

$$\Delta\mathrm{GINI}(A) = \mathrm{GINI}(L) - \mathrm{GINI}_A(L)$$

III. Literature Review

This section briefly presents some of the empirical studies that compared the performance of decision tree algorithms which use different impurity metrics for feature selection at non-leaf nodes. An attempt is made to find out if the choice of these feature selection metrics has any impact on the accuracy of the model from past studies.

Mingers [11] tested different feature selection measures empirically, and reported that choice of the feature selection measure affects the size of the tree but not its accuracy. The accuracy remained the same even when attributes are randomly selected. Patil [12] studied the two decision tree based classification algorithms C5.0 and CART. C5.0 uses information gain and CART algorithm uses GINI index to select the features for split conditions. Their study was an experiment to compare C5.0 and CART classification algorithms to classify if a customer qualifies for membership card or not. The study revealed that C5.0 gives higher classification accuracy of 99.6% than CART algorithm with 94.8% accuracy.

A study empirically compared different feature selection measures and proposed a variant of GINI index which uses GINI index ratios for feature selection. In this study they compared the classification accuracy of modified GINI with other classification algorithms ID3, C4.5 and GINI. The results show that ID3 and C4.5 based on Information gain have low classification and prediction accuracy than GINI index and modified GINI index. Modified GINI index is reported to obtain the highest accuracy among all algorithms that were compared [13]. Adhatrao et.al [14] present experiments to compare the performance of two decision tree algorithms, ID3 and C4.5 in predicting the performance of first year engineering students based on the performance achieved by old students who are now in second year engineering. The results show that both the algorithms give same accuracy. In a study Hssina, et.al [15] compared different decision tree algorithms viz. ID3, C4.5, C5, CART and the results reported show that C4.5 has achieved the highest classification accuracy. C4.5 uses information gain to evaluate goodness of split.

Above discussed studies give varied results on the performance of Information gain and GINI index. Moreover, the empirical studies compared the models that were built using different tree based algorithms. These algorithms differ in splitting attribute selection, number of splits (binary /ternary), order of splitting attribute (splitting the same attribute only once or multiple times), stopping criteria and pruning technique (pre/post) [14]. All these factors contribute to the performance of the models built using these algorithms.

The present study is unique as it focuses only on finding the impact of GINI index and Information gain on classification. Therefore, unlike other studies, this study develops classification models using single algorithm called decision tree classifier on which GINI index and information gain are applied individually. This neutralizes the impact of all other factors on models.

IV. Experimental Setup

This section gives the details of data and experimental procedure.

A. Dataset Description

The experiment is conducted using real data provided by UCI Machine Learning repository [16]. The data was collected by Portuguese banking institution by making phone calls to customers. The dataset is relatively a large dataset with 41187 rows and 21 columns. One input variable, ‘duration’ is discarded, as it is highly multi valued and should be avoided for good prediction. Details of the remaining variables are given in Table I. The classification goal is to predict whether customer will subscribe for a term deposit (y) based on remaining 19 input variables. The dataset is clean; it doesn’t have Null values. Term deposit (y) is the outcome variable with two class labels (yes or no). Therefore, it is a binary classification problem.

TABLE. I. Description of the Dataset

Variable	***Description***	***Type***
age	Age of the customer	numeric
job	Type of job of customer	categorical
marital	marital status	categorical
education	Educational qualification	categorical
default	Has credit in default	categorical
housing	Has housing loan	categorical
loan	Has personal loan	categorical
contact	Contact communication type (cell, telephone)	categorical
month	Last contact month of the year	categorical
day_of_ week	Last contact day of the week	categorical
campaign	number of contacts performed during this campaign and for this client	numeric
pdays	number of days that passed by after the client was last contacted from a previous campaign	numeric
previous	number of contacts performed before this campaign and for this client	numeric
poutcome	outcome of the previous marketing campaign	categorical
emp.var.rate	employment variation rate - quarterly indicator	numeric
cons.price.index	consumer price index - monthly indicator	numeric
cons.conf.index	consumer confidence index - monthly indicator	numeric
euribor3m	euribor 3 month rate - daily indicator	numeric
nr.employed	number of employees - quarterly indicator	numeric
y (outcome variable)	has the client subscribed a term deposit? (binary: 'yes','no') (Yes=1, No=0)	categorical

When developing a decision tree, the goal at each node is to identify the attribute and a split condition of the attribute that best divides the training set into pure subsets at that node [17].

Given a dataset with input variables and an outcome variable with a class label, the decision tree algorithm recursively divides the training set until each division contains examples of same class label. If all the observations of the division belong to one class, then it is homogenous subset and if they belong to multiple classes it is impure or heterogeneous [18]. To evaluate the goodness of the split, two splitting indices, GINI index and Information gain are used. Both GINI index and Information gain are applied on Decision tree classifier algorithm and models are developed.

The dataset is split into two parts, training and test. The general practice is to divide the dataset into 80:20 ratios, 80 % training data and 20% test data (unseen data). Using the decision tree classifier algorithm, a classification model built recursively from the training data, dividing the data until each division is pure (homogenous class) and then its prediction accuracy is tested on the unseen test data. In this experiment, the classification model is trained to predict whether customers would subscribe for a term deposit (Yes or No) using the 19 input variables.

A k-fold cross validation method minimizes the bias associated with random sampling of the training and hold out of data samples while comparing the predictive accuracy of two or more methods [3]. In our experiment classification model is trained and tested 10 times where the training set is split into 10 exclusive subsets of equal size and each time, the model is trained on all 9 leaving 1 subset which will be used for testing. Overall accuracy is simply average of the 10 individual accuracies obtained.

B. *Decision Tree Classifier*

Many algorithms have been proposed for creating decision trees. In this experiment, Decision tree classifier, a supervised learning algorithm is used. It is based on CART and can be used for creating both classification and regression trees [19]. *rpart* is a package in *R* programming, which implements many of the ideas found in CART model. Different splitting criterions can be applied while splitting the nodes of the tree using rpart function [20]. The classification models built by applying Information gain and GINI index are shown in Fig. 2 and Fig. 3, respectively.

It is noted that both the splitting measures select the same feature, 'Number of employees' with same split condition at the root node. 'Number of employees' which is a numeric attribute is selected with split condition nr.employees >=5088.

C. *Performance Evaluation Metrics*

Classification is technique where the model is developed using a labeled dataset. It means each record in the training dataset has a class label associated with it. The model is later used to predict the class labels of new/unseen data. Predictive accuracy of classification model is its ability to correctly predict the class label of an unseen data. The common metrics for measuring the accuracy of classification models are confusion matrix, overall accuracy, per-class accuracy, recall and precision [3] [21]. First confusion matrix is created using which all other metrics are easily calculated.

- Confusion matrix

Confusion matrix gives detailed view of the performance with breakdown of correct and incorrect predictions for each class. The performance is measured by comparing the predicted outcome values with actual values. The information is tabulated in the form of a confusion matrix as shown in Table II.

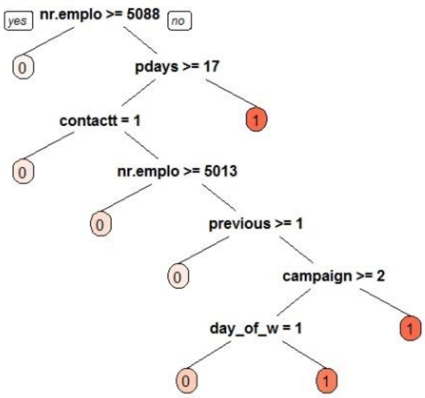


Fig. 2. Decision Tree Visualization using Information Gain.

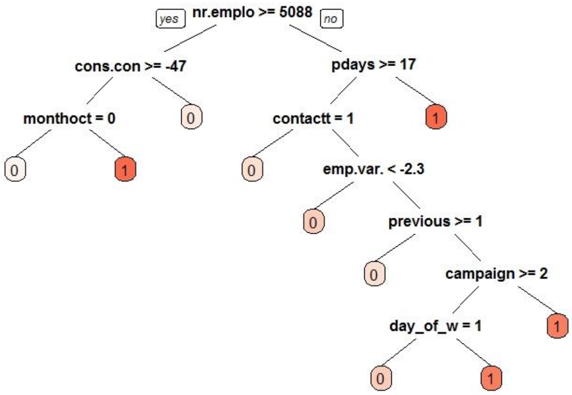


Fig. 3. Decision Tree Visualization using GINI Index.

TABLE. II. CONFUSION MATRIX

Actual			
		Positive	Negative
Predicted class	Positive	True positive count (TP)	False Positive count (FP)
	Negative	False Negative count (FN)	True Negative count (TN)

where True positives (TP) corresponds to the number of positive examples correctly predicted by the model, False negatives(FN) represents number of positive examples wrongly predicted as negative, False positive(FP) refers to number of negative examples wrongly predicted as positive and True negative (TN) is number of negative examples correctly predicted [22]

- Overall Accuracy

Overall classifier accuracy is the rate at which the model makes accurate predictions. It is the ratio of number of correct predictions to total number of predictions made.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

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

- Per-class accuracy

Per class accuracy gives the average of accuracy of prediction of each class. It is particularly useful when the data sets are imbalanced. Overall accuracy is micro average and per-class accuracy is macro average.

$$\text{Per class accuracy} = \frac{\text{Number of correct predictions of that class}}{\text{Total count of predictions of that class}}$$

$$\text{Majority (positive) class accuracy} = \frac{\text{TP}}{(\text{Sensitivity})\text{TP} + \text{FN}}$$

$$\text{Majority (Negative) class accuracy} = \frac{\text{TN}}{(\text{Sensitivity})\text{TN} + \text{FP}}$$

- Precision is defined as the ratio of correctly classified majority class values (True positives) divided by sum of correctly classified majority class values (True positives) and incorrectly classified majority class values (False positive). It should be high.

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

- Recall is defined as the ratio of correctly classified majority class values (True positives) divided by sum of correctly classified majority class values (True positives) and incorrectly classified minority class values (False Negatives). Recall estimates the classifiers accuracy in predicting the majority class. It should be high.

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

D. Performance Evaluation on the Test Set

The test set has a total of 8237 observations. Confusion matrix of Decision tree classifier with Information gain and GINI Index are shown in Table III and Table IV. Positive/majority class is represented as 0 negative/minority class is represented using 1.

TABLE. III. CONFUSION MATRIX OF CLASSIFICATION RESULTS OBTAINED BY DECISION TREE CLASSIFIER WITH INFORMATION GAIN

	Actual			Total
Predicted		0	1	
	0	7198	718	7916
	1	119	202	321
Total		7317	920	8237

TABLE. IV. CONFUSION MATRIX OF CLASSIFICATION RESULTS OBTAINED BY DECISION TREE CLASSIFIER WITH GINI INDEX

	Actual			Total
Predicted		0	1	
	0	7190	709	7899
	1	127	211	338
Total		7317	920	8237

TABLE. V. RESULTS OF OTHER PERFORMANCE EVALUATION METHODS

Methods	Overall classifier Accuracy	Majority class accuracy	Minority class accuracy	Recall (sensitivity)	Precision
Information gain	89.84	98.3	21.9	98.3	90.9
GINI Index	89.85	98.2	22.9	98.2	91.0

Accuracy, recall, precision and F1 score values are shown in Table V.

Results in Table V, quite clearly show that there is no significant difference between the classification accuracy obtained by the two feature selection measures. Overall accuracy as well as per class accuracy values remain approximately the same. Other observations are in line with literature which says, classifiers trained on low dimensional, imbalanced data classify most of the samples to majority class [23]. Therefore, it is deceivingly simple to achieve high overall accuracy, although it is difficult to classify the data reliably. This is evident from the results obtained, where the majority class accuracy is too high (98.3%) when compared to minority class accuracy (22% approx.). With imbalanced data set, even when the minority class accuracy is very low, the overall accuracy would be high because of high True positive count as in our case. Hence, kappa statistic is measured which takes in to account the chance agreement.

- Kappa Coefficient:

Kappa coefficient is an interesting alternative to measure the accuracy of classifier models. It is particularly useful when the data sets are imbalanced [24]. It is used to quantify the reproducibility of discrete variable.

Originally Cohen's Kappa(κ) coefficient was introduced to measure the level of inter-observer agreement, its value ranging from 0 to 1 [25]. If κ is 0 then the agreement between observed and expected is only by chance; if it 1, it is a perfect agreement. κ value between 0 and 0.2 indicates slight agreement, 0.2 to 0.4 says fair agreement, 0.6 to 0.8 is substantial agreement. [26]. The Kappa (κ) statistic takes into account the chance agreement and is defined as.

$$\text{Kappa}\ (\kappa) = \frac{\text{Observed Agreement} - \text{Expected Agreement}}{1 -- \ \text{Expected Agreement}}$$

Kappa coefficient is used to evaluate the accuracy of models by measuring agreement between predicted values and true values. Using the confusion matrix in Table III and Table IV, kappa values for the classifiers are generated as

Kappa value of the classifier model based on Information gain, Kappa (κ) =

$$\frac{\big((7198) + (202)\big|(8237)\big) - \big((7916)(7317) + (321)(920)\ \big|(8237)^2\big)}{1 - \big((7916)(7317) + (321)(920)\ \big|(8237)^2\big)}$$
=0.284

Kappa value 0.28 indicates that observed agreement is 28% of the way between chance and perfect agreement.

Kappa value of the classifier model based on GINI index,

Kappa (κ) =

$$\frac{\big((7190) + (211)\big|(8237)\big) - \big((7899)(7317) + (338)(920)\ \big|(8237)^2\big)}{1 - \big((7899)(7317) + (338)(920)\ \big|(8237)^2\big)}$$
=0.293

Kappa value 0.29 indicates that observed agreement is 29% of the way between chance and perfect agreement.

It is clearly evident from the results obtained that both the classifier models obtained near to equal results. In other words, the results clearly show that the classification accuracy of decision trees is not sensitive to choice of feature selection measures.

High overall accuracy (89% approx.) and very low minority class accuracy (22%) show that the data is not classified reliably. This could be because the dataset used in the experiment is highly imbalanced with 29231 positive (majority) samples and 3719 negative (minority) samples. In next section we provide the details of methods for balancing the dataset and discuss the results of the experiment conducted after balancing the dataset.

V. Balancing the Dataset

Imbalanced datasets have imbalanced class distribution; where by more observations belong to one class than other. Classification algorithms suffer from the problem of imbalanced dataset which leads to biases and poor generalizations. Sometimes, in real world applications, minority class would be of most interest and classifying them correctly should be given high importance, allowing small error rate in classification of majority class since the cost of misclassifying them could be relatively very [27].

For a binary classification problem, if S is the training data, y is the response variable, [28] defines imbalanced classification problem as follows:

S = $\{(x_1, y_1) \ldots (x_m, y_m)\}$, where $y_i \in \{-1, 1\}$ will be data labels.

$S^+ = \{(x, y) \in S: y = 1\}$ be the positive or minority instances.

$S^- = \{(x, y) \in S: y = -1\}$ be the negative or majority instances.

In the test set if, $|S^+| > |S^-|$, the performance of classification algorithm will be very poor, and misclassification rate will be high especially when it comes to the minority class. Therefore, to improve the performance, resampling methods are applied on the training dataset to generate a new set E with synthetic instances of minority class, transforming the training dataset into, $S = (S^+ \cup E) \cup S^-$

A. *Resampling*

Imbalanced datasets have imbalanced class distribution. The dataset used for the study is imbalanced with 29231 positive samples and 3719 negative samples. In such situations, it is difficult to classify the data reliably, although it is simple to attain high accuracy. It is quite essential to balance the dataset to classify reliably. Distribution of classes can be balanced by random oversampling minority class observations or random under sampling majority class observations or by combining both over and under in a systematic manner [29]. Random oversampling creates the problem of over fitting the classifiers and under sampling suffers from loss of useful observations. Another heuristic method, SMOTE (Synthetic Minority Oversampling Technique) based on oversampling is widely used which reduces the over fitting to certain extent and performs better than random over sampling. SMOTE generates synthetic observations of minority class [27] [23].

Before applying any of the resampling techniques training and test data must be split to avoid over fitting and poor generalizations. After resampling we have nearly equal ratio of observations for each class in the training set. The number of observations after applying the resampling methods on the training set can be seen in Table VI.

B. *Results: Performance Evaluation after Resampling*

After balancing the dataset with resampling techniques, the experiment described in section IV is repeated and accuracy is measured. Confusion matrix created after applying resampling techniques is shown in Table VII.

TABLE. VI. Number of Observations after Applying Resampling Techniques

Dataset	Number of features	Training set size	Number of positive samples	Number of Negative samples	Imbalance ratio
Original	20	32950	29231	3719	89:11
Over	20	58462	29231	29231	Equal
Under	20	7438	3719	3719	Equal
Both	20	32950	16556	16394	50.2 : 49.8
SMOTE	20	26033	14876	11157	57 : 43

TABLE. VII. CONFUSION MATRIX WITH DIFFERENT RESAMPLING TECHNIQUES

	OVER			UNDER			BOTH			SMOTE		
Information gain		0	1		0	1		0	1		0	1
	0	5858	311	0	6122	332	0	6041	322	0	6720	459
	1	1459	609	1	1195	588	1	1276	598	1	597	461
GINI index		0	1		0	1		0	1		0	1
	0	5858	311	0	6139	339	0	6064	330	0	6720	459
	1	1459	609	1	1178	581	1	1253	590	1	597	461

Tables VIII and IX summarizes the results obtained by the classification models after applying different resampling techniques. The results in the tables show that balancing the data set has decreased the majority class accuracy but improved the minority class accuracy. Balancing the data set has improved the minority class accuracy by increasing the count of true negative. As discussed earlier it is relatively simple to achieve high overall accuracy with imbalanced data sets, but classifying data reliably is difficult. Thus, after balancing the dataset the objective of classifying data reliably is achieved as the minority class accuracy has improved.

TABLE. VIII. RESULTS OBTAINED WITH DIFFERENT RESAMPLING TECHNIQUES USING INFORMATION GAIN

Metric	Overall Accuracy	Majority class accuracy	Minority class accuracy	Recall	Precision	Kappa
Over	78.5	80.0	66.2	80.0	94.9	29.9
Under	81.4	83.6	63.9	83.6	94.8	33.7
Both	80.6	82.5	65	82.5	94.9	32.7
SMOTE	87.18	91.8	50.1	91.8	93.6	39.3

TABLE. IX. RESULTS OBTAINED WITH DIFFERENT RESAMPLING TECHNIQUES USING INFORMATION GAIN

Metric	Overall Accuracy	Majority class accuracy	Minority class accuracy	Recall	Precision	Kappa
Over	78.5	80.0	66.2	80.0	94.9	29.9
Under	81.5	83.9	63.1	83.9	94.7	33.6
Both	80.7	82.8	64.1	82.8	94.8	32.6
SMOTE	87.18	91.8	50.1	91.8	93.6	39.3

Further analysis of results show that, SMOTE has achieved highest overall accuracy among all the resampling methods. Also, with Smote technique kappa value is 39%. It shows that SMOTE technique is relatively more reliable technique for balancing the dataset than other three methods studied.

VI. CONCLUSIONS

The empirical results reported in this paper show that both Information gain and GINI index produce the same accuracy for classification problems. The experiment is conducted before and after the data set is balanced. The results obtained prove that there is no significant difference in the performance of models using GINI index and Information gain before and after the data set balanced. The results are in line as stated by Mingers [11] that splitting indices have no impact on accuracy. In summary, the results obtained in this paper show that classification accuracy of decision trees for both balanced and imbalanced data sets, is not sensitive to the choice feature selection metrics that were studied.

Another interesting observation is balancing the dataset has lowered the majority class accuracy with decrease in count of true positives and minority class accuracy has improved with increase in the true negative count. In other words, the sensitivity decreased and specificity improved after the data set is balanced. Despite the fact that there is a decrease in overall accuracy, there is clearly a significant rise in the minority class accuracy. This proves that classification accuracy is sensitive to number of positive and negative samples in the data set and type of data, balanced or imbalanced.

REFERENCES

[1] James, G., Witten, D., Hastie, T., and Tibshirani, R.: ‘Tree-based methods’: ‘An introduction to statistical learning’ (Springer, 2013), pp. 303-335.

[2] Doherty, C., Camina, S., White, K., and Orenstein, G.: ‘The path to predictive analytics and machine learning’ (O'Reilly Media, 2017. 2017).

[3] Turban, E., Sharda, R., and Delen, D.: ‘Business intelligence and analytics: systems for decision support’ (Pearson Higher Ed, 2014. 2014).

[4] Loh, W.-Y., and Shih, Y.-S.: ‘Split selection methods for classification trees’, Statistica sinica, 1997, pp. 815-840.

[5] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J.: ‘Classification and regression trees. Belmont, CA: Wadsworth’, International Group, 1984, 432, pp. 151-166.

[6] Raileanu, L.E., and Stoffel, K.: ‘Theoretical comparison between the gini index and information gain criteria’, Annals of Mathematics and Artificial Intelligence, 2004, 41, (1), pp. 77-93.

[7] Shannon, C.E.: ‘A note on the concept of entropy’, Bell System Tech. J, 1948, 27, (3), pp. 379-423.

[8] Wang, Y., Li, Y., Song, Y., Rong, X., and Zhang, S.: ‘Improvement of ID3 algorithm based on simplified information entropy and coordination degree’, Algorithms, 2017, 10, (4), pp. 124.

[9] Fan, R., Zhong, M., Wang, S., Zhang, Y., Andrew, A., Karagas, M., Chen, H., Amos, C., Xiong, M., and Moore, J.: ‘Entropy - based information gain approaches to detect and to characterize gene - gene and gene - environment interactions/correlations of complex diseases’, Genetic epidemiology, 2011, 35, (7), pp. 706-721.

[10] Lefkovits, S., and Lefkovits, L.: ‘Gabor feature selection based on information gain’, Procedia Engineering, 2017, 181, pp. 892-898.

[11] Mingers, J.: ‘An empirical comparison of selection measures for decision-tree induction’, Machine learning, 1989, 3, (4), pp. 319-342.

[12] Patil, N., Lathi, R., and Chitre, V.: ‘Comparison of C5. 0 & CART classification algorithms using pruning technique’, Int. J. Eng. Res. Technol, 2012, 1, (4), pp. 1-5.

[13] Suneetha, N., Hari, V., and Kumar, V.S.: 'Modified gini index classification: a case study of heart disease dataset', International Journal on Computer Science and Engineering, 2010, 2, (06), pp. 1959-1965.

[14] Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., and Honrao, V.: 'Predicting students' performance using ID3 and C4. 5 classification algorithms', arXiv preprint arXiv:1310.2071, 2013.

[15] Hssina, B., Merbouha, A., Ezzikouri, H., and Erritali, M.: 'A comparative study of decision tree ID3 and C4. 5', International Journal of Advanced Computer Science and Applications, 2014, 4, (2), pp. 13-19.

[16] Moro, S., Cortez, P., and Rita, P.: 'A data-driven approach to predict the success of bank telemarketing', Decis Support Syst, 2014, 62, pp. 22-31.

[17] SHARDA, R.D.: 'BUSINESS INTELLIGENCE AND ANALYTICS: Systems for Decision Support' (PRENTICE HALL, 2016. 2016).

[18] https://people.revoledu.com/kardi/tutorial/DecisionTree.

[19] https://dataaspirant.com/2017/02/03/decision-tree-classifier-implementation-in-r/.

[20] Therneau, T., Atkinson, B., Ripley, B., and Ripley, M.B.: 'Package 'rpart'', Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016), 2015.

[21] Zheng, A.: 'Evaluating machine learning models: a beginner's guide to key concepts and pitfalls', 2015.

[22] Tan, P.-N., Steinbach, M., and Kumar, V.: 'Introduction to data mining' (Pearson Education India, 2016. 2016).

[23] Blagus, R., and Lusa, L.: 'Improved shrunken centroid classifiers for high-dimensional class-imbalanced data', BMC bioinformatics, 2013, 14, pp. 64-64.

[24] McHugh, M.L.: 'Interrater reliability: the kappa statistic', Biochemia medica: Biochemia medica, 2012, 22, (3), pp. 276-282.

[25] McGee, S.: 'Evidence-based physical diagnosis e-book' (Elsevier Health Sciences, 2012. 2012).

[26] Ensrud, K.E., and Taylor, B.C.: 'Epidemiologic Methods in Studies of Osteoporosis': 'Osteoporosis' (Elsevier, 2013), pp. 539-561.

[27] Zheng, Z., Cai, Y., and Li, Y.: 'Oversampling method for imbalanced classification', Computing and Informatics, 2016, 34, (5), pp. 1017-1037.

[28] Cordón, I.: 'Working with imbalanced datasets'.

[29] Wasikowski, M.: 'Combating the class imbalance problem in small sample data sets', University of Kansas, 2009.

A Review of Data Gathering Algorithms for Real-Time Processing in Internet of Things Environment

Atheer A. Kadhim[1], Norfaradilla Wahid[2]
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
Johor, Malaysia

***Abstract*—Today, Wireless Sensor Network (WSN) has become an enabler technology for the Internet of Things (IoT) applications. The emergence of various applications has then enabled the need for robust and efficient data collection and transfer algorithms. This paper presents a comprehensive review for the existing data gathering algorithms and the technologies adopted for that applications. After reviewing the algorithms and the challenges related to them, which extend the physical reach of the monitoring capability; they possess several constraints such as limited energy availability, low memory size, and low processing speed, which are the principal obstacles to designing efficient management protocols for WSN-IoT integration.**

Keywords—Internet of Things (IoT); Wireless Sensor Network (WSN) and Data Gathering; Virtual Machine (VM); Virtualization Cloud (VC); Data Reduction (DR); Access point (AP); Mobile Ad hoc Network (MANET)

I. Introduction

The Internet of Things (IoT) is one of the emerging technologies in the area of information technology [1]. It is widely called IoT which means that many things or objects are interconnected to each other through the Internet [2]. Internet technology is known for a long time and has been used for connecting computers using Internet protocol (TCP/IP) so that millions of networks are interconnected around the world [2]. Those networks are used for different kinds of purpose such as private, public, academic, business, and government networks. Technically, these networks might be connected to each other using fiber optics or wirelessly [3].

The availability of Internet networks has triggered an interest in connecting all the objects using Internet networks [4]. As such, several researchers have paid attention over the last ten years to enable connectivity for worldwide networks. This, in turn, has enhanced the vision for global networking for all the objects. IoT is virtual shall provide unlimited opportunities and connections to occur [5]. Nowadays, IoT research and development has become one of the hot topics in many countries around the world. Although IoT has provided a lot of opportunities, many challenges have been arisen such as security considering the huge number of devices connected to each other for achieving a certain type of function.

The next generation of IoT is required to provide new services to meet the demand of the fourth industrial revolution. Therefore, it must be able to deal effectively in the data transfer process without any human involvement in the interconnected objects such as computing devices and digital machines [3]. Although there are tremendous works have been done by several international standardization bodies, industry players, researchers, developer and other parties, there are several issues need to be addressed to reach the peak of IoT capabilities [6].

However, in this work, several studies are reviewed, discussed and critically analyzed to providing a solid literature review for future research. Additionally, a wide range of case studies from the past up to date is presented for a better understanding of the theory related to each proposed algorithm and the applicable technology in each study. Articles, journals, books, and previous works had been listed in the following tasks.

In the manner, Dias et al., [6] review numerous techniques that are used for Prediction-Based Data Reduction in Wireless Sensor Networks. Meanwhile, the work of Maraiya et al., [7], reviews the most common Data Aggregation methods in WSN. In the work of Cheng et al., [8], reviewed the state of the art of approximate data collection algorithms in IoT and WSN. Fig. 1 shows the number of reviewed and discussed articles in this work based on the years, note, * represents the number of review articles.

A total of 41 research articles and four review articles are covered in this work. The review emphases on the fundamental of IoT based on WSN which are Sensor Technology, Characteristic Features, Overview of IoT Sensor, IoT Hardware Prototype and Saving Energy Techniques. This review approach allows us to improve the scope and shape the direction of IoT based on WSN.

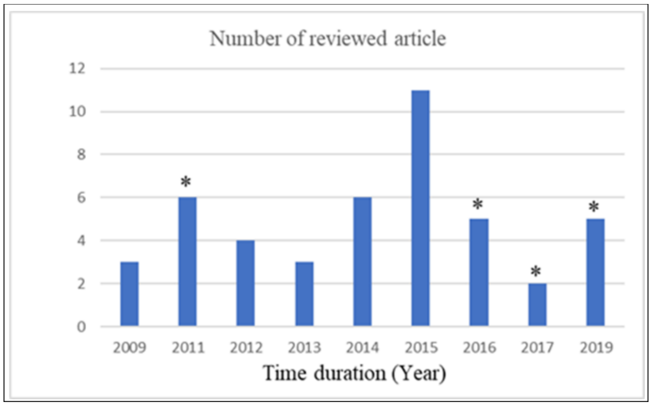


Fig. 1. An Overview of the Reviewed Articles.

This research is partly sponsored by Universiti Tun Hussein Onn Malaysia through Tier 1 Grant H108.

This paper segmented into four parts starting with the section of introduction which describes the IoT, WSN, and IoT based on WSN. Furthermore, an overview of the IoT based on WSN has been discussed under Section II. Then, Section III a reviews the IoT sensor node energy-saving methods in WSN is given. Whereas, the analysis and discussion of this work have been given in Section IV. Finally, Section V presents the conclusion of this paper.

II. Overview of IoT based on WSN

The vision for the Internet in the future is to be a global network consisting of many objects connected together using a specific IP address based on the relevant standard. Accordingly, all the devices including computers, sensors, RFID tags or mobile phones will have the capability to access the network and then communicating with other devices for the purpose of performing a certain task [9]. The WSN technology has increased the importance of IoT by combining the technologies of WSN and the internet resulting in IoT infrastructure. WSN is widely utilized for different kinds of applications. It is used mainly for gathering data from the field or environment through sensors. The key element for the IoT paradigm is RFID and WSN. RFID is used for identification purposes and tracking. Meanwhile, WSN is a very good option for proving sensing actuation function to IoT [3]. WSN has been adopted in many applications as an effective solution in several applications and research.

Unfortunately, the adoption of WSN in different types of applications has made a lot of challenges to specify the WSN requirement to be used for IoT. Generally, the common WSN platform can be applied with reasonable results in a wide range of IoT monitoring applications [6]. Furthermore, it is required to deploy WSN not only in monitoring applications but also in many applications such as security, biomedical research and tracking [10]. So, IoT is expected to play a significant role in very vital applications for emergency services. Based on the type of network, IoT networks can be classified into many types according to its intended application such as environmental data collection, military applications, security monitoring, health applications, home applications, and so on.

Although the generic form of WSN is applicable for monitoring applications, it still requires tough requirements such as employing a huge number of nodes with very low cost [9, 10],. The designer needs to consider that the nodes have to stand alone for a long time before the service time. Other factors also need to consider such as simplicity positioning the nodes with cheap maintenance cost. These requirements made generic WSN platforms less preferred [7]. This is would limit the applicability while it is suitable for many applications including but not limited to agricultural, medical and military applications as depicted in Fig. 2.

WSN is characterized as a two-direction system where the data on long-distance can be transferred between the sink node and sensors through a jumping mechanism. As such, the sensed data such as temperature, moistness, light and so on can be sent to the destination point and afterward pass on to the preparing gear [11]. The detecting hubs convey in multi-jump. Each sensor is a handset having reception apparatus, a small-scale controller and an interfacing circuit for the sensors as a correspondence, activation and detecting unit individually alongside a wellspring of intensity which could be both battery or any vitality reaping innovation However, in this work we emphases on the fundamental parts of IoT based on WSN as shown in Fig. 3.

A. Sensor Technology

As a keyway to obtain information, sensor generation and communicator generation, personal computer technology constitutes the three pillars of data technology. Sensors are the main components of the Internet of Things awareness layer, which can assist the Internet of Things data and access the external physical world in a timely manner [9]. The sensor network technology formed by the sensor network technology's sensor and communication network has laid the foundation for the development of the Internet of Things. The sensor warehouse is mainly used for the environment and small parts tracking to meet the requirements of the product to the surrounding environment and safety monitoring. Inductive devices, including in particular sensors that select sensing gadgets (including light sensors, temperature sensors) are widely used in warehouse control of multipurpose systems [11].

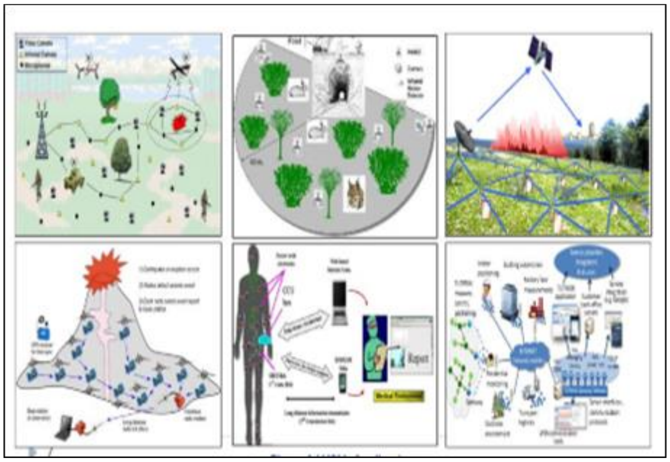


Fig. 2. WSNs Applications [10].

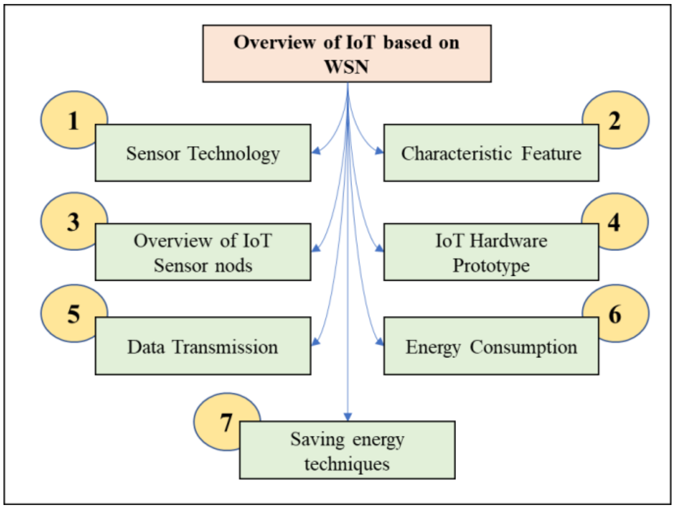


Fig. 3. Overview IoT based on WSN.

A sensor is a tool that converts the physical measurement sign. The sensor's selection must conform to the requirements of the surrounding environment or the object. It can be collected by the body, and the organic chemistry effect is associated with the measurement sensor as the development of the system. In well-known circumstances, there are no special requirements for the environment of daily warehouse items (including temperature, humidity, etc.), so special sensors should be installed near the well-known warehouses. This type of sensor may be a combination of various sensing environments [10]. In the measurement to meet the needs of general merchandise. For unique items or unique business needs, often more feature definition areas are needed, and integrated sensors are set up, including gravity, strain, fuel density, and sensor noise. The steeper price included in the number of sensors, not always the same as the use of smooth, can be placed when the product is replaced or the items necessary to alternate, move the sensor, and open and close the component provider as required to increase the sensor usage rate [8].

B. Characteristic of WSNs

WSN can be generally defined as a set of nodes that are connected to each system so that all the nodes perform together with the sensing and control of the environment. This process enables the communication between humans or computers and the environment under control [12]. WSNs these days, as a rule, incorporate sensor hubs, actuator hubs, passages, and customers. An expansive number of sensor hubs conveyed haphazardly within or close to the observing region (sensor field), structure arranges through self-association. The gathered information is then checked by sensor hubs and pass forward to other sensor hubs via the jump mechanism [13-15]. For the purpose of reliable transfer of data, the hubs need to take care of the availability of jump to the portal hub which then transfers the data to the administration hub via the web or satellite [7]. The client designs and deals with the WSN with the administration hub, distribute checking missions and gathering of the observed information. As related advancements develop, the expense of WSN hardware has dropped drastically, and their applications are slowly growing from the military regions to mechanical and business fields. In the interim, guidelines for WSN innovation have been all around grown, for example, ZigBee.

C. Overview of IoT Sensor Nodes

Although WSN provides good service for sensing areas with difficult to access, it experienced a lot of challenges which limit its deployment. One of those challenges is to provide a sustainable power supply for the wireless sensor nodes [6], [9]. Up to our knowledge, there are a large number of published papers have proposed different kinds of solutions to overcome this drawback. Recently, it has been proposed to use energy harvesting technology and wireless charging technology to provide sustainable energy resources [16].

One of the basic elements in IoT based WSN is the IoT sensor hub which composes of four sections (i) power module, (ii) a sensor, (iii) a microcontroller, and (iv) a remote handset. The function of the power module is to provide power for the framework whereas the sensor is used for capturing the status of the sensed raw data. It is in charge about data transfer and exchange to other sensors. In fact, it converts the sensed data such light into an electrical signal and then passing it to the microcontroller [8]. The microcontroller is employed to receive the information sent by the sensor and perform the required operations on it. The RF module is the last element which is located at that point where the information is exchanged so physical acknowledgment can be accomplished in this element. It is highly required to consider all these elements need to be with small size and low power consumption.

D. IoT Hardware / Prototype

The rapid advancement of computing hardware technology has resulted in developing small scale devices at a reasonable cost. Consequently, IoT has got attention for various applications [3].

For IoT, the microcontrollers are integrated with processors, wireless chips and other components to form the Prototyping Development Kit through embedded software packages with reprogramming capability. Other studies, as in this future research, employ Arduino hardware since it is an open-source device for controlling a large number of sensors rather than personal computers [17]. Briefly, Arduino has the capability to be registered based on the microcontroller board and the programming composer of the board. Arduino hardware is programmed using C or C++ programming language. Based on the handling sight and the programming condition, Arduino load represent logical wiring similar to physical connections, [18]. The microcontroller has a very important microchip called AVR provided by Atmel organization. This chip operates at 16 MHz with an 8-bit, but it has limited and has limited accessible memory (32 Kbytes of capacity) and 2 Kbytes of irregular access memory. Due to this Atmel chip, Arduino became popular for work especially for many DIY applications [16].

With respect to the product advancement, programming of the shield was begun earlier with sensors programming and further included clock module. DHT22 and RTC modules are used to provide information with aid of the libraries of Arduino IDE [18]. The data are read in real-time through the analog pins of sensors and then send a copy of data to IDE’s serial port for testing purposes. Arduino gives an SD card library, which was utilized to make a capacity SD Card Data-log. The capacity spares sensor information and the opportunity to the SD card on the W5100 arrange shield. To give clients remotely screen their nursery through site page, the webserver should have been built up. Webserver libraries were made for Arduino however they didn't meet the necessities of the framework. Better web content help was expected to the venture, so new webserver was intended to fit exactly the framework prerequisites. Live to outline was intended to show information for the client on the site [18]. Without JavaScript support on the webserver, web association would have been required. The new webserver empowers the framework to be utilized disconnected in a neighborhood without web association.

E. Data Transmission in IoT based WSN

Al-Fagih, [19] has reviewed the data transmission in IoT in detail. The data transmission in IoT-WSN is directly related to the type of application for data transfer from the sensor toward the access point. Generally, data transmission can be performed through one of the following types: continuous data transmission where the data are sent by the sensors in specific times. The second type is event-dependent where the transmission is enabled when such an event occurs. The last is query-based data transmission where data are sent once the access point transmits query. Despite these three types are different in operation, the continuous type can be jointly applied with an event or query-based data transmission forming a hybrid model. It's observed that the introduced framework facilitates applications that are more relevant to the query-based model.

In the literature, several protocols are proposed for data delivery in WSN such as, [5, 6, 19]. These protocols are arranged in three different architectures: a hierarchical structure, a data-centric structure, or location-based structure as depicted in Fig. 4. In the hierarchical structure, nodes are arranged in a clustering paradigm so that the head of the cluster collected together for the purpose of minimizing the transferred data and thus reducing the cost of energy.

The protocols related to the data-centric paradigm are query-based and as such only, the desirable data will be transmitted. In this process, the duplication of transmission for the same data is avoided. In the third type, the data is sent selectively to the targeted location. This is a good alternative to the transmission of data to all the locations. As a result, the bandwidth and power are saved significantly.

It is vital to take in mind that IoT is strongly dependent on WSN. In WSN, remote communication has existed between the sensors and the network hub. The major function of IoT is to construct an overall system among all the conceivable articles. Besides, WSN is a genuinely enhanced innovation that guides the client to accomplish the importance of IoT [8]. The primary thought of WSN is to associate the detecting layer and system layer in the IoT.

Fig. 5 shows one scenario of WSN where the operation is event related. Once the sensor node identifies an even, it gathers the data and send it to the next nearest node and so on until reaches to the destination node. This is the simple structure of independent WSN. In this scenario, data are delivered only to a single gateway. In addition to that, the data can be transmitted in access point scheme or hybrid scheme as shown in Fig. 6, respectively.

The independent WSN has been improved and enhanced. The improved version of the independent WSN is called hybrid WSN. In contract to independent WSN, hybrid WSN has multiple gateways for the purpose of data transmission. In this case, WSN will maximize its performance.

In the third scheme, Access Point-WSN is not similar to the other schemes. Basically, this scheme adopts WLAN structure [20]. As shown in Fig. 7, there are many of hubs in WSN, each one of them compose with the other to make the connections. The system is interfacing WSN and Web through one portal. When the entryway is separated, there is an alternative method to interface the two systems. Nonetheless, it is divides the nodes into two parts to support and fix the lacking that may happen during data transmission among the nodes. It is more grounded from the system of one portal in first scenario. In contrast to the past strategy, pathway arrange the fills as a self-arranged in WSN.

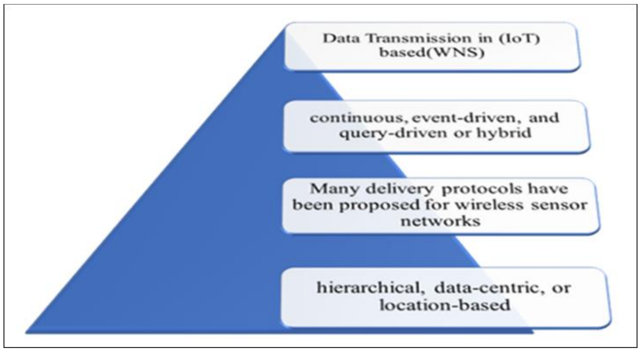

Data Transmission in (IoT) based(WNS)
continuous, event-driven, and query-driven or hybrid
Many delivery protocols have been proposed for wireless sensor networks
hierarchical, data-centric, or location-based

Fig. 4. Data Transmission in (IoT) based (WNS).

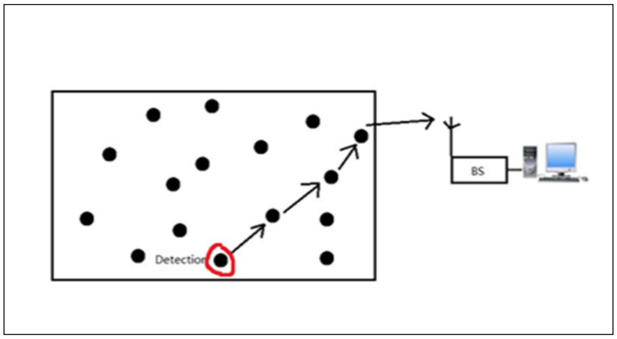


Fig. 5. Independent WSN Scheme.

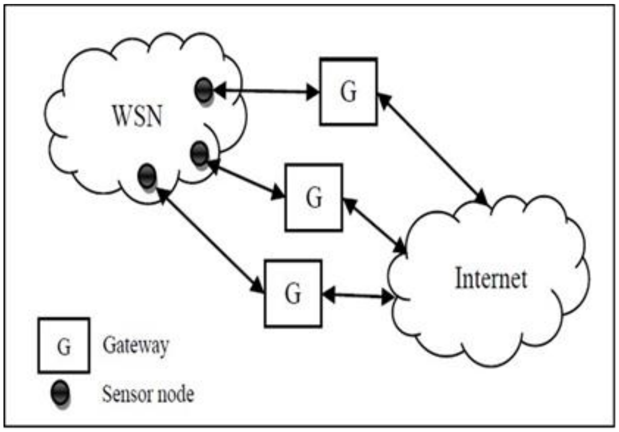


Fig. 6. Scheme of Hybrid WSN Data Transmission.

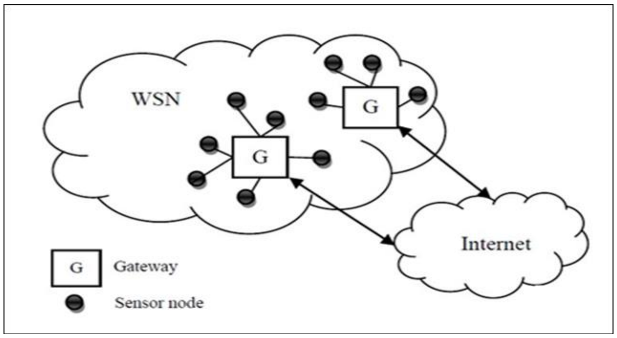


Fig. 7. Scheme of hybrid WSN data transmission

In WSN, the applicable protocols related to data routing are based on the data transmission model specifically in managing network active status and power consumption. In this regard, it is found that hierarchical scheme protocols are the most proper option for environmental monitoring applications since data are sent unremittingly to the AP. This is expected since applications can produce a large number of duplicated data which gathered together to the receiver. This, in turn, would shrink data overwhelming on the route and thus optimize the energy consumption. However, the typical WSN is too limited to meet the needs of the Internet of Things in terms of heterogeneity and transmission/processing load. Therefore, they adopted an extended concept of a sensor network that contains MANET nodes.

F. IoT Energy Consumption

energy waste has been described in [21]: In IoT based WSNs, sensors consume energy while sensing, handling, transferring and transferring or getting data to complete the tasks. Data gathering are normally collected by sensor dedicated for that job. Obviously, as the produced data are minimized, the saved energy would be increased. The inherent redundancy of a wireless sensor network generates a huge similar report, and the network is responsible for routing to the receiver. It has been proven practically, that the communication system is responsible for the significant amount of power wastage. In terms of communication, from the perspective of applications, useless countries have also wasted a lot of energy [9].

G. Energy Saving Techniques in WSN

There are various solutions for tacking the issue of energy consumption. Some of those solutions are to minimize the sent data, to reduce the overhead which accompanies the data as well as applying a smart and efficient routing algorithm. Additionally, increasing the time for idle status with no transmission which implies a limited number of transmissions. The other solution is to focus on topology control as described by [9]:

- Minimize of data transmission: the transfer of data duplication or unnecessary data can be avoided by applying these certain types of solutions. Many related works have been discussed in the related works section in this chapter.

- Decreasing the overhead: The data transmission will be maximized by minimizing the overhead attached to the data. Several existing techniques have been discussed in the related works section in this paper.

- Optimizing the routing: For better performance, the routes need to be "available" during all the time but in such a way to keep minimum power consumption. In this regard, some routing protocols are utilizing the node movement or "send to all" characteristics of WSN. In the other type, the GPS coordinate of the nodes is employed to create a path to the destination. In a different way, a hierarchy structure of nodes is used for a proper path with less overhead. Additionally, many paths can be employed for reaching the destination by balancing the data on different paths resulting in more. Finally, a statistics-centric protocol transmits the most effective information to the involved nodes to avoid costly transmissions.

- Duty cycling: The duty cycle refers to the proportion of time that a node is active during its life cycle. Nodes should coordinate and adapt to specific application needs during sleep or activity. The nodes with High granularity technology are the active nodes in the WSN while the nodes with low granularity imply shut down the RF transmission of the active nodes since there is no data to send. They are highly relevant to media access agreements.

III. Review of the IoT Sensor node Energy Saving Methods in WSN/IoT

Chaudhary et al., [20] praised the improved technologies of energy-saving technologies for data gathering in WSN. In that article, the battery has been used wisely and effectively and thus low power consumption is identified. As such, people can maximize the benefit of using the same resource but with a large number of messages. The production of cluster overhead and node selection methods are based on certain parameters in order to use the correct decisions in this case, using global weight calculations such as nodes, cluster head data collection, and data aggregation using data cube clustering.

Rohankar et al., [21] had made reviewed the latest developments in WSN including data gathering methods. The review paper had categorized each consideration technology based on the underlying topology. The second classification is based on the saved energy scheme. Those technologies are assessed qualitatively by comparing one to each other. At the end of the review, the limitations of the technologies considered were discussed.

In the previous work of Quwaider and Jararweh, [22], they proposed a cloud-based and efficient data collection system in a cloud-based WBAN. The main objective was to provide a wide range of WBAN monitoring data to end-users or service providers in a robust way. In this case, the data gathering in WBAN has been simulated using a prototype composed of a virtual machine (VM) and a virtualization cloud (VC). Using this prototype system, they provide a scalable storage and processing infrastructure for large WBAN systems.

In Xu and Song [23], the authors studied instantaneous periodic query scheduling in multi-hop data gathering (WSNs). Given a set of heterogeneous data gathering queries in the WSN, each query desires that data from the source node be assembled into the control center. Firstly, they proposed a series of almost urgent requirements for different queries that can be scheduled by the WSN. Then, three effective data gathering algorithms were developed to meet instantaneous requirements under the limitation of resources. The issue has been addressed through three vital tasks: (1) data gathering, building of routing tree, (2) schedule of path activity, and 3)) scheduling on Packet-level.

In Luo et al., [24], the work investigated the speed of raw data gathering from all nodes to the receiver. In this case, the TDMA technique has been employed on the same frequency channel. Additionally, a centralized and distributed fast data gathering algorithm is employed to find the optimal solution in polynomial time when no interfering links occur. The study also proposed the RCTS algorithm for identifying the best solution. It is found that RCTS is time efficient and it is a good candidate for eliminating the major parts of the interference captured in indoor and outdoor environments.

Guo et al. [25] proposed a new method called Event-based Data Aggregation (EDA) which uses using the fuzzy logic cloud member model to gather the data of events in the WSN. In this method, the base station has the capability to restore the whole event data once it receives the data packets of the event. EDA method provides a degree of balance between delay factor, savings of flow and the accuracy of the restored events. The performance evaluation of this method has been conducted by Guo using both analysis and simulations.

Jacques et al., [26], proposed a new filtering technique for specifying the redundancy of data received in constant time slots. Further to that, the data gathering method has been proposed based on grouping data which share most of the same features as such the data integration would be maintained.

In the work of Pfletschinger et al., [27], a network coding scheme in WSN has been proposed and its effectiveness has been evaluated through the following factors: reliability enhancement, power efficiency and resilience to network protocol failures. The main challenge in that work was to identify the number of eavesdropping so that space diversity can be employed but with low power consumption.

In the prior work of Raza, [17], the purpose of this work is to explore the complex interactions between application features and adaptive mechanisms across the network stack by using specific real-world deployments. Moreover, the paper proposed a generic framework that integrates adaptation into near-optimal energy efficiency for heterogeneous applications.

In Bahi et al., [28], the author introduced an energy-saving technique for data gathering in structure with constant on time slots transmission. This study has discussed the issue of locating the pairs of nodes that generates redundant data. In addition, this study provides a frequency filtering method to solve this problem. In Enam et al., [29], the work was to build up a vitality productive information gathering condition for a substantial scale, haphazardly sent group based remote sensor organizes by utilizing a virtual lattice-based instrument to limit the bunches and balance out the group bulks in the system. This was one of the requirements for implementing the proposed differential information total plan for the spatially related information in a bunch.

In the work of Laiymani and Makhoul [30], they displayed a productive versatile testing approach dependent on the reliance of restrictive difference on estimations shifts after some time. At that point, in extra they proposed a various dimensions action display that utilizes conduct capacities demonstrated by altered Bezier bends to characterize application classes and take into consideration examining versatile rate. The proposed strategy was effectively tried in a genuine sensor informational collection as the researcher said.

In Enam [31], the author built up a novel and a versatile technique for information conglomeration that abuses the spatial connection between the sensor hubs. The primary element of the proposed accumulation technique is that notwithstanding lessening the expense of excess information move in the system, it additionally ideally uses the accessible space in a bundle at each group head.

Trade-off between them depends to a large extent on the certain application. In this regard, one of the techniques for data gathering is called Prefix Frequency Filtering (PFF) where power consumption and data accuracy are targeted in that study. The main target of PFF is to identify the data groups produced by neighboring nodes with shared features resulting in canceling the redundant data and thus avoiding energy dissipation. Although this method is simple it requires tedious computational time. In prior work of Harb, Makhoul, and Laiymani [32], PFF has been improved by integrating K-means of clustering algorithm as such it is called KPFF. The KPFF was able to minimize the time needed for detecting similar pairs and therefore the data latency is minimized as well.

Li et al., [33] had analyzed the complexity for many factors such as data message complexity and energy cost complexity. In this work, the lower bound of the complexity of the optimal method has been employed but for the other factors, an efficient distributed algorithm has been used. This in result provided gradually matching with the upper bound of complexity.

In Carlos-Mancilla et al., [34], this work proposes and builds up a proficient information collection technique for remote sensor systems (WSN). In the proposed information collection strategy, each bunch of head (CH) hub contains a nearby sending history to choose whether to advance or dispose of the most as of late gotten parcel. At the point when the new parcel touches base at the CH hub, the limit is determined dependent on the data of the sending history; at that point, an arbitrary number is created and contrasted with the edge an incentive with deciding if the information bundle ought to be disposed of. Truth be told, the CH hub advances the new bundle with a likelihood of 1-p and disposes of it with the likelihood p that decides the parameter p dependent on the sending history.

The energy-saving with optimism use for a long time has become a recent challenge facing the researchers around the world. It is required to minimize the power usage in WSN but has to ensure the reliable and robust functional performance of WSN. It has to meet the minimum requirement of normal operation without failure due to power supply. Whereas, the authors proposed a reactive data acquisition scheme called SWIFTNET in [35]. It is based on the synergetic effect of a combination of data reduction methods and energy-saving data compression schemes. In particular, it combines compressive sensing, data prediction, and adaptive sampling strategies.

The Internet of Things represents advances in miniaturization, wireless connectivity, and increased data storage, driven by various sensors. Sensors detect and measure any changes in location, temperature, light, etc.; in addition, they need to convert billions of objects into data-generating “things” to report their status and often interact with their environment. Application and service development methods and frameworks are needed to support the implementation of solutions that cover data collection, transmission and data processing, analysis, reporting, and advanced querying. In the previous work of Lengyel et al. [3], this article introduced the Sensor HUB framework, which utilizes the most advanced open source technologies and provides a unified toolchain for IoT related applications and service development. Sensor HUB is both a method and an environment that supports the development of applications and services related to the Internet of Things. In addition, it supports data monetization methods that provide a way to define data views and analyze data on different data sources. The framework uses the platform-as-a-service (PaaS) model and has been applied in the areas of vehicles, health, production lines, and smart cities.

Data collection and propagation in the Internet of Things (IoT) Wireless Sensor Network (WSN) requires a stable multi-hop network path from source to sink. However, due to limited energy, the battery consumption of the interrupting node can cause the path to be disconnected and result in the end-to-end data transmission failure of the WSN-based IoT. Therefore, in addition to its own energy-saving, each sensor involved in multi-hop transmission activity also needs a feasible strategy for selecting a relay node through utilization. Its remaining energy and multi-hop IoT network connection. In Luo et al. [5], the author first analyzed the energy consumption model and data relay model in wireless sensor networks, and then proposed the concept of "equivalent node" to select relay nodes to achieve data transmission and Energy-saving optimization. A probabilistic propagation algorithm called ENS PD is designed to select the best energy strategy and extend the lifetime of the entire network. Extensive simulations and actual test results show that our models and algorithms can minimize power consumption compared to other methods while ensuring the quality of communications in WSN-based IoT.

In the previous work of Rault et al. [36], the paper proposes a novel sensor network data acquisition framework using flight sensor nodes. Since sensor nodes are usually limited by energy, efficient data communication within the network is required. In contrast to its conventional role in sensor networks, the proposed framework utilizes various entities that form networks for different utilities. The use of flight sensor nodes is often considered the traditional purpose of sensing and monitoring. Flight sensing nodes are commonly used in the form of an airborne sensor network and they cannot be used as data collection entities as proposed in this framework. Similarly, it is often desirable for a cluster head (CH) to transmit aggregated data to a neighboring CH or directly to a base station (BS). In the proposed framework, the CH transmits data directly to the flight sensor nodes, avoiding the need for energy-intensive multi-hop inter-cluster communication to communicate information to the BS. Flight sensor nodes are called sensor flights.

In Mudgule et al., [37], the author focuses on the data redundancy and energy of sensor nodes. Data simplification is one of the data pre-processing techniques for data mining, which can improve storage efficiency and reduce costs. Data Reduction (DR) is designed to remove unnecessary data when transferred. For this reason, according to WSN, many data reduction strategies will be introduced. This survey introduces the latest data reduction-based algorithms and techniques that help increase the network's energy and longevity.

In Maraiya et al., [7], discussed the data aggregation approaches based on the routing protocols, the algorithm in the wireless sensor network. And also discuss the advantages and disadvantages of various performance measures of the data aggregation in the network.

In Pandey and Kaur [38], the authors' attention to various data aggregation algorithms in a wireless sensor network. Data aggregation technique increases the lifetime of sensor networks by decreasing the number of packets to be sent to the sink or base station. Here, they first explore the data aggregation algorithms on the basis of network topology, then they explored various tradeoffs in data aggregation algorithms and finally they highlighted security issues in data aggregation.

The work in (Hung et al., [39], proposed a centralized algorithm to determine a set of representative nodes with high energy and wide data coverage. Here, the sensor node's data coverage is considered to be a set of sensor nodes that have very close reading behavior to a particular sensor node. In order to further reduce the extra cost in the messages used to select representative nodes, a distributed algorithm was developed. In addition, when the energy of the original representative node is insufficient or cannot capture the spatial correlation within its respective data coverage, a maintenance mechanism is proposed to dynamically select the alternative representative node. Through experimental research on synthetic and actual data sets, the proposed algorithm has been proved to be able to effectively provide approximate data collection while extending the network lifetime.

A performance assessment of information reduction techniques for IoT based WSN Multimedia applications has been provided in [40]. In this article, the authors study the performance of various BS algorithms and compression techniques in computing and communication energy, time, and quality. They have chosen five different BS algorithms and two compression techniques and implemented them on the Android platform. Considering the fact that these BS algorithms operate under the WMSN environment where data is subject to packet loss and error, they also studied the packet loss rate performance of the network under various packet sizes. Experimental results show that the highest energy efficiency BS algorithm can also provide the best prospect detection quality. The results also show that data compression techniques including BS algorithms and compression techniques can provide significant energy savings in terms of transmission energy costs.

In the prior work of Dias et al., [6], the authors analyzed and classified the existing prediction-based data reduction mechanisms for wireless sensor network design. Their meaning is based on the constraints of the wireless sensor network, the characteristics of the prediction method, and the monitoring data, and a systematic procedure for selecting the prediction scheme in the wireless sensor network. Finally, this article concludes this article and discusses future challenges and open research directions for the use of predictive methods to support the development of wireless sensor networks.

A data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications have been proposed in [41], In this work, the author proposes another structure that consolidates information expectation, pressure, and recuperation together to accomplish the exactness and effectiveness of grouped remote sensor arrange information handling. The principle reason for this system is to diminish correspondence costs while guaranteeing the precision of information handling and information expectation. In this system, information expectation is accomplished by executing the Least Mean Square (LMS) bi-forecast calculation with the best advance size with the least mean squared subsidiary (MSD), where the group head (CH) can get a decent guess from the sensor the genuine information of the hub. On this premise, the brought together primary segment investigation (PCA) innovation is utilized to pack and recoup the forecast information of the CHS and the sink individually, in order to spare the correspondence cost and wipe out the spatial repetition of the detecting information on the earth. Every one of the blunders that outcome from these procedures will, in the long run, be assessed hypothetically and these are controllable. In light of hypothetical examination, the creators structured some usage calculations. Reproductions utilizing certifiable information have demonstrated that (LMS) system gives a financially savvy answer for natural checking applications in bunch based WSNs.

In recent work of Alduais et al. [17], the authors have strategized to diminish the number of information transmissions and lessen the measure of information that prompts an all-encompassing system lifetime. The proposed strategy intends to diminish the number of transmitted messages by means of hubs supporting single and different sensors depending on the relative or relative contrasts between the transmitted present and last sensor estimations. The outcomes demonstrate that the proposed technique appears to demonstrate the best execution in decreasing the number of message transformations and parcel measures. In that article, the normal level of decrease in message transmission times is 74%, and 80% is the normal rate decrease of hub information. Bundle measure. From the outcomes, it tends to be obviously observed that diminishing the number of bundle transformations and decreasing the extent of hub information parcels lessens vitality utilization and broadens the administration life of the framework.

Raza et al. [11], depicts subsidiary based expectation (DBP), another kind of information forecast method that is less complex than the writing. Assessments utilizing genuine datasets from various WSN arrangements demonstrate that DBPs, for the most part, perform superior to contenders, with information pressure rates as high as 99% and great expectation exactness. In any case, tests led on genuine remote sensor arranges in expressway burrows have appeared considering the system stacking, DBP just triples its lifetime - a critical outcome in itself, however, it is a long way from the above information concealment rate. So as to completely understand the vitality funds acknowledged by information forecast, the information layer and the system layer must be together enhanced. In that review test explore, considering the activity of DBP, a basic change of the MAC and wiring stack can altogether expand the lifetime by multiple times.

In Aït-Sahalia, and Xiu, [42] this work proposes an algorithm based on “Principal Component Analysis” to perform multivariate data reduction. It was considered an air quality monitoring scenario as a case study. The results showed that using the proposed technique, the outcome of the study reduced the data sent preserving its representativeness. Moreover, it’s showed that energy consumption and delay were reduced proportionally to the amount of reduced data.

In prior work Mccorrie et al. [43], another strategy for specifically sifting detected information dependent on state acknowledgment has been concocted that utilizes skewed twofold exponentially weighted moving normal channels for exact state expectation. This is genuine regardless of whether a critical temperature step change happens. A test system was executed to create a flight temperature profile as the flight temperature experienced, all things considered, so the calculation could be balanced and assessed. The outcomes abridged a reenacted trip of 280 variable lengths (from roughly 58 minutes to 14 hours). The outcomes demonstrated that in the departure, cruising and landing stages, the number of transmissions was diminished by a normal of 95, 99.8 and 91% with the detecting and transmitting framework. Correlation of the transmissions experienced when examining at the same rate. The algorithm produces an average error of 0.11 ± 0.04 °C in the 927 °C range.

In de Carvalho et al. [44], the authors proposed to use a method based on multiple linear regression to improve prediction accuracy. The improvement is achieved by the multivariate correlation of readings gathered by sensor nodes in the field. The authors claimed that the solution has outperforms some current solutions adopted in the literature.

A survey for approximate sensory data collection has been presented in the recent work of Cheng et al., [45], that survey reviews the state of the art approximated by a collection algorithm. They classified the min into three categories: the model-based ones, the compressive sensing-based ones, and the query-driven ones. For each category of algorithms, the advantages and disadvantages are elaborated, some challenges and unsolved problems are pointed out.

In the recent work of Alduais et al., [1], that work displayed another pointer to assess the execution of various multivariate information decrease models in remote sensor systems (WSNs). The proposed measurement is known as the update recurrence metric (UFM), which is characterized as the recurrence of refreshing model reference parameters amid information accumulation. A strategy for evaluating the mistake limit amid the preparation stage has additionally been proposed. The prescribed blunder edge is utilized to refresh the show reference parameters when vital. Numerical examination and recreation results demonstrate that the proposed measurement confirms the adequacy of the multivariate information decrease show in vitality utilization of sensor hubs. Furthermore, the proposed versatile limit improves the execution of the model more than the non-versatile edge in decreasing the recurrence of refreshing model reference parameters, which correctly extends the lifetime of the node.

Compared to the non-adaptive thresholds of the multivariate data reduction model of MLR-B and PCA-B, the adaptive threshold increases the frequency of parameter updates by 80% and 52%, respectively.

IV. CONCLUSION

The vast usage of the Internet of Things (IoT) innovation for different applications has empowered the requirement for hearty and effective information accumulation and exchange calculations. This paper introduced a complete audit for the current information gathering calculations and the innovations received for those applications. It reviewed the proposed algorithm for tracking this issue. Although the existing algorithms for data gathering can perform well, it still needs to be further enhanced in the future to overcome all the deficiencies such as low power consumption for standalone sensors. This paper has covered a comprehensive review for those algorithms. However, this paper is a platform for developing many solutions by the researchers for an efficient algorithm for wireless sensors. That solution will be investigated when proposed in future work.

ACKNOWLEDGMENT

The authors would like to thank Universiti Tun Hussein Onn Malaysia for supporting this research through Tier 1 Grant H108.

REFERENCES

[1] Alduais, N., J. Abdullah, A. Jamil, and H. Heidari, Performance evaluation of real-time multivariate data reduction models for adaptive-threshold in wireless sensor networks. IEEE sensors letters, 2017. 1(6): p. 1-4.

[2] Moreno, M.V., J. Santa, M.A. Zamora, and A.F. Skarmeta. A holistic IoT-based management platform for smart environments. in 2014 IEEE International Conference on Communications (ICC). 2014. IEEE.

[3] Lengyel, L., P. Ekler, T. Ujj, T. Balogh, and H. Charaf, SensorHUB: an IoT driver framework for supporting sensor networks and data analysis. International Journal of Distributed Sensor Networks, 2015. 11(7): p. 454379.

[4] Noori, a., et al., Wireless Sensor Network Deployment Based On Machine Learning For Prolonging Network Lifetime And pdr. Journal of Theoretical and Applied Information Technology, 2019. 97(14).

[5] Luo, J., D. Wu, C. Pan, and J. Zha, Optimal energy strategy for node selection and data relay in WSN-based IoT. Mobile Networks and Applications, 2015. 20(2): p. 169-180.

[6] Dias, G.M., B. Bellalta, and S. Oechsner, A survey about prediction-based data reduction in wireless sensor networks. ACM Computing Surveys (CSUR), 2016. 49(3): p. 1-35.

[7] Maraiya, K., K. Kant, and N. Gupta, Wireless sensor network: a review on data aggregation. International Journal of Scientific & Engineering Research, 2011. 2(4): p. 1-6.

[8] Rohankar, R., C. Katti, and S. Kumar, Comparison of energy efficient data collection techniques in wireless sensor network. Procedia Computer Science, 2015. 57: p. 146-151.

[9] Minet, P., Energy efficient routing. Ad Hoc and Sensor Wireless Networks: Architectures: Algorithms and Protocols, 2009.

[10] Luo, S., Y. Sun, and Y. Ji, Data collection for time-critical applications in the low-duty-cycle wireless sensor networks. International Journal of Distributed Sensor Networks, 2015. 11(8): p. 931913.

[11] Raza, U., A. Camerra, A.L. Murphy, T. Palpanas, and G.P. Picco, Practical data prediction for real-world wireless sensor networks. IEEE Transactions on Knowledge and Data Engineering, 2015. 27(8): p. 2231-2244.

[12] Bröring, A., et al., New generation sensor web enablement. Sensors, 2011. 11(3): p. 2652-2699.

[13] Shujaa, m., lagged multi-objective jumping particle swarm optimization for wireless sensor network Deployment. Journal of Theoretical and Applied Information Technology, 2019. 97(2).

[14] Hammid, A.T. and M.H.B. Sulaiman, Series division method based on PSO and FA to optimize Long-Term Hydro Generation Scheduling. Sustainable Energy Technologies and Assessments, 2018. 29: p. 106-118.

[15] Hammid, A.T., M.H.B. Sulaiman, and O.I. Awad, A robust firefly algorithm with backpropagation Neural networks for solving hydrogeneration prediction. Electrical Engineering, 2018. 100(4): p. 2617-2633.

[16] Alduais, N.A.M., J. Abdullah, and A. Jamil, RDCM: An Efficient Real-Time Data Collection Model for IoT/WSN Edge With Multivariate Sensors. IEEE Access, 2019. 7: p. 89063-89082.

[17] Alduais N. A. M., J. Abdullah, A. Jamil, and L. Audah., An Efficient Data Collection and Dissemination for IOT Based WSN. In Proceeding of the 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, IEEE,Vancouver, 2016: p. 1-16.

[18] Yu, S. Arduino based Balancing Robot. in 2019 International Conference on Electronics, Information, and Communication (ICEIC). 2019. IEEE.

[19] Al-Fagih, A.E., A framework for data delivery in integrated Internet of Things architectures. 2013: Queen's University (Canada).

[20] Chaudhary, S., N. Singh, A. Pathak, and A. Vatsa, Energy Efficient Techniques for Data aggregation and collection in WSN. International Journal of Computer Science, Engineering and Applications, 2012. 2(4): p. 37.

[21] Reinhardt, A., D. Christin, M. Hollick, and R. Steinmetz. On the energy efficiency of lossless data compression in wireless sensor networks. in 2009 IEEE 34th Conference on Local Computer Networks. 2009. IEEE.

[22] Quwaider, M. and Y. Jararweh, Cloudlet-based efficient data collection in wireless body area networks. Simulation Modelling Practice and Theory, 2015. 50: p. 57-71.

[23] Beaulah, H.L., K. Thangaraj, and M. Chitra, Design of a New Energy Efficient L4 Leach Protocol based Visual Sensor Network for Forest Monitoring System. Asian Journal of Research in Social Sciences and Humanities, 2016. 6(7): p. 662-671.

[24] Palpanas, T., Real-time data analytics in sensor networks, in Managing and Mining Sensor Data. 2013, Springer. p. 173-210.

[25] Guo, Y., F. Hong, Z. Guo, Z. Jin, and Y. Feng. EDA: Event-oriented data aggregation in sensor networks. in 2009 IEEE 28th International Performance Computing and Communications Conference. 2009. IEEE.

[26] Bahi, J.M., A. Makhoul, and M. Medlej. Data aggregation for periodic sensor networks using sets similarity functions. in 2011 7th International Wireless Communications and Mobile Computing Conference. 2011. IEEE.

[27] Pfletschinger, S., M. Navarro, and C. Ibars. Energy-efficient data collection in WSN with network coding. in 2011 IEEE GLOBECOM Workshops (GC Wkshps). 2011. IEEE.

[28] Bahi, J.M., A. Makhoul, and M. Medlej. Frequency filtering approach for data aggregation in periodic sensor networks. in 2012 IEEE Network Operations and Management Symposium. 2012. IEEE.

[29] Enam, R.N., R. Qureshi, and S. Misbahuddin, A uniform clustering mechanism for wireless sensor networks. International Journal of Distributed Sensor Networks, 2014. 10(3): p. 924012.

[30] Laiymani, D. and A. Makhoul. Adaptive data collection approach for periodic sensor networks. in 2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC). 2013. IEEE.

[31] Enam, R.N. and R. Qureshi. An adaptive data aggregation technique for dynamic cluster based wireless sensor networks. in 2014 23rd International Conference on Computer Communication and Networks (ICCCN). 2014. IEEE.

[32] Harb, H., A. Makhoul, D. Laiymani, A. Jaber, and R. Tawil. K-means based clustering approach for data aggregation in periodic sensor networks. in 2014 IEEE 10th international conference on wireless and mobile computing, networking and communications (WiMob). 2014. IEEE.

[33] Li, X.-Y., Y. Wang, and Y. Wang, Complexity of data collection, aggregation, and selection for wireless sensor networks. IEEE Transactions on Computers, 2010. 60(3): p. 386-399.

[34] 34.Carlos-Mancilla, M., E. López-Mellado, and M. Siller, Wireless sensor networks formation: approaches and techniques. Journal of Sensors, 2016. 2016.

[35] Aderohunmu, F.A., D. Brunelli, J.D. Deng, and M.K. Purvis, A data acquisition protocol for a reactive wireless sensor network monitoring application. Sensors, 2015. 15(5): p. 10221-10254.

[36] 36.Rault, T., A. Bouabdallah, and Y. Challal, Energy efficiency in wireless sensor networks: A top-down survey. Computer Networks, 2014. 67: p. 104-122.

[37] C. B. Mudgule, U.N., & P. D. Ganjewar, Data Compression in Wireless Sensor Network: A Survey. International Journal of Innovative Research in Computer and Communication Engineering, 2014. 2: p. 1-10.

[38] Pandey, I.K., M. Natarajan, and S. Kaur-Ghumaan, Hydrogen generation: aromatic dithiolate-bridged metal carbonyl complexes as hydrogenase catalytic site models. Journal of inorganic biochemistry, 2015. 143: p. 88-110.

[39] Hung, C.-C., W.-C. Peng, and W.-C. Lee, Energy-aware set-covering approaches for approximate data collection in wireless sensor networks. IEEE Transactions on Knowledge and Data Engineering, 2011. 24(11): p. 1993-2007.

[40] Sarisaray-Boluk, P. and K. Akkaya, Performance comparison of data reduction techniques for wireless multimedia sensor network applications. International Journal of Distributed Sensor Networks, 2015. 11(8): p. 873495.

[41] Wu, M., L. Tan, and N. Xiong, Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications. Information Sciences, 2016. 329: p. 800-818.

[42] Aït-Sahalia, Y. and D. Xiu, Principal component analysis of high-frequency data. Journal of the American Statistical Association, 2019. 114(525): p. 287-303.

[43] McCorrie, D.J., E. Gaura, K. Burnham, N. Poole, and R. Hazelden, Predictive data reduction in wireless sensor networks using selective filtering for engine monitoring, in Wireless Sensor and Mobile Ad-Hoc Networks. 2015, Springer. p. 129-148.

[44] de Carvalho, C.G.N., D.G. Gomes, J.N. de Souza, and N. Agoulmine. Multiple linear regression to improve prediction accuracy in WSN data reduction. in 2011 7th Latin American Network Operations and Management Symposium. 2011. IEEE.

[45] Cheng, S., Z. Cai, and J. Li, Approximate sensory data collection: a survey. Sensors, 2017. 17(3): p. 564.

An Investigation of a Convolution Neural Network Architecture for Detecting Distracted Pedestrians

Igor Grishchenko[1], El Sayed Mahmoud[2]
Faculty of Applied Science and Technology
Sheridan College, Oakville, Canada

***Abstract*—The risk of pedestrian accidents has increased due to the distracted walking increase. The research in the autonomous vehicles industry aims to minimize this risk by enhancing the route planning to produce safer routes. Detecting distracted pedestrians plays a significant role in identifying safer routes and hence decreases pedestrian accident risk. Thus, this research aims to investigate how to use the convolutional neural networks for building an algorithm that significantly improves the accuracy of detecting distracted pedestrians based on gathered cues. Particularly, this research involves the analysis of pedestrian' images to identify distracted pedestrians who are not paying attention when crossing the road. This work tested three different architectures of convolutional neural networks. These architectures are Basic, Deep, and AlexNet. The performance of the three architectures was evaluated based on two datasets. The first is a new training dataset called SCIT and created by this work based on recorded videos of volunteers from Sheridan College Institute of Technology. The second is a public dataset called PETA, which was made up of images with various resolutions. The ConvNet model with the Deep architecture outperformed the Basic and AlexNet architectures in detecting distracted pedestrian.**

***Keywords*—Convolutional neural networks; computer vision; cognitive load; distractive behavior**

I. Introduction

Pedestrians are the most vulnerable objects observed by autonomous vehicles because they travel along streets, roads, sidewalks, alone and with others in both busy and idle areas. According to the Canadian Motor Vehicle Traffic Collision Statistics 2015 [1], pedestrians accounted for 15.4% of fatalities and 14.3% of serious injuries in all motor vehicle accidents. Pedestrians can make changes in their path because many roads and streets cannot have physical constraints that ensure pedestrians use the appropriate behavior all the time. This makes planning a safe route challenging even with all the current technologies equipped to self-driving cars today.

One of the main reasons for the difficulties in detecting and predicting pedestrian behavior is attributed to the use of mobile devices while walking. Pedestrians who use handheld devices tend to walk blindly into the path of a moving vehicle. Doing so increases the likelihood of a collision. Using devices while walking limits pedestrian cognitive functions which in turn could lead to walking with high risk to cause the accident [2].

The use of handheld devices by pedestrians affects their cognitive load and the ability to pay close attention to the road, thus, increases the car accident risk. This creates a further challenge for the self-driving car to plan the safest route because the walking path of a distracted pedestrian is not related to the current road conditions. Identifying pedestrians who use cell phones during their walking will significantly decrease the number of injuries and deaths due to distracted pedestrians. This study developed and trained a Convolutional Neural Network (CNN) to detect pedestrians who use handheld devices while crossing the road. Ultimately, this work developed the distracted pedestrian detector, based on convolutional neural networks, which is able to analyze whether the pedestrian is distracted or not in real-time.

A. Motivation

The motivation of this thesis is to improve the safety of pedestrians by leveraging the convolutional neural networks. The application of convolutional neural networks could improve the accuracy of detecting pedestrians and identify if they are distracted. The ConvNet investigates image structural information and builds the neural network model in a more insightful manner than non-deep neural networks [3]. Today, research on the detection of pedestrian motions and route planning is conducted frequently with many readily available publications. However, only few mention the fact that pedestrians can be distracted and how their behavior and movement can and may change unexpectedly due to cognitive dissonance. This study investigates the problem of distracted pedestrians by implementing the detector based on the ConvNets, which can identify whether the observed pedestrian is holding the handheld device or not. Stakeholders who benefit from the proposed algorithm are the vehicle manufactures, smart cities project teams, and researchers. As mentioned previously, drivers are also distracted by handheld devices as well. Thus, the developed algorithm can also be applied to warn a driver if a pedestrian is distracted and the chance of accident will overall decrease. The main goal of this work is to improve the accuracy of automated vehicles to make their choices safer and minimize the possibility of injury.

B. Organization of Paper

The rest of this paper is organized as follows, the literature review chapter covers prior researches related to autonomous vehicles and the detection of pedestrians by examining different techniques such as neural nets (MLP), knowledge extractions, and model tuning. It consists of studies that focus on human cognition research and how handheld devices can lead to unwilling motions while walking. The methodology chapter focuses on describing what methodology was used and how it was applied in detail. This involves the selection of

ConvNet architecture, model training and tuning as well as testing the detector on the videos of participants. Lastly, the results chapter presents the gathered experimental findings, a review of the findings with analysis and future research opportunities.

II. Literature Review

With the sharp growth of self-driving cars in the automotive industry and the increasing usage of handheld devices by pedestrians, the ability for autonomous vehicles to detect distracted pedestrians has become prevalent, hence receiving a considerable amount of attention and extensive research on determining whether the pedestrian is distracted or not [3] [4].

Many research groups concentrated on the challenge of determining the limb positioning of a pedestrian for a long time and introduced a variety of models. Some studies applied classical machine learning algorithms by fitting labeled data into models, such as Gaussian process (GP) regression [5], Support Vector Machines (SVM) [6], and Mixed Markov-Chain Model (MMCM) [7]. Other groups conducted research considering deep neural networks. Dominguez-Sanchez et al. conducted research for the improvement of pedestrians' motions detection by leveraging convolutional neural network (CNN) [3]. Another approach proposed by Yamashita et al. involves the use of Multi-Task Convolutional Neural Network for the detection of pedestrians and the position of their limbs simultaneously [8]. The latter two approaches will be considered the closest to this study and will be the focus of this study's research.

It is essential to detect distracted pedestrians since it can help to prevent vehicle conflicts and reduce vehicle traffic due to indecisions when crossing and overall slower crossing speed [9]. According to Zaki et al., this type of research would benefit multiple domains which include road safety which extends the application of computer vision (CV). The potential improvement of the current methodology for identifying distracted pedestrians would be the exploration of head and hands positional tracking [9].

With the growth of autonomous cars in the motor vehicle industry and the increasing number of distracted pedestrians, the importance of this research as well as the understanding and analysis of the distracted walking behavior of pedestrians have been more than reaffirmed. Recent studies about the exploration of pedestrians' gait benchmarks for the identification of whether they are distracted or not has been completed [9].

A survey of theory and practice in the interaction between self-driving cars and pedestrians conducted by Rasouli et al. showed that pedestrians who are distracted by handheld devices are 75% more likely to display unintentional blindness [10]. Another study conducted by Neider et al. investigated that distraction arising from the cell phone usage challenges pedestrians' ability to estimate the time-to-contact of traffic accurately, which increases the odds of failing to cross a road safely. Fig. 1 visualizes the results gathered by Neider et al. during the research experiments and shows the percentage of attempts in which participants successfully crossed the street [11]. Fig. 1 demonstrates that pedestrians who were talking on the phone while crossing the street were less likely to successfully cross the road compared to non-distracted pedestrians [11].

Distracted pedestrians tend to change their walking direction more often and on average, cross the street slower than undistracted pedestrians, which can lead to unwilling accidents [10] [9]. The ability of autonomous cars to detect pedestrians who are not paying attention while crossing the road can improve road safety. Since the motor vehicle industry is steadily shifting towards self-driving cars, these autonomous cars must recognize if a pedestrian is not paying attention to the road, in order to prevent any hazards associated with distraction [12]. Current studies focus on analyzing pose and extracting gait parameters of pedestrians to determine whether the pedestrian(s) is distracted or not [12] [9].

This study's intention is to improve self-driving cars' accuracy in collisions detection and path planning by identifying whether the pedestrians are distracted or not. The main goal of this work is to use a convolutional neural network model to detect distracted pedestrians by examining specific distracted behavior scenarios of pedestrians.

A. Convolutional Neural Networks in Computer Vision

Deep Convolutional Neural Networks (ConvNet) has demonstrated amazing performance in several computer vision tasks, including face recognition, digits recognition, and image classification, due to the ability to extract visual benchmarks from the pixel-level content [13]. However, it was a great challenge to train the deep ConvNets due to the lack of training data and computational power in the past, but many methods had been proposed to overcome this problem since 2006 [14]. In 2012, Krizhevsky et al. proposed a classic ConvNet architecture, AlexNet, and demonstrated notable improvements in the image classification tasks [15]. AlexNet showed high levels of accuracy in image recognition applications and received considerable attention from the community, and therefore, many studies were conducted to improve or even surpass AlexNet's performance. Subsequently, more effective and deeper ConvNet architectures were proposed: ZFNet, VGGNet, GoogleNet, and ResNet [14]. The typical modification of these new architectures was the increased depth in order to extract even more features from the input. Furthermore, deep ConvNets were successfully applied for pedestrians' detection problems by estimating the movement of their limbs [16] [3].

The research by Lu et al. examined the application of convolutional neural networks for player detection and team classification in group sports such as basketball, ice hockey, and soccer from broadcasting videos [17]. They also experimented on a pedestrian dataset to evaluate the generality of their approach. Their model performed very well and was able to classify each team in different sports with 97% accuracy. Table I shows the confusion matrix of the percentage of players being classified by teams in the 4 different data sets [17]. Table I represents the proportion of players in each team being classified into the corresponding team. Classes TA, TB, and O refer to Team A, Team B, and Others accordingly.

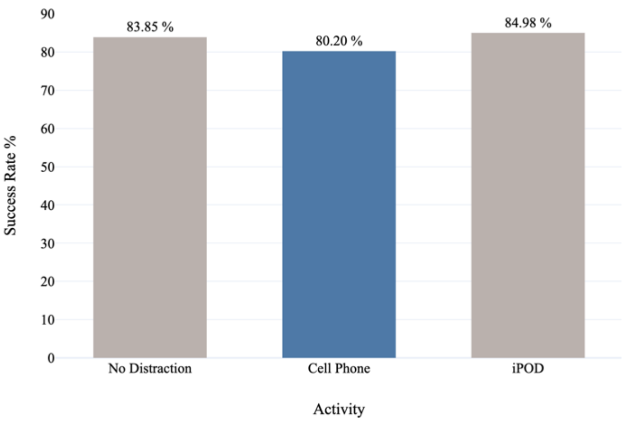


Fig. 1. The Percentage of Pedestrians' Success Crossing While being Distracted or not.

The study conducted by Dominguez-Sanchez et al. evaluated the ability and performance of the current convolutional neural networks and proved how CNNs can impressively perform an estimation task of determining limbs movement of a pedestrian. During the research, they trained their networks with their own novel video dataset which was processed into frames through the image preprocessing pipeline. Only one of every six frames were used for the input during experiments of pedestrians' limb position and movement detection. After the evaluation of AlexNet, GoogleNet, and ResNet architectures, they identified that ResNet was the best for pedestrians' movement recognition and demonstrated 79% accuracy in the test set. Table II illustrates the results obtained by the ResNet in the test set [3].

TABLE. I. CONFUSION MATRIX OF TEAM MEMBERSHIP CLASSIFICATION IN 4 DATASETS

Dataset	Classes	**TA**	**TB**	**O**
Basketball	TA	**99.65**	0.18	0.17
	TB	0.91	**97.88**	1.21
	O	0.86	1.71	**97.43**
Ice Hockey	TA	**98.91**	0.72	0.37
	TB	1.33	**97.99**	0.68
	O	0.69	1.36	**97.95**
Soccer Set 1	TA	**98.63**	0.24	1.13
	TB	0.83	**98.23**	0.94
	O	2.08	1.41	**96.51**
Soccer Set 2	TA	**98.33**	0.44	1.23
	TB	0.91	**97.78**	1.31
	O	2.46	1.37	**96.17**

TABLE. II. CONFUSION MATRIX OF RESULTS OBTAINED BY RESNET

	Front	**Left**	**Right**
Front	**0.980**	0.011	0.008
Left	0.058	**0.841**	0.100
Right	0.081	0.265	**0.652**

Abdulnabi et al. introduced a modified deep convolutional neural network architecture that enables multitasking, so different CNNs can share knowledge among each other [18]. Their learned Multi-Task CNN demonstrated better performance in predicting semantic binary attributes by sharing visual knowledge between tasks. The results obtained from experiments on two different datasets and multiple different CNNs shows that Multi-Task CNN used by Abdulnabi et al. outperformed single-task neural networks and achieved 92% accuracy in attribute predictions in images [18] Deep convolutional neural networks demonstrated amazing performance in pedestrians and attribute detection and were selected as the approach for this research.

III. DATA SOURCES

One of the data sources used in this research was built by recording student volunteers from the Sheridan College Institute of Technology (SCIT dataset). Recording students' videos to create a dataset was approved by the Sheridan Research Ethics Board. The total number of participants was 15 with different demographics such as gender, race, and age which allowed us to construct a good quality diverse dataset. The videos were recorded in an enclosed environment where each participant was asked to mimic a distracted/non-distracted pedestrian, based on the attributes listed in Table III while crossing the road. These video recordings of their walk were incorporated into the training set and further used for this study. The volunteers were recorded from three different positions for both front and rear views in order to capture every possible angle, direction, and position. Then, all the video footage was split into frames and labeled based on the participants' behavior to differentiate distracted and non-distracted scenarios. Each participant had around 350 frames per each activity, thus, we formed 350 × 15 ≈ 5,000 images per activity after data preprocessing.

Another data source was built from Composition of PEdesTrian Attribute (PETA) dataset with 19000 images, with the image size ranging from 17 × 39 pixels to 169 × 365 pixels, which were released by Deng et al. during their research [19]. They also provided attribute annotations for each image in order to perform benchmarks detection. Yet, their dataset did not provide any labels whether the person on an image is distracted or not. Thus, all the images were reviewed and classified manually to fit the purpose of this research.

A. Determining Pedestrians Distracted behavior Scenarios

After collecting data based on walking pedestrians, all the images were broken down into two classes: *distracted* and *non-distracted* pedestrians. The literature has been explored to identify what type of behavior can cause cognitive load and result in an unsafe road crossing. According to research conducted by Mwakalonge et al., 75% of pedestrians who were walking while taking on a cell phone displayed inattention blindness and failed to notice unusual activity [20]. Another study by Neider et al. performed the experiment in a virtual pedestrian environment and determined that participants who were distracted by music or texting were more likely to be hit by an automobile [5]. 5 different scenarios were identified where a pedestrian is considered to be distracted based on their hands and head positioning. Table III provides an overview of those scenarios as well as example images from the SCIT dataset. Then, PETA dataset images that fall under the identified scenarios were manually moved to a different directory to be separated from the images that were identified as non-distracted pedestrians. As for the SCIT dataset, all the videotaped volunteers were asked to mimic distracted and non-distracted behavior before the recording, thus, all the data were already structured and easily distributed in two classes. Also, each distracted and non-distracted scenario was recorded from different views to simulate real-life situations as much as possible.

IV. Methodology

The development phases for the proposed detector include: (i) identifying the appropriate sample size to train an accurate ConvNet image recognition classifier, (ii) datasets preprocessing to improve the quality of the data, and (iii) designing a ConvNet architecture and fine-tuning hyper-parameters to get the accurate classifier.

A. Identifying Appropriate Sample Size

The most effective dataset size to accurately train a ConvNet model is determined iteratively and can be guided by the distribution of classes and their behaviors. Therefore, it is not clearly defined which sample size would to train an accurate ConvNet pedestrian classifier. Li et al. used the Caltech-101 dataset which contains 9,144 images with a variety of classes to train and test their CNN image classifier and achieved 89% accuracy [21]. The samples of 4,000 images and 30,000 images of distracted and non-distracted pedestrians were gathered from the PETA and SCIT datasets accordingly. However, the whole number of images in the SCIT dataset was not used in the experiments since this number is calculated based on the number of images for each behavior example where we have 5,000 images per scenario. Therefore, we used all the images from the non-distracted scenario set to create the first-class and randomly selected 1,000 images from each of the distracted scenarios sets to create the second class. Eventually, we constructed the dataset of 10,000 images of distracted and non-distracted classes based on the SCIT data.

B. Preprocessing of SCIT and PETA datasets

Before training the detector and conducting different experiments, people were cropped from the frames in the SCIT dataset gathered by our experiment. A pretrained Mask R-CNN object detector was used to detect people in each image and annotate their bounding boxes to perform the cropping. The resolution of the cropped pedestrian images is ranging from 62 × 224 pixels to 494 × 987 pixels in the SCIT dataset. The amount of blur in each image was also computed in order to remove images with excessive amounts of blurring that improved the dataset quality. Further, data augmentation techniques were applied to both PETA and SCIT datasets in order to increase the size of the datasets. Particularly, we augmented our data by rescaling, zoom-range, and fill-mode.

C. Determining CNN Architecture and Fine-Tuning

Convolutional Neural Networks have been selected due to their convolution layers which extract features from an input image and learn from them by exploiting small chunks of input data in order to preserve the spatial relationship between them.

TABLE. III. Description of Scenarios when Walking Pedestrian is Distracted

Scenario Description
Head down and holding the phone with the left hand. A participant is chatting on the phone.
Head down and holding the phone with the right hand. A participant is chatting on the phone.
Head down and holding the phone with both hands. A participant is chatting on the phone.
The left hand is near the head. A participant is speaking over the phone.
The right hand is near the head. A participant is speaking over the phone.

We proposed two architectures Basic and Deep with 3 and 5 convolutional layers accordingly to undertake the problem of distracted pedestrian detection:

The first architecture has the following structure: The first convolutional layer has 16 filters of size 3 with ReLU activation function followed by batch normalization and max-pooling layer of size 2×2; the second convolutional layer has 32 filters of size 3 with Tanh activation function followed by batch normalization and max-pooling layer of size 2×2; the third convolutional layer has 64 filters of size 3 with ReLU activation function followed by batch normalization and max-pooling layer of size 2×2. The last max-pooling layer is followed by the dropout layer with a 25% dropout rate. After the aforementioned layers, we have flatten layer followed by two dense which also called fully connected layers. The first dense layer has 64 nodes with the ReLU activation function and the second has only 2 nodes with Sigmoid activation function since we need to find a probability of the pedestrian being distracted or not. This architecture is presented on the left side of Fig. 2.

The second architecture is the modification of the above one where the second and third layers were duplicated such that two convolutional layers are stacked together before every max-pooling layer. Multiple stacked convolutional layers can be able to learn more complex features from the input before the destructive max-pooling layer [22]. We considered this technique to be promising in the detection of distracted pedestrian problem. The second architecture is shown on the right side of Fig. 2.

We applied the same hyper-parameters to both architectures; we used RMSprop optimizer with default parameters: learning rate = 0.001 and $\beta = 0.9$. The loss function we selected was the binary cross-entropy since this function better suits classification tasks with 2 classes [23]. All the convolutional layers were preceded by the zero or “same” padding to preserve the size of post convolution. Finally, we applied the early stopping regularization technique to prevent the model from overfitting.

D. Testing Strategy

The detector was tested with randomly selected images of distracted and non-distracted pedestrians which have not been seen by the model during training. Since SCIT data consists of 15 different participants, we randomly selected 4 participants and their images to generate the test set. The data of the other 11 participants were used for training. This 11/4 split is equivalent to a 75/25 data split, where 75% of data was used to train the model and the other 25% was used to test the model. This approach allowed us to always test our model on the people's data which the model had never seen before. Regarding the PETA dataset, since most of its data points represent a unique pedestrian, we randomly split data following the same 75/25 approach. Besides, the data in both datasets was always shuffled every time when we trained a new version of the model in order to reduce variance, make sure that the model remains general, and prevent overfitting. We conducted an experiment to examine how both our architectures can perform on different combinations of datasets, which drastically different in the resolution of the images. AlexNet architecture was also evaluated on the same datasets to compare it with our proposed architectures.

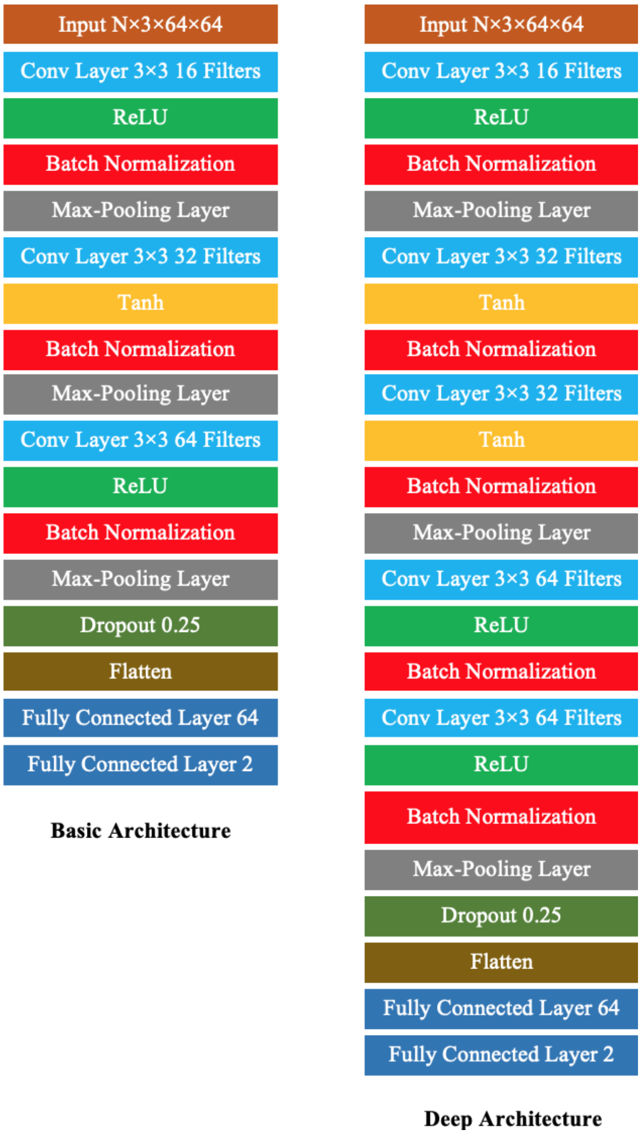


Fig. 2. Architectures of Distracted Pedestrians Detector.

E. *Proposed Experiment*

The purpose of the experiment was to see how the quality of the images would affect the performance of the ConvNet based on different architectures. Therefore, we created three different sample sets from the SCIT and PETA datasets for this test. The first sample was made of only the SCIT dataset where all the images had high resolution (62 × 224 pixels to 494 × 987 pixels) and distraction scenarios were equally distributed. The second sample was constructed from the PETA dataset and its images had a relatively low number of pixels (17 × 39 pixels to 169 × 365 pixels). The third data sample was created using both SCIT and PETA dataset where high and low image resolution (17 × 39 pixels to 494 × 987 pixels) were combined. The purpose of the third sample was to see whether the ConvNet accuracy would degrade or not if we feed data to it which has a huge range in quality to it.

The models with Basic and Deep architectures were trained and tested on the aforementioned datasets. We also investigated how AlexNet architecture that achieved state-of-the-art results in many computer vision tasks would tackle the distracted pedestrian detection problem [24]. AlexNet is a much deeper network with more filters in each convolutional layer. The model with AlexNet architecture was also trained on the same data samples, so we could compare its performance with our Basic and Deep architectures. The reason why the AlexNet had been also evaluated was to examine if the deeper network with more filters would be smarter in the feature extraction related to our problem and would have better accuracy in distracted pedestrian detection. Fig. 3 illustrates the design of the experiment.

V. Results and Analysis

This section shows the experimental results of building the Distracted Pedestrian Detector based on different combinations of the datasets: SCIT, PETA, and a combination of both. This work tested two different ConvNet architectures. The first is called Basic, and the second is called Deep, which duplicates the second and third layers of the Basic architecture. Additionally, we examined how the AlexNet model would tackle the distracted pedestrian detection problem based on the combinations of the aforementioned datasets. The Deep ConvNet architecture was more efficient than the Basic and AlexNet architectures in detecting the distracted pedestrians based on all three datasets.

A. *Effect of the Image Resolution on the Performance*

The highest accuracy of the Distracted Pedestrian Detector with Deep architecture for the SCIT dataset was 95.11%. Fig. 4 shows the average accuracies of the Deep, Basic, and AlexNet architectures trained and tested on the SCIT data sample. Since the SCIT datasets had the highest resolution, this particular evaluation demonstrates how the architectures behave on images with a big number of pixels. The Deep architecture also showed the highest average 94.02% accuracy. The Basic architecture was the second in the accuracy and achieved 90.00% on average. Lastly, the performance of AlexNet was close to the Basic architecture but demonstrated lower average accuracy – 89.23%. Based on the high precision and recall scores, shown in Table IV, we can see that all the models trained on the SCIT data were able to correctly classify a high number of the relative data points. This is supported by the f1 score since it was also relatively high too, meaning that models were general and unbiased. This was due to the SCIT dataset being well distributed and provided the models with balanced training and testing data. We can see that all the architectures performed relatively well on the dataset which contains images with high resolution.

When trained and validated on the PETA dataset, all the architectures demonstrated lower accuracies. This can be explained by a certainly low resolution of images in the PETA dataset. Fig. 5 visualizes the average accuracies achieved by the Deep, Basic, and AlexNet architectures trained and tested on the PETA data sample. The Deep architecture maintained the first place and showed an average 85.44% accuracy. The AlexNet architecture had the 83.67% accuracy on average what was close to the Deep one. Yet, the Basic architecture demonstrated the biggest reduction in accuracy and achieved 78.01% what notably different from the score of Deep and AlexNet architectures. The Basic ConvNet had the smallest number of convolutional layers and, therefore, the minimum number of filters. It performed relatively bad in distinguishing between distracted and non-distracted pedestrians. We tried to increase the number of filters in each convolutional layer by 4 times such that it had 64 filters in the first layer, 128 filters in the second layer, and 256 in the third layer. Unfortunately, this only worsened the architecture, because the high number of filters caused model overfitting since the training accuracy was 97.15% while the validation accuracy was only 76.34%. This indicates that the three convolutional layers are not enough to deal with images with a small number of pixels.

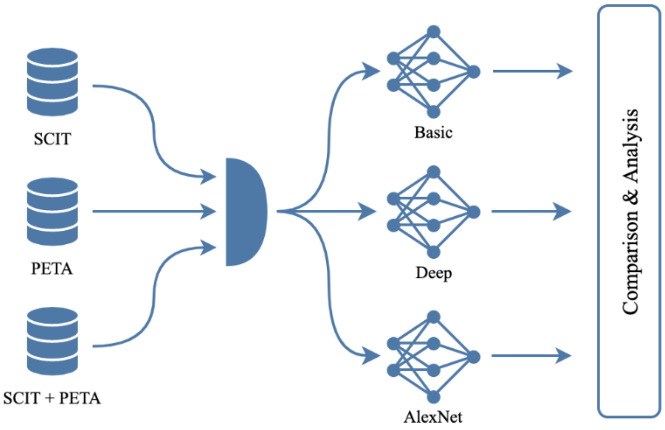


Fig. 3. Experiment Design.

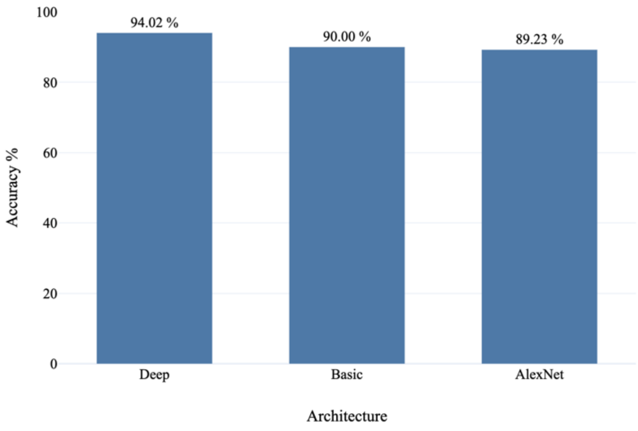


Fig. 4. Average Accuracy of Architectures for SCIT Dataset.

TABLE. IV. AVERAGE PRECISION, RECALL, AND F1 SCORE METRICS OF MODELS FOR SCIT DATASET

	Precision	Recall	F1 Score
Deep	0.9434	0.9374	0.9403
Basic	0.9246	0.8812	0.9024
AlexNet	0.8826	0.9000	0.8912

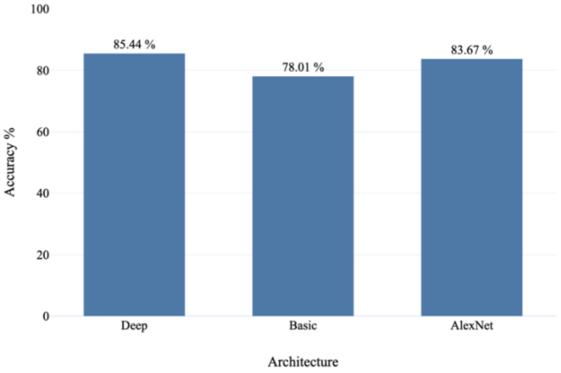


Fig. 5. Average Accuracy of Architectures for PETA Dataset.

If we analyze the precision, recall, and f score metrics, demonstrated in Table V, we can see that the recall metric significantly dropped compared to the precision metric. It means that the models evaluated on the PETA data classified more distracted pedestrians as non-distracted. We then can conclude that the data with low-quality images did not allow models to learn enough patterns, since it was relative to the distracted behavior. Also, some of the images were captured from a distance making it really difficult for the models to detect if an observed pedestrian is holding a handheld device or not.

The third dataset which was used for the evaluation of the architectures was the combination of both SCIT and PETA data. The highest accuracy was demonstrated by the Deep architecture which achieved 88.78%. The average accuracy of the Deep, Basic, and AlexNet architectures trained and evaluated on the combination of SCIT and PETA datasets is shown in Fig. 6. The Deep architecture, again, showed the best average accuracy – 87.01%. The accuracies of AlexNet and Basic architectures were 84.32% and 80.56%, respectively. All the architectures did not improve much, and their average accuracies were approximately 2% better compared with the models trained and tested only on the PETA dataset. These results illustrate that even if we combine the images with low and high resolutions, the images with a low number of pixels in the set still affects the ability of ConvNet accurately detect distracted pedestrians. Besides, the big range of the resolution could also be a reason for the not significant improvement of the architectures. ConvNets could not establish a clear pattern from the extracted features to find the difference between distracted and non-distracted scenarios.

Table VI shows the precision, recall, and f1 score metrics obtained by the ConvNet models trained and tested on the combination of both SCIT and PETA datasets. It is clear that if we add high-quality images to the dataset that contains images with a low number of pixels, the models can learn more features and distinguish distracted and non-distracted pedestrians with better accuracy. However, the following metrics are still lower compared to the obtained metrics in Table IV, which demonstrates again, that data with low-quality images has a big influence on the architectures, even if data points with a big number of pixels are dominant in this dataset.

Since the model based on the Deep architecture demonstrated higher accuracy across all three datasets, the one-way analysis of variance (ANOVA) test was used to determine if the Deep architecture's score is significantly different from the Basic and AlexNet models. The ANOVA test was conducted on three different sets of models trained on the different datasets as shown in Fig. 4, Fig. 5, and Fig. 6. The *p-value* from the three test results was the following: 0.00003, 0.000025, 0.000027 for the sets of models trained on the SCIT, PETA, and combination of SCIT and PETA datasets, accordingly. Since the *p-value* across all the datasets was less than 0.05, this indicates that the models' accuracies were significantly different and not from the same. Thus, we can conclude that the difference in the model's scores is significant showing that the Deep actually had the highest accuracy.

B. Impact of Architecture Design

We also inspected the filters and feature maps during the layers' convolution of Basic and Deep ConvNet architectures. Since Deep architecture was designed to have the second and third layers combined together followed by the max-pooling layer, the third layer was able to receive a more precise feature map where we still can recognize the original image as shown in Fig. 7. In contrast, all the convolutional layers in Basic architecture are split by max-pooling layer, therefore, the feature map of the third layer in the Basic architecture is less interpretable and contains high-level concepts as displayed in Fig. 8. From Fig. 7 and Fig. 8, we can see that the feature map in the third convolutional layer of the Deep architecture still contains visual concepts like edges, which are useful for our problem since the detector needs to evaluate the position of the pedestrian limbs to differentiate distracted and non-distracted behavior. While the feature map in the third layer of the Basic architecture looks more like the abstraction of the original image and contains high-level features that might have more information about small parts of the image such as a mobile device in the hands. Of course, both low and high-level features are highly important to accurately detect distracted pedestrians. Though, the design of the Deep architecture allowed filters to extract more low-level features that helped ConvNet to characterize the position of pedestrian limbs and better recognize the distractive action. This explains why ConvNet with Deep architecture outperformed the Basic ConvNet across all the three datasets since the Basic architecture could not extract enough features related to the pedestrians' actions.

TABLE. V. AVERAGE PRECISION, RECALL, AND F1 SCORE METRICS OF MODELS FOR PETA DATASET

	Precision	Recall	F1 Score
Deep	0.8990	0.8251	0.8604
Basic	0.8780	0.7341	0.7996
AlexNet	0.8915	0.8024	0.8446

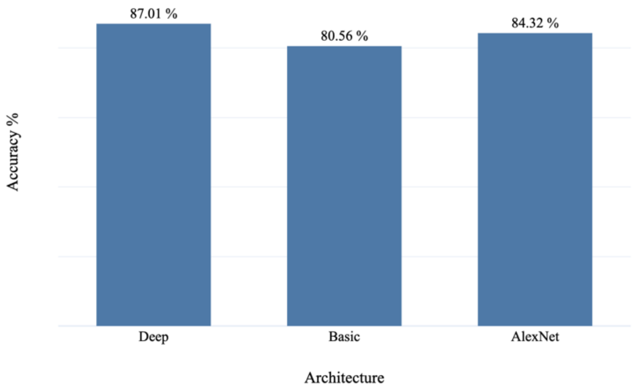


Fig. 6. Average Accuracy of Models for SCIT and PETA Datasets.

TABLE. VI. AVERAGE PRECISION, RECALL, AND F1 SCORE METRICS FOR THE COMBINATION OF SCIT AND PETA DATASETS

	Precision	Recall	F1 Score
Deep	0.8888	0.8566	0.8724
Basic	0.8272	0.7929	0.8097
AlexNet	0.8685	0.8266	0.8470

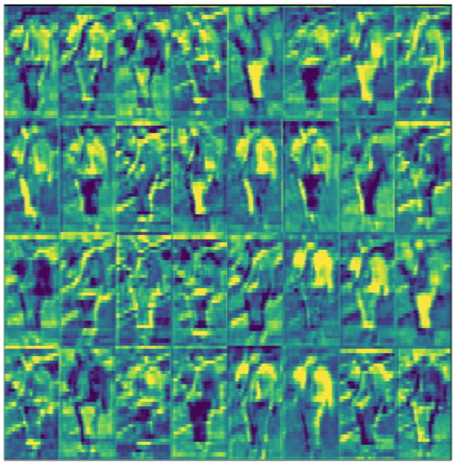


Fig. 7. Visualization of the Filters in the Third Conv Layer of the Deep Architecture.

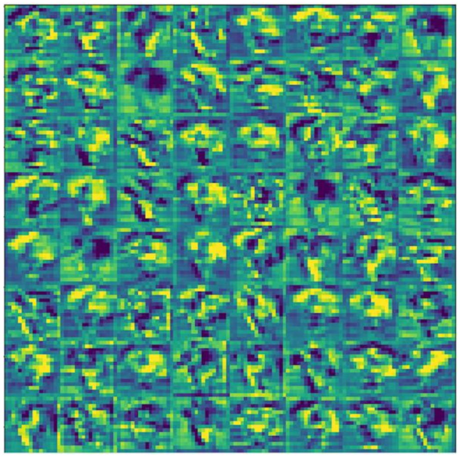


Fig. 8. Visualization of the Filters in the Third Conv Layer of the basic Architecture.

Interestingly enough that Deep and AlexNet architectures had a similar design in terms of the combined convolutional layers. While the Deep architecture combined the second with third and the fourth with fifth layers, the AlexNet architecture design combined the third, fourth, and fifth convolutional layers without max-pooling layers between them. But based on the gathered results demonstrated above, the Deep architecture achieved higher average accuracies across all the three datasets. Despite the fact, that even if AlexNet has a similar structure to the Deep architecture, its combined convolutional layers focused mostly on the extraction of the high-level features since they were the last group and received feature maps that already got through multiple max-pooling layers. Therefore, AlexNet could not extract more low-level features like the Deep architecture. This derives the conclusion that the low-level features which are responsible for the detection of edges and shapes played a very important role in the distracted pedestrian detection problem and allowed the Deep architecture to outperform the AlexNet and Basic ConvNets.

VI. Conclusion

This research aimed to explore the application of convolutional neural networks to address the problem of detecting distracted pedestrians automatically. This work investigated various combinations of CNN architectures and datasets to build an effective distracted pedestrian detector. A novel training dataset was created from video recordings of volunteer participants from the Sheridan College Institute of Technology when they acted as distracted and non-distracted pedestrians. This dataset is called SCIT and could be used for further research in various computer vision research problems related to human detection. Three ConvNet models were implemented with different architectures: Basic, Deep, and AlexNet. Each model was trained and tested on three different datasets: SCIT, PETA, and the combination of both. The results from the experiment had indicated that the model that utilized the Deep architecture had outperformed the other models that used the Basic and AlexNet architectures when applied to all the datasets. The developed detector could be used for autonomous vehicles and driver alert systems to identify distracted pedestrians who cross the street and minimize the probability of injury. The detector would also be useful for the variety of stakeholders including the vehicle manufactures, researchers, and smart cities project teams.

VII. Future Work

The detector currently takes an entire image and makes a prediction based on the extracted features. The next step will be to modify the algorithm so that it would extract pedestrian limbs such as head and hands from each image and evaluate them independently instead of analyzing a complete image. This modification will increase the efficiency of the system because it will minimize the misclassification of handheld devices with other potential objects in the pedestrian’s hands. An analysis of how a pedestrian’s head direction changes would also create a meaningful impact on when identifying if a pedestrian is distracted.

Predicting the route of a distracted pedestrian will be another perspective addition to the system. Distracted pedestrians tend to change their route unexpectedly what increases the possibility of an accident. With the knowledge that a pedestrian is distracted, his/her long-term path could be predicted more accurately. The information about pedestrians’ future path and if they are distracted or not could advance the safe route planning for self-driving cars.

Sequential frame classification can be another improvement to the detector. In this case, extraction of the sequence features, which are also called temporal or time-related features, will be required in addition to the features of the images. This approach could help identify when a pedestrian had acted similar to a distracting behavior for a short period of time when the pedestrian's action was not an actual distraction. This could reduce the number of false positives that would improve the reliability of the detector.

Acknowledgment

This work has resulted in Igor Grishchenko's undergraduate thesis of the Honours Bachelor of Computer Science (Mobile Computing). The collection of data from human participants was approved by the Sheridan Research Ethics Board. We are grateful for the insightful feedback on the work received from the thesis advisory committee members at Sheridan College.

References

[1] Transport Canada, “Canadian Motor Vehicle Traffic Collision Statistics: 2017,” Transport Canada, 27-Feb-2019. [Online]. Available: https://www.tc.gc.ca/eng/motorvehiclesafety/canadian-motor-vehicle-traffic-collision-statistics-2017.html. [Accessed: 18-Nov-2019].

[2] G. Yogev-Seligmann, J. M. Hausdorff, and N. Giladi, “Do we always prioritize balance when walking? Towards an integrated model of task prioritization,” Movement Disorders, vol. 27, no. 6, pp. 765–770, 2012.

[3] A. Dominguez-Sanchez, M. Cazorla, and S. Orts-Escolano, “Pedestrian Movement Direction Recognition Using Convolutional Neural Networks,” IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 12, pp. 3540–3548, 2017.

[4] Y. Tang, L. Ma, W. Liu, and W.-S. Zheng, “Long-Term Human Motion Prediction by Modeling Motion Context and Enhancing Motion Dynamics,” Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018.

[5] Y. Chen, M. Liu, S.-Y. Liu, J. Miller, and J. P. How, “Predictive Modeling of Pedestrian Motion Patterns with Bayesian Nonparametrics,” AIAA Guidance, Navigation, and Control Conference, 2016.

[6] J.-T. Wang, D.-B. Chen, H.-Y. Chen, and J.-Y. Yang, “On pedestrian detection and tracking in infrared videos,” Pattern Recognition Letters, vol. 33, no. 6, pp. 775–785, 2012.

[7] A. Asahara, K. Maruyama, A. Sato, and K. Seto, “Pedestrian-movement prediction based on mixed Markov-chain model,” Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS 11, 2011.

[8] T. Yamashita, H. Fukui, Y. Yamauchi, and H. Fujiyoshi, “Pedestrian and part position detection using a regression-based multiple task deep convolutional neural network,” 2016 23rd International Conference on Pattern Recognition (ICPR), 2016.

[9] M. H. Zaki and T. Sayed, “Exploring walking gait features for the automated recognition of distracted pedestrians,” IET Intelligent Transport Systems, vol. 10, no. 2, pp. 106–113, Jan. 2016.

[10] A. Rasouli and J. K. Tsotsos, “Autonomous Vehicles That Interact With Pedestrians: A Survey of Theory and Practice,” IEEE Transactions on Intelligent Transportation Systems, pp. 1–19, 2019.

[11] M. B. Neider, J. S. Mccarley, J. A. Crowell, H. Kaczmarski, and A. F. Kramer, “Pedestrians, vehicles, and cell phones,” Accident Analysis & Prevention, vol. 42, no. 2, pp. 589–594, 2010.

[12] A. Rangesh, E. Ohn-Bar, K. Yuen, and M. M. Trivedi, “Pedestrians and their phones - detecting phone-based activities of pedestrians for

autonomous vehicles," 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), 2016.

[13] D. Tomè, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, and S. Tubaro, "Deep Convolutional Neural Networks for pedestrian detection," Signal Processing: Image Communication, vol. 47, pp. 482–489, 2016.

[14] 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, 2018.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, Nov. 2015.

[16] Y.-L. Hou, Y. Song, X. Hao, Y. Shen, and M. Qian, "Multispectral pedestrian detection based on deep convolutional neural networks," 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 2017.

[17] K. Lu, J. Chen, J. J. Little, and H. Hea, "Lightweight convolutional neural networks for player detection and classification," Computer Vision and Image Understanding, vol. 172, pp. 77–87, 2018.

[18] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-Task CNN Model for Attribute Prediction," IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 1949–1959, 2015.

[19] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian Attribute Recognition At Far Distance," Proceedings of the ACM International Conference on Multimedia - MM 14, 2014.

[20] J. Mwakalonge, S. Siuhi, and J. White, "Distracted walking: Examining the extent to pedestrian safety problems," Journal of Traffic and Transportation Engineering (English Edition), vol. 2, no. 5, pp. 327–337, 2015.

[21] Q. Li, Q. Peng, and C. Yan, "Multiple VLAD Encoding of CNNs for Image Classification," Computing in Science & Engineering, vol. 20, no. 2, pp. 52–63, 2018.

[22] J. B. Ahire, Artificial Neural Networks: The brain behind AI. 2018.

[23] P. Lakhani, D. L. Gray, C. R. Pett, P. Nagy, and G. Shih, "Hello World Deep Learning in Medical Imaging," Journal of Digital Imaging, vol. 31, no. 3, pp. 283–289, Mar. 2018.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.

Performance Tuning of Spade Card Antenna using Mean Average Loss of Backpropagation Neural Network

Irfan Mujahidin[1], Dwi Arman Prasetya[2], Nachrowie[3], Samuel Aji Sena[4], Putri Surya Arinda[5]
Department of Electrical Engineering, University of Merdeka Malang,
Malang, Indonesia [1,2,3,4,5]

***Abstract*—The microstrip antennas have different dimensions to get the desired performance, especially for microstrip antennas that have complex components and dimensions with the performance: the range of frequency at 2.4 GHz until 3.6 GHz, Maximum Power of Gain value is 5.83 dB and the minimum value is 3 dB and Maximum Directivity Value is 6.22 and the minimum value is 3.32. in consequence, needs to fill the demand for a new and the corresponding design as solvent to adaptive matching as tuner the frequency on antenna design that needs requires a complex mathematical method and simulation. This paper has the novel design to tune the performance of spade card microstrip antenna that can operate on the single, dual or multi-band and able to produce circular or linear polarization using Backpropagation Neural Network in order to obtain an optimum design with a backpropagation algorithm as a solution to simplify the design process. As a result, after 20000 epochs the training loss is around 0.044 and the testing loss is around 0.058. The model has a good performance despite only using a few numbers of training data.**

Keywords—Spade card antenna; mean average loss; neural network; performance tuning antenna

I. Introduction

The progress of telecommunication technology has developed very rapidly. This can be seen from the increasing need to obtain information, whether the information in the form of sound, data, pictures, or video, with communication equipment that can be used anywhere and anytime. The most developed technological advancement today is wireless communication, which can support the implementation of a global telecommunication system. One example of the application of wireless communication is WLAN (Wireless Local Area Network) that can meet the needs of access to information and communication that can be used anywhere and anytime[1]. Compared to the Wired Local Area Network that used to use the cable as its transmission medium, WLAN technology is easier in terms of installation, practical and efficient[2].

In the application of wireless communication, the role of the antenna as an electromagnetic component is necessary because of the antenna attend as a means to emit and take electromagnetic waves in which the equipment formation signal is contained. The antenna is a very important component to support the wireless communication system[3]. The latest developments in wireless communication systems require antenna characteristics that have relatively small, flexible and practical shapes.

The microstrip antenna is one kind of the antenna with the exact quality of the required telecommunication circuit. The microstrip antenna is made on a specific substrate material with a radiating element located on one side of the substrate and the other is a conductor layer that acts as a ground plane[4]. Microstrip antennas work on UHF frequency allocations (300 MHz – 3000 MHz) up to X Band (5200 MHz - 10900 MHz) so that microstrip antennas can be used for wireless communication.

Microstrip antennas have different components and dimensions to get different performance after what is needed, especially for microstrip antennas that have complex dimensions and components[5]. in consequence, needs to fill the demand for a new and the corresponding design as solvent to adaptive matching as tuner the frequency on antenna design that needs requires a complex mathematical method and simulation. This paper has the novel design to tune the performance of spade card microstrip antenna that can operate on the single, dual or multi-band and able to produce circular or linear polarization using Backpropagation Neural Network [6][7]. The model has a good performance despite only using a few numbers of training data.

II. Research Method

A. Antenna Design

The antenna constructed in this paper is a microstrip circuit antenna with the following component specifications:

- Dielectric material : FR4

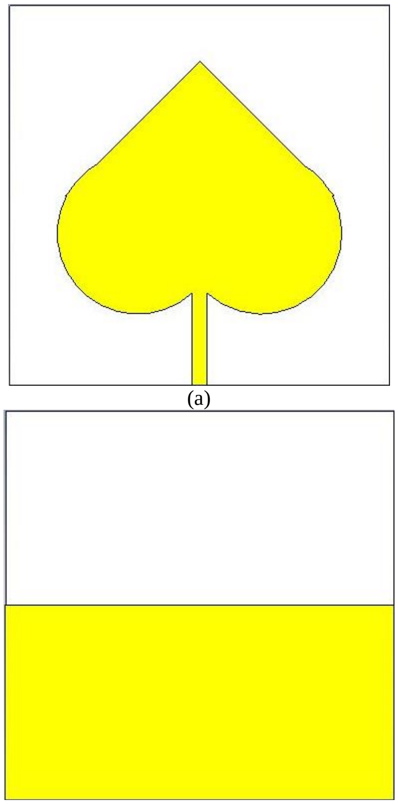
The dielectric value (εr)	= 4.2
The dielectric components thickness (h)	= 1.6 mm

- Seam substrate copper :

The conductor material thickness (t)	= 0.000001 mm
Copper conductivity (σ)	=5.80x107mho m-1
Material size	= 29.7 x 21 mm

After mathematically calculated this research did a simulation and optimization to get the expected results using

CST software. The shape of the ground plane and patches are made to have different sizes to obtain the desired frequency[8]. Shape design and simulation of this antenna is important because it will affect the antenna performance.



(a)

(b)

Fig. 1. Geometric Spade Card Shapes Patch Microstrip Antenna Front (b) Rear

The main part of the antenna is a triangle shape joined with a double circle at the bottom that resembles the shape of a reversed heart, shown in Fig. 1. Part circle on the patch directly connect the feeder line as a link to the patch on the antenna connector[9]. Then the ground plane consists from one form and forms ground plane is basically that is rectangular in addition to the connection with the connector.

In this paper, basically the patch of the antenna has the circular shape, so the component of patch element dimensions with the equation:

$$\alpha = \frac{F}{\left\{1+\frac{2h}{\pi\varepsilon_r F}\left[\ln\left(\frac{\pi F}{2h}\right)+1.7726\right]\right\}^{\frac{1}{2}}}$$

The notation above can be explained that a is the radius of the antenna patch component, h is the substrate thickness, εr is the dielectric permittivity of the substrate and F is the logarithmic function. The transmission line circuit of microstrip element design is finished theoretical based on the matter and any literature [10][11].

B. Datasets

The data used in this paper have been obtained from the CST design results. The antenna simulation using a different combination of ground plane height and the triangle and circle dimensions resulted in 123 samples which have been classified to be more precise. in this research, picked random 90 samples which have been measured and used it as the training data and the rest as the testing data[12][13]. The frequency, return loss, gain and directivity will be saved for further use.

C. Network Achitecture

The implementation uses a three-layer of a Multilayer Perceptron in this paper. In Fig. 2, The input layer has only one node as the frequency, 16 nodes inside the hidden layer to add a non-linearity to model and 6 nodes in the output layer as the representation of the ground plane height, triangle side length, circle diameters, return loss, gain and directivity of the antenna[14][15].

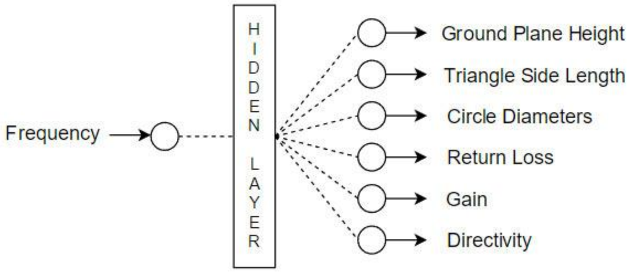


Fig. 2. Neural Network Architecture

The algorithm was experimented using a different-numbers of hidden layers and neurons[16][17]. However, the architecture explained above is the best possible combination and has the best result in minimizing the loss and converge faster than any other combination.

D. Implementation

The training process of backpropagation network consists of two main step. First, training features will be fed to the input layer and will flow through all neuron in the hidden layer. The interconnection between neuron is called synapse. All synapse in the network has its own weight[18][19]. The features will be multiplied with its synapse weight and add a bias value at the same time. After the features arrive at the output layer, the loss will be computed using Mean Average Loss (MAE) (1). Second, the loss value will be minimized using stochastic gradient descent. The loss value will be backpropagated to compute the gradient to minimize the loss and update all synapse weights.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|$$

Neural network model was created using Python and Tensor flow framework. The activation function of all neuron inside the hidden layer are ReLU and the activation function of the output are Linear, because the method do regression task in this experiment[20][21]. Stochastic Gradient Descent (SGD) used as the optimizer with learning rate of 0.0001 and repeated until 20.000 epochs. Training data will feed into the network in batch of 10 data. Thus to run experiments, we compare the results of several experimental samples with the accuracy of the prediction algorithm used. Thus, to measure the success of the implementation of the method used by comparing the results of experiments and comprehensive algorithm predictions.

III. Result and Analysis

After 20.000 Epochs, the analysis evaluates the model using the test data. The analysis evaluates the MAE value and compare that value with the loss of the training process. The training loss is around 0.044 and the testing loss is around 0.058.

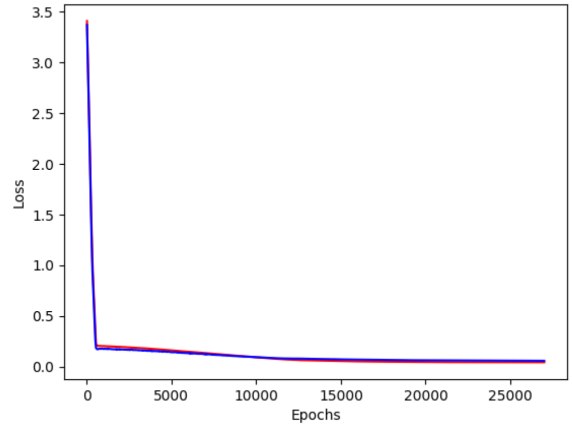


Fig. 3. Training and Testing Loss

Based on Fig. 3 the results provide the comparison of testing data with the output of the neural network for input frequency of 2913MHz in Table I. The model can predict the dimension of the antenna with a very low error [22][23]. But there are some gap in the S-Parameter, Gain and Directivity estimation. Thus, a better quantity and quality of a new dataset are required for further research. To know more results from the results of sampling data can be seen in Fig. 4 in the attachment.

TABLE I. Comparison of Prediction and Truth for 2913 Mhz Input

Parameters	Truth	Prediction
Ground Plane Height	40.1	40.1
Triangle Side Length	20	20
Circle Diameters	17.300	17.333
S-Parameter	-20.230	-21.076
Gain	4.4	4.821
Directivity	5.420	5.364

IV. Conclusion

Backpropagation neural network has been used to design an antenna. By using a correct activation function and proper learning rate, backpropagation neural network shows a promising result for the tune of a spade card shape microstrip antenna: the range of frequency at 2.4 GHz until 3.6 GHz, Maximum Power of Gain value is 5.83 dB and the minimum value is 3 dB and Maximum Directivity Value is 6.22 and the minimum value is 3.32. The model has a good performance despite only using a few numbers of training data. This research can be pushed further to learn using a different shape, substrate, frequency range and other parameters of a microstrip antenna. For future work it is necessary to develop a wider range of antenna frequencies and the need for a combination of algorithms to improve the accuracy of the results.

References

[1] A. Ghosh, R. Ratasuk, B. Mondal, N. Mangalvedhe, and T. Thomas, “LTE-advanced: Next-generation wireless broadband technology,” IEEE Wirel. Commun., 2010.

[2] H. J. Visser and R. J. M. Vullers, “RF energy harvesting and transport for wireless sensor network applications: Principles and requirements,” Proceedings of the IEEE. 2013.

[3] I. Mujahidin, S. H. Pramono, and A. Muslim, “5.5 Ghz Directional Antenna with 90 Degree Phase Difference Output,” 2018.

[4] D. M. Pozar, “New results for minimum Q, maximum gain, and polarization properties of electrically small arbitrary antennas,” in 3rd European Conference on Antennas and Propagation, 2009.

[5] R. Yuwono and I. Mujahidin, “Rectifier using UWB microstrip antenna as electromagnetic energy harvester for GSM, CCTV and Wi-Fi transmitter,” J. Commun., 2019.

[6] D. J. Norris, Beginning Artificial Intelligence with the Raspberry Pi. 2017.

[7] I. Mujahidin, “Directional 1900 Mhz Square Patch Ring Slot Microstrip Antenna For WCDMA,” JEEMECS (Journal Electr. Eng. Mechatron. Comput. Sci., 2019.

[8] Y. J. Cheng, W. Hong, K. Wu, and Y. Fan, “A hybrid guided-wave structure of half mode substrate integrated waveguide and conductor-backed slotline and its application in directional couplers,” IEEE Microw. Wirel. Components Lett., 2011.

[9] R. Yuwono, I. Mujahidin, A. Mustofa, and Aisah, “Rectifier using UFO microstrip antenna as electromagnetic energy harvester,” Adv. Sci. Lett., 2015.

[10] C. E. Balanis, “Antenna Theory: Analysis and Design, 3rd Edition - Constantine A. Balanis,” Book. 2005.

[11] I. Mujahidin and B. F. Hidayatulail, “2.4 GHz Square Ring Patch With Ring Slot Antenna For Self Injection Locked Radar,” JEEMECS (Journal Electr. Eng. Mechatron. Comput. Sci., vol. 2, no. 2, 2019.

[12] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, “Deep learning in spiking neural networks,” Neural Networks. 2019.

[13] D. A. Prasetya, A. Sanusi, G. Chandrarin, E. Roikhah, I. Mujahidin, and R. Arifuddin, “Small and Medium Enterprises Problem and Potential Solutions for Waste Management,” J. Southwest Jiaotong Univ., vol. 54, no. 6, 2019.

[14] R. Shokri and V. Shmatikov, “Privacy-preserving deep learning,” in 2015 53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015, 2016.

[15] D. A. Prasetya, A. Sanusi, G. Chandrarin, E. Roikhah, I. Mujahidin, and R. Arifuddin, “Community Culture Improvisation Regarding Waste Management Systems and Per Capita Income Increase,” J. Southwest Jiaotong Univ., vol. 54, no. 6, 2019.

[16] C. Cortes, X. Gonzalvo, V. Kuznetsov, M. Mohri, and S. Yang, “AdaNet: Adaptive structural learning of artificial neural networks,” in 34th International Conference on Machine Learning, ICML 2017, 2017.

[17] M. Kalash, M. Rochan, N. Mohammed, N. D. B. Bruce, Y. Wang, and F. Iqbal, “Malware Classification with Deep Convolutional Neural Networks,” in 2018 9th IFIP International Conference on New Technologies, Mobility and Security, NTMS 2018 - Proceedings, 2018.

[18] Y. Bengio, “Learning deep architectures for AI,” Found. Trends Mach. Learn., 2009.

[19] D. L. Deng, X. Li, and S. Das Sarma, “Quantum entanglement in neural network states,” Physical Review X. 2017.

[20] M. M. Bronstein, J. Bruna, Y. Lecun, A. Szlam, and P. Vandergheynst,

"Geometric Deep Learning: Going beyond Euclidean data," IEEE Signal Processing Magazine. 2017.

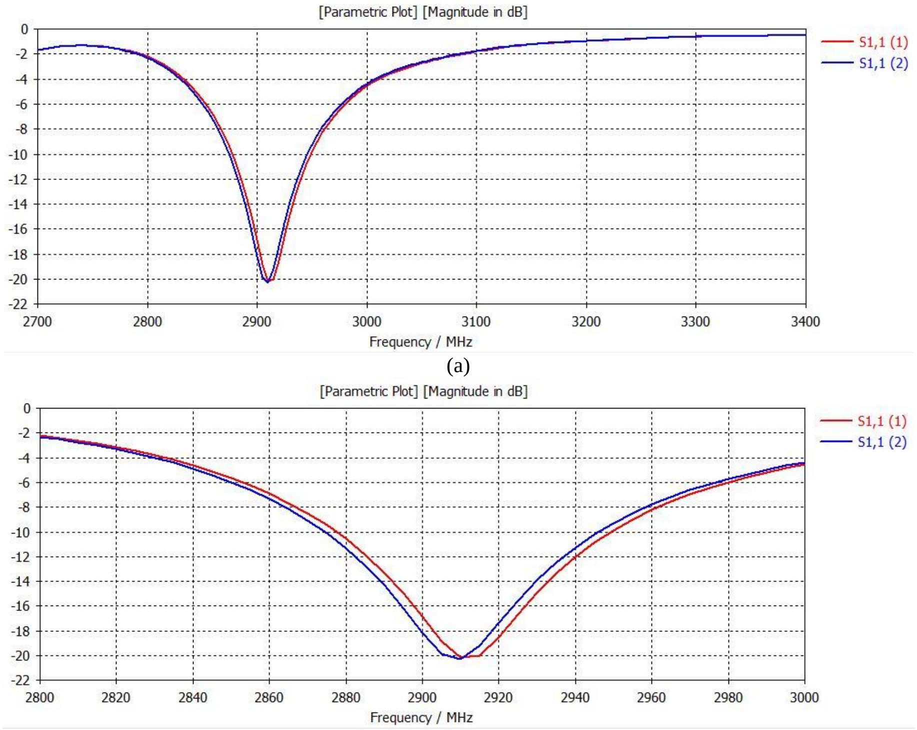
[21] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," Neurocomputing, 2016.

[22] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," npj Digit. Med., 2018.

[23] I. Mujahidin, D. A. Prasetya, A. B. Setywan, and P. S. Arinda, "Circular Polarization 5.5 GHz Double Square Margin Antenna in the Metal Framed Smartphone for SIL Wireless Sensor," in 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA), 2019, pp. 1–6.

APPENDIX

The appendix contains sample results from the results of experiments that have been carried out. The sample provides an overview of the accuracy of the prediction algorithm used. Thus algorithmic implementation can be implemented more efficiently in the various antenna sizes that are proposed. From some of the samples displayed, all value is the good agreements that based on existing minimum standard parameters.



(a)

(b)

Fig. 4. Comparison of prediction and truth for 2913 MHz input

TABLE II. PREDICTION RESULT AND TRUTH OF THE TESTING DATA

No	Ground Plane Height		Triangle Side Length		Circle Diameters		S-Parameter		Gain		Directivity	
	T	P	T	P	T	P	T	P	T	P	T	P
1	40	40	20	20	16	15.981	-20.820	-18.981	4.8	4.678	4.8	5.066
2	40	40	20	20	16.1	16.077	-20.510	-19.120	4.8	4.688	4.84	5.085
3	40	40	20	20	16.2	16.212	-20.270	-19.314	4.32	4.701	4.93	5.113
4	40	40	20	20	16.3	16.270	-20.100	-19.397	4.47	4.706	4.98	5.125
5	40	40	20	20	16.4	16.385	-20.270	-19.564	4.47	4.718	5.04	5.149
6	40.01	40.01	20	20	17.3	17.333	-20.230	-20.076	4.4	4.821	5.42	5.364
7	40.01	40.01	20	20	17.4	17.431	-20	-21.215	4.83	4.830	5.41	5.384
8	40.01	40.009	20	20	17.5	17.556	-22.130	-21.395	4.87	4.843	5.46	5.410
9	40.01	40.009	20	20	17.6	17.662	-21.950	-21.548	4.96	4.853	5.5	5.432
10	40.01	40.009	20	20	17.7	17.710	-21.580	-21.617	4.96	4.858	5.5	5.442
11	40.02	40.009	20	20	17.8	17.855	-21.227	-21.825	4.96	4.873	5.54	5.471
10	40.02	40.009	20	20	17.9	17.942	-21.450	-21.950	4.92	4.880	5.62	5.489
11	40.02	40.009	20	20	18	18.040	-22.090	-22.103	4.89	4.891	5.61	5.511
Sum	0.306		0		1.254		4.52		0.69		0.391	
MAE	**0.027**		**0**		**0.114**		**0.410**		**0.062**		**0.035**	

A Review of Vision and Challenges of 6G Technology

Faiza Nawaz[1], Jawwad Ibrahim[2], Maida Junaid[4], Sabila Kousar[5], Tamseela Parveen[6]
Department of computer Science & IT
University of Lahore Gujrat Campus
Gujrat, Pakistan

Muhammad Awais ali[3]
Department of computer Science
NUCES, FAST Islamabad, Pakistan

***Abstract*—With the accelerated evolution of smart terminals and rising fresh applications, wireless information traffic has sharply enhanced and underway cellular networks (even 5G) can't entirely compete the rapidly emerging technical necessities. A fresh framework of wireless communication, the sixth era (6G) framework, by floating aid of artificial intelligence is anticipated to be equipped somewhere in the range of 2027 and 2030. This paper presents critical analysis of Vision of 6G wireless communication and its network structure; also outline a number of important technical challenges, additionally some possible solutions related to 6G, as well as physical layer transmission procedures, network designs, security methods.**

Keywords—Wireless communication; visions; 6G; cellular network; generations; digital technology; satellite networks; cell less architecture

I. Introduction

In spite of the fact that 5G is still in the underlying phase of commercial scale, i.e., related technical features need to be further enhanced and the business model of Internet of Things and vertical industry application situations should be additionally investigated, it is also mandatory for us to synchronously look forward to the communication needs of the future information society and start the idea and technology research for the next generation mobile communication system [1]. Here we try to analyze the necessity of the immediate start-up of the concept and technology research on the next generation mobile communication system referred to as 6G.

A 6G mobile network system requires to deliver extremely fast speed, increase capacity then non-proximity in order to support the likelihood of fresh applications, as vigorous medicine, computer disaster forecasting plus virtual reality (VR). In light of the previous pre-regulation on mobile networks, the first 6G links will be based mostly on the current 5G structure, benefiting from the advantages obtained in 5G (for example, the increase in allowed frequency bands also optimized the design of a decentralized system) and change the means we chore and play [2]. About 2030, our audiences are probable to be affected by data, allowing immediate and limitless wireless connectivity. As an outcome, 6G should promote wireless technologies we know nowadays and attain system prosecution. As an idea aimed at the upcoming, in terms of speed, 6G is likely to use an advanced frequency spectrum than preceding generations to advance the data throughput estimated at 100 to 1000 times quicker than 5G [3]. To be precise, 6G systems will let hundreds of GBs per second to connect to the second using broadband spectrum; for instance, the combined usage of a band from 1 to 3 GHz, a millimeter wave band (mm wave) (30 to 300 GHz) and a terahertz band (0.06 to 10 THz).

In 1926, the visionary Nikola Tesla announced: "When the wireless connection is fully applied, the Earth will be full of incredible brains ..." In 2030, inspired by the basic needs on a personal and social level, and depends on is the expected progress. Information and Communication Technology (ICT), Tesla predictions can come from all over the world and 6G will perform a specific part in this progress by giving an ICT framework that will allow end users to be enclosed by a "huge artificial brain." It offers virtual storage services, unlimited storage and mass cognition capabilities [4].This document presents a 6G vision and also analyzes the possible challenges associated with 6G.

II. 1G to 5G

The cellular wireless Generation (G) usually mention to an alteration in the complexion of framework, speed, technology and frequency. Individually generation devours roughly standards, capacities, techniques, also novel characteristics which separate it as of the past one.

A. First generation (1G analog technology)

The first generation mobile (1980-1990): It assisted data rates beginning 1 KBps to 2.8 KBps and utilize a circuit switch. It used an output technology called Analog Phone Service. It used a bandwidth of 40 MHz and a frequency range of 800 to 900 MHz, Only the sound will support. It used Frequency Division Multiplexing. It delivered little quality calls. The energy consumption was high. It distressed from some disadvantages, as deprived sound connections, poor data capacity, lack of security and untrustworthy transfer [5].

B. Second generation (2G digital technology)

It depends on the GSM or, in other words, on the global mobile communications system. It was promoted in Finland in 1991. These were the first digital cellular networks, with some evidence of the output networks they replaced: better standard, better safety. 2G technologies have been replaced by digital technologies for digital communication by providing services for example text messaging, photo messaging and MMS. Entirely text messages are digitally encoded in 2G technology.

This digital encryption permits you to exchange your data so that the intended recipient does not understand them and does not understand them. There are 3 dissimilar kinds "FDMA, TDMA / GSM and CDMA" of 2G mobile techniques offered with diverse operational techniques, characteristics and terms [6].

C. *Third generation (3G)*

The third generation of mobile transmission systems offers 144kbps great speeds and more for high speed data. It conforms to improvements in older wireless technologies, such as "high speed transmission, high multimedia access and global roaming". 3G is commonly utilized for mobile phones and headphones as a way to link the telephone to the Internet or other IP networks to provide voice and video calls, download and data, plus surf the web. 3G will help multimedia applications such as complete video movement, videoconferencing as well as Internet access. Data is directed via technology named packet switch. Telephone calls decrypted by circuit switch. It is a very modern process of communication that has evolved over the past era [7].

D. *Fourth generation (4G)*

4G mobile communication framework was presented in the late 2000s and was an IP organize framework. The primary objective of 4G innovations is to give great quality, great capacity, and minimal effort security administrations for voice and information, sound and Internet services through IP administrations. The aim of modifying all IP addresses is to have a shared platform for all the innovations advanced to date. It has a capacity of 100 Mbps and 1 Gbps. To utilize the 4G mobile network, multimodal user terminals must be clever to select the wireless destination system. To deliver wireless service anytime, anyplace, terminal portability is an important influence in 4G. Terminal mobility suggests automatic roaming among dissimilar wireless networks. 4G technology coordinates a number of present and future wireless techniques such as "OFDM, MC-CDMA, LAS-CDMA and Network-LMDS" to deliver liberty of drive and continuous roaming from one technology to a different. LTE "long-term evolution" and Wi-MAX "wireless interoperability for microwave access" are pondered 4G technologies. The initial triumphant fourth generation field test was coordinated in Japan in 2005 [8].

E. *Fifth generation (5G)*

In 5G, research focuses on the progress of a "World Wide wireless Web (WWWW), dynamic ad-hoc wireless networks (DAWN) and real wireless communication". The utmost significant techniques for 5G technologies are "802.11 wireless networks in local areas (WLAN) and wireless networks in an urban area (WMAN), an ad hoc wireless personal area network (WPAN) and networks wireless for digital communications". 5G feature provides AI capabilities to portable devices [9].

III. 6G Visions

Since 5G has not yet been launched economically and there are still no excellent applications, it seems very convincing to discuss the 6G prerequisites [10]. The necessities for 6G will track the next technological models, as "smart cars and smart assembly". On transportation as well as manufacturing, a large network of smart means of transportation and robots will need a mobile broadband network as well as an "ultra-high rate wireless with excellent reliability and ultralow latency, giving new kinds of mobile-as-a-service and mobile-as-manufacturing applications".

There are some visions described in this paper according to different prospective and requirements of 6G wireless technology.

A. *A framework of 6G founded on the space resource use, frequency, time*

Fig. 1 describes 6G will use an advanced frequency spectrum than preceding generations to advance speed of data in the frequency dimension. On one needle, great frequency bands, as "mm wave band, terahertz band and visible light frequency band" will be utilized for a transmission of 100 GB / s + in 6G. On the next needle, in the forthcoming, "mobile networks with satellite systems and the Internet may be integrated to build integrated networks", which will gain the frequency, sorts on behalf of services from a point of view of individual mobile communication. In the spatial dimension, to benefit more from multi-channel, the number of antenna devices prepared in equally the transmitter then the receiver will be amplified. MIMO residue techniques (PM-MIMO), as the ultra-huge MIMO (UM-MIMO) for terahertz communication, can help hundreds or thousands of antennas to transmit / receive. In the time dimension, 6G will provide promising change to weak and structural. In addition, the 6G time slot unit can be coherently trodden to further proficiently utilize the higher frequency bands and respond to subtle services. As particle time improves, the flexibility also versatility of the systems will be better also thus ease their compatibility with 2G to 5G [11].

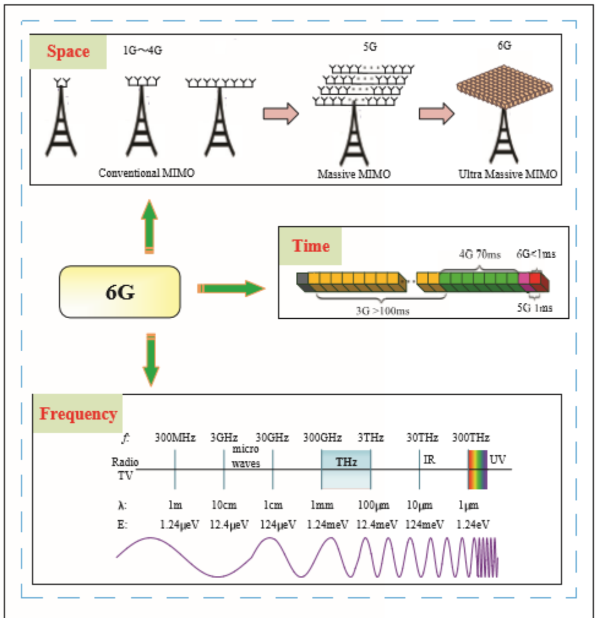


Fig. 1. A framework of 6G founded on the space resource use, frequency, time [11]

B. 6G with Satellite Network

The 6G mobile system for global coverage will be coordinated 5G mobile wireless system plus the satellite network. Such satellite networks include a "satellite communications network, an Earth image satellite network and navigation satellite network". The telecommunications satellite is utilized to transmit voice, data, the Internet and video; the data from the earth imaging satellite network is used to collect climate and environmental data and the satellite navigation network to the global positioning system (GPS). The four republics that built such satellite systems are GPS by the USA. The Compass framework built by China, the "Galileo system" by EU plus the Russian Glonass system [12]. The core objective of 6G is to deliver mobile phone users with a variety of services as a network identifier in a variety of locations, multimedia applications and accessibility to the Internet association for mobile users with a speed of high data without interrupting the network. Below Fig. 2 shows the satellite network with 6G.

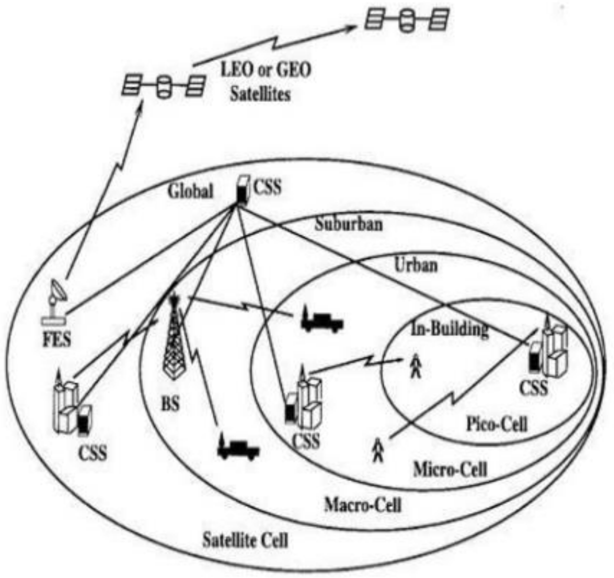


Fig. 2. 6G with Satellite Network [13]

It will provide extremely fast access to Internet services with fantastic data rates of up to 10-11 Gbps. It offers a complete wireless network without borders. Inadequate connectivity between network clients to provide incredible transmission speeds in the Terabit range. Maximize data and IOPS (input and output per second operation) [13].

C. 6G connectivity vision

According to Fig. 3 6G vision can be summarized into four main words “Intelligent Connectivity”, “Deep Connectivity”, “Holographic Connectivity” also “Ubiquitous Connectivity”. These four main words organized establish the 6G general vision of “Wherever you think, everything follows your heart”.

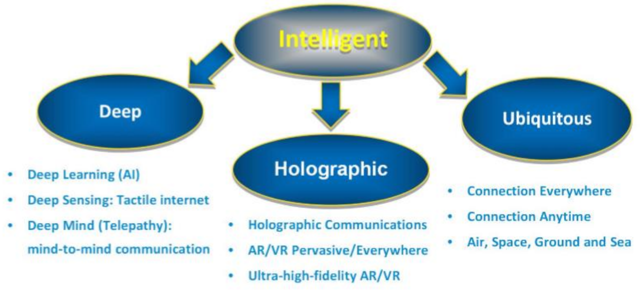

Intelligent
Deep
Holographic
Ubiquitous
• Deep Learning (AI)
• Deep Sensing: Tactile internet
• Deep Mind (Telepathy): mind-to-mind communication
• Holographic Communications
• AR/VR Pervasive/Everywhere
• Ultra-high-fidelity AR/VR
• Connection Everywhere
• Connection Anytime
• Air, Space, Ground and Sea

Fig. 3. 6G vision [15]

We believe that building 6G network based on AI technology will be an inevitable choice, and "Intelligent" will be the inherent feature of 6G network, namely the so-called "**Intelligent connectivity**". 6G networks will face many challenges: more complex and huge networks, more types of terminals and network devices, and more complex and diverse business types. "Intelligent Connectivity" will meet two requirements at the same time: on the one hand, all the related connected devices in the network itself are intelligent, and the related services have been intelligent; on the other hand, the complex and huge network itself needs intelligent management. "Intelligent Connectivity" will be the basic characteristics supporting the other three major features of 6G network: Deep Connectivity, Holographic Connectivity and Ubiquitous Connectivity. We expect that in the next 10 years (2030 ~) of 6G systems, access requirements will evolve from deep coverage to "**Deep connectivity**". Its characteristics can be summarized as follows: Deep Sensing: Tactile Internet, Deep Learning/AI: Deep Data Mining, DeepMind: Telepathy, Mind-to-Mind Communication. It can be expected that in ten years (2030 ~), the media communication will be mostly planar multimedia, higher fidelity AR/VR interaction, even holographic information interaction, also wireless holographic communication will become a genuineness. High fidelity AR/VR will be ubiquitous, as well as holographic communication and display can also be carried out at anytime and anywhere, so that people can enjoy fully immersed holographic interactive experience at any time and place, that is, to realize the communication vision of so-called "**holographic connectivity**" [14]. "Anytime, anywhere" connection requirement ten years later (2030 ~), that is, to achieve real **"Ubiquitous connectivity”,** a vast" world will become more and more accessible. Future 6G visions, "Intelligent Connectivity" is the brain then nerve of the 6G network, while the other three characteristics of "Deep Connectivity", "Holographic Connectivity" and "Ubiquitous Connectivity" establish the trunk of the 6Gnetwork. These four characteristics together make the future 6G network a complete organic whole with "soul". In the future, the communication system will be further developed and enhanced on the basis of the existing 5G. The information will break through the curb of time and space, the network will close the distance between all things, the seamless integration of human and all things will be realized [15].

D. Cell less Architecture for 6G Networks

Following generation wireless networks have to assist an enormous figure of terminal users within minor geographical extents, and this will stretch increase to dense or ultra-dense placement of APs/BSs with overlapping coverage zones. In such a case, a different AP / BS will be served simultaneously on the devices (for example, by multiple transmissions and multi-client affiliations), which will be necessary for well-organized handover, frequency distribution and interference management. When an extremely quick backup between certain AP and BS is used, the general network, from the point of view of the end device, will be presented as a major distribution system, without cells, with multiple inputs and with multiple output (MIMO). In particular, all the APs will know all the devices active in their region. Access points can be thought of as remote radio headers (RRHs), as is the situation with cloud radio access networks (CRAN) [16]. More than one AD can accommodate each device, thanks to the coordination of the transmission or via a transmission multiplex. It can be convenient to see this cell less architecture as a general form of the famous Comp transmission, where collaborative access points come together to respond to all the gadgets in their inclusion regions (cellular devices and cell replacement). This can be improved via the usage of much quick centralized processing units which allocate resources to diverse terminal devices, and the CRANs can target the processing of data to what is called the group of baseband units. Complete coordination between numbers of DAs can lead to interference management ideally, or almost ideally, through centralized or distributed improvement techniques.

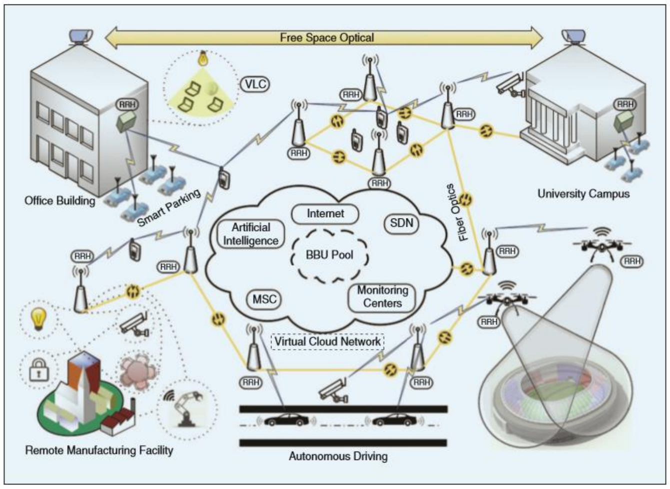


Fig. 4. The Cell-Less 6g Network Architecture [17].

Fig. 4 shows for network architecture with significant access necessities, novel spectrum management also multiple access strategies will be requisite. The selection of advanced frequency bands such as "millimeter waves and beyond" will assess problematic spectrum shortages; although these bands are not perfect, particularly in medium and huge communication areas, owing to the relaxation and very high preconditions of the beam direction. In the case of multiple access, the selection of full-time OMA schemes in the obtainable spectrum is not enough. On the second hand, untainted NOMA techniques will not contain the flexibility to facilitate wireless connectivity aimed at gadgets by diverse service needs. Therefore, new access and resource allocation and multiple access management techniques will be needed to interfere with these cell-free networks, provided the restricted spectrum resources [17].

E. 6G communication architecture scenario

Some of the main inspiring developments at back the development of 6G transmission framework are as shadows, "high bit rate, high reliability, low latency, high energy efficiency, high spectral efficiency, new spectrums, green communication, intelligent networks, network availability and convergence of communications, localization, computing, control and sensing" 6G will be a completely computerized, linked globe.

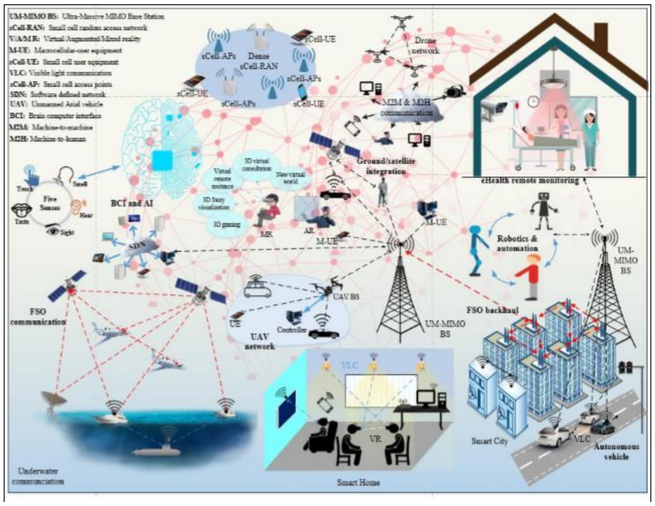

UM-MIMO BS: Ultra-Massive MIMO Base Station
sCell-RAN: Small cell random access network
V/A/MR: Virtual/Augmented/Mixed reality
M-UE: Macrocellular-user equipment
sCell-UE: Small cell user equipment
VLC: Visible light communication
sCell-AP: Small cell access points
SDN: Software defined network
UAV: Unmanned Arial vehicle
BCI: Brain computer interface
M2M: Machine-to-machine
M2H: Machine-to-human
sCell-APs
Dense sCell-RAN
sCell-UE
Drone network
M2M & M2H communication
Ground/satellite integration
eHealth remote monitoring
Touch
Smell
Five Senses
Taste
Hear
Sight
BCI and AI
SDN
3D virtual consultation
Virtual remote assistance
New virtual world
3D Story visualization
3D gaming
MR
AR
M-UE
Robotics & automation
UM-MIMO BS
FSO communication
UAV BS
UE
Controller
UAV network
FSO backhaul
VLC
VR
Smart City
Autonomous vehicle
Underwater communciation
Smart Home

Fig. 5. Possible 6G communication architecture scenario [19]

The Fig. 5 demonstrate the communication architecture setup to imagining the 6G communication systems. Approximately important predictions as well as applications of 6G wireless communication are fleetingly defined beneath.

Super smart society: The specific structures of 6G will quicken the structure of smart societies prompting "life class developments, environmental observing and robotics using AI-based M2M communication and energy harvesting" [20]. **Extended reality:** Augmented reality services (hereinafter referred to as XR), "counting augmented reality (AR), mixed reality (MR) and VR", are essential components of 6G communication systems [19]. **Connected robotics and autonomous systems:** 6G systems help deploy linked robots and autonomous systems. The automatic means of transportation founded on 6G wireless communication can significantly modify our everyday lives. The 6G network will indorse the actual use of cars without a driver [18]. **Wireless brain computer interactions:** It is a means of straight communication among the brain and outside devices. The BCI receives signals from the brain that they are moving to a digital device and analyzes and interprets the signs in additional orders or actions. 6G wireless communications elements will facilitate the actual application of BCI networks to live a smart life. **Haptic Communication:** The sense of touch is use by non-verbal communication. The proposed 6G wireless communication will support random communication. **Smart healthcare:** 6G systems will ease a consistent remote monitoring system in the healthcare system. Even remote surgery will be possible thanks to 6G communication. Lager data speed, less failure, and a very consistent 6G network will aid transport large amounts of medical data quickly and reliably, which can advance access to upkeep and eminence of care. **Automation and industrial:** The term automation references to "automatic control of processes, devices and systems ". 6G automated systems will offer "highly reliable, scalable, and secure communications using high-speed, low-intelligence networks ".

Information transfer in the five senses: This technique applies from the neurological procedure to sensory integration. It regains the feelings of the human physique as well as the environment and utilizes the body efficiently in the surroundings and in native conditions. BCI technique will efficiently improve this application. **Internet of everything:** The 6G system will support the complete IoE system. It is essentially an Internet of Things (IoT), then it is a general word that assimilates four characteristics, as "data, people, processes also physical devices", into a framework [19].

IV. Comparison between 5G and 6G

A valued comparison of 5G and 6G communications is shortened in Fig. 6 [21]. We first assume that the electrical competence of 5G was previously close to the border with progress in a huge MIMO, network compaction and millimeter wave transmission, for example, similarly a series of legacy multiplex methods acquired from 4G. As the limits of Shannon are limited, it is unlikely that the spectral efficiency in 6G will improve on a large scale. On the contrary, 6G communications should significantly improve security, privacy and confidentiality with novel techniques. In 5G networks, customary encryption algorithms founded on the main "Rivest-Shamir-Adleman (RSA)" public crypto-sms are still used to ensure the security and confidentiality of transmissions. RSA crypto-spores are uncertain about the pressure of Dig Data and artificial intelligence technologies, much less than the privacy mechanisms that weren't developed in the 5G era.

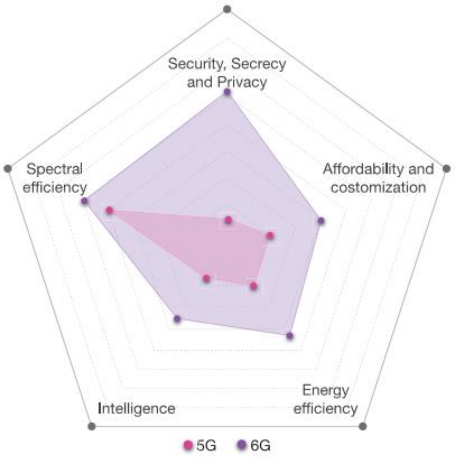


Fig. 6. Comparison among 5G and 6G communications [21]

TABLE I. A Comparisons of 5G and 6G Kpis [22]

KPI	5G	6G
Traffic size	10 Mb/s/m^2	~ 1–10 Gb/s/m^3
Downlink data rate	20 Gb/s	1 Tb/s
Uplink data rate	10 Gb/s	1 Tb/s
Uniform user experience	50 Mb/s, 2D everywhere	10 Gb/s, 3D everywhere
Latency (radio interface)	1 ms	0.1 ms
Jitter	NS	1 µs
Reliability (frame error rate)	$1{-}10^{-5}$	$1{-}10^{-9}$
Energy/bit	NS	1 pJ/b
Localization accuracy	10 cm in 2D	1 cm in 3D
NS: not specified		

The KPIs summarized in Table I [22], which highlights essential improvements through admiration to 5G KPIs. More or less KPIs, as delay jitter and energy per bit, are not definite in 5G, as they do not truly signify an emphasis of 5G, whereas they are important KPIs for 6G.

V. 6G Issues and Solutions

A. Limits on Flexible Radio Access

Cell size plus carrier frequency may limit the OFDM numerology option [23]. On one arrow, utilize of numerology with the extensive subcarrier spacing is usually further appropriate for minor cell sizes owing to its smaller delay extensions than large cell increases. On the other hand, large quantities of digital cells can be used with a larger space subcarrier, but at a lower performance cost. It is similarly significant to remember that the size of cells with high frequency carriers is restricted because of problems of route propagation and Doppler propagation in case of high mobility [24].

B. Network security issue

Security is a serious concern for 6G wireless networks, specifically when using the Terrestrial Space Integrated Network (STIN) technique. In 6G, in addition to traditional physical series safety, other forms of privacy, as cohesive network security, must be measured together. A novel approach to security, which depends on little difficulty and a high level of security, must therefore be intensified. To this conclusion, certain physical layer security techniques intended for 5G can be protracted to 6G systems, for example, a secure MIMO mass based on low density parity control (LDPC); Mm-Wave Safe techniques can also be used for "UM-MIMO and THz band applications". When it comes to integrated network security, it is very important that there is an appropriate management purpose for diverse function means for diverse security domains. A central distribution management mechanism is a promising mechanism for STIN which takes into account the management of multicultural and certificate less communication keys. With the efficient administration and application, these physical and network layer security methods can combine this integrated security solution, which effectively protects confidential information and confidentiality on 6G networks [11].

C. Resource as a Service (RaaS)

The advent of software networking (SDN) and network functionality verification (NFV) eases the evolution towards an integrated resource-oriented resource allocation called RaaS. The result is a perception of network splitting to generate virtual networks across the physical infrastructure. It permits mobile operators or service suppliers to assign virtual network resources to encounter precise service needs. Programmable metro conditions and software specificities will probably be part of the network's resources. Therefore, only NFV development trends during the 6G cycle will contain the network screening with programmable software and surfaces defined by the metro, from a machine-activated cloud access network (C-RAN) to free [25].

D. Heterogeneous High Frequency Bands

The use of mm-Wave and THz in 6G presents a number of new open problems. For mm-Wave, support for high movements at mm-Wave frequencies will be an open central problem. In the case of THz, new models of architecture and propagation are necessary [26]. The great power, great sensitivity and low noise of the transmitter required to overcome the THz loss on the high path are key features. Once these elements of the physical series are well understood, network layer protocols and new connections must be developed to enhance the utilization of cross-frequency resources, captivating into account the extremely variable also inexact nature of mm-wave and THz environments. Alternative significant way is the learning of the coexistence of THz cells, millimeter waves and microwaves in each series [27].

E. Tactile Communications

Next using holographic communication to translate virtual views near to the reality of people, actions, environments, etc. It is advantageous to remotely exchange a physical communication via a real-time Internet connection [28]. The expected services include telecommunications, the automated collaborative reader and interpersonal communication, which should allow the application of random control across communication networks. The efficient design of the communication system between the rows must be carried out to meet these strict requirements. For instance, new physical layer diagrams (PHY) must be established, for the design of signaling systems, the congestion of waveforms, etc. to improve transfer and motivated protocol. Wireless communication systems cannot meet these requirements and wireless fiber communication systems are required [29].

VI. Conclusion

In this document, we have discussed current then upcoming generations of wireless communications. We present a general description of the technologies that will characterize 6G networks. 6G networks will embrace new spectrum bands, combining advances across the network, from the circuit and antenna design to network architecture, protocols and artificial intelligence. Finally, we highlight a number of issues in 6G communication systems, which we hope will inform future development. We noticed that 6G networks considered by their flexibility also versatility and that the sketch of 6G networks is a very ordered scientific arena.

References

[1] Zhao, Y., Yu, G., & Xu, H. (2019). 6G Mobile Communication Network: Vision, Challenges and Key Technologies. *arXiv preprint arXiv:1905.04983*.

[2] David, K., & Berndt, H. (2018). 6G vision and requirements: Is there any need for beyond 5G? *ieee vehicular technology magazine*, *13*(3), 72-80.

[3] Cacciapuoti, A. S., Sankhe, K., Caleffi, M., & Chowdhury, K. R. (2018). Beyond 5G: THz-based medium access protocol for mobile heterogeneous networks. *IEEE Communications Magazine*, *56*(6), 110-115.

[4] Calvanese Strinati, E., Mueck, M., Clemente, A., Kim, J., Noh, G., Chung, H., ... & Destino, G. (2018). 5GCHAMPION–Disruptive 5G Technologies for Roll-Out in 2018. *ETRI Journal*, *40*(1), 10-25.

[5] Goyal, P., & Sahoo, A. K. A Roadmap towards Connected Living: 5G Mobile Technology.

[6] sah, d. h. n. (2017). A brief history of mobile generations and satellite wireless communication system.

[7] T. Arunkumar and L. Kalaiselvi, “Latest Technology of Mobile Communication and Future Scope of 7.5 G”, International Journal of Engineering & Technology Research, Volume 2, Issue 4, pp. 23-31, July 2014

[8] Gawas, A. U. (2015). An overview on evolution of mobile wireless communication networks: 1G-6G. *International Journal on Recent and Innovation Trends in Computing and Communication*, *3*(5), 3130-3133.

[9] Gill, J., & Singh, S. (2015). Future prospects of wireless generations in mobile communication. *Asian J. Comp. Sci. Technol*, *4*(2), 18-22.

[10] Zong, B., Fan, C., Wang, X., Duan, X., Wang, B., & Wang, J. (2019). 6G Technologies: Key Drivers, Core Requirements, System Architectures, and Enabling Technologies. *IEEE Vehicular Technology Magazine*, *14*(3), 18-27.

[11] Yang, P., Xiao, Y., Xiao, M., & Li, S. (2019). 6g wireless communications: Vision and potential techniques. *IEEE Network*, *33*(4), 70-75.

[12] Khutey, R., Rana, G., Dewangan, V., Tiwari, A., & Dewamngan, A. (2015). Future of wireless technology 6G & 7G. *International Journal of Electrical and Electronics Research*, *3*(2), 583-585.

[13] Kalbande, D., Haji, S., & Haji, R. (2019, June). 6G-Next Gen Mobile Wireless Communication Approach. In *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1-6). IEEE.

[14] Bastug, E., Bennis, M., Médard, M., & Debbah, M. (2017). Toward interconnected virtual reality: Opportunities, challenges, and enablers. *IEEE Communications Magazine*, *55*(6), 110-117.

[15] Yajun, Z., Guanghui, Y., & Hanqing, X. U. (2019). 6G mobile communication networks: vision, challenges, and key technologies. *SCIENTIA SINICA Informationis*, *49*(8), 963-987.

[16] A. Checko et al., “Cloud RAN for mobile networks—A technology overview,” IEEE Commun. Surveys Tut., vol. 17, no. 1, pp. 405–426, 2015. doi: 10.1109/COMST.2014.2355255.

[17] Al-Eryani, Y., & Hossain, E. (2019). The D-OMA Method for Massive Multiple Access in 6G: Performance, Security, and Challenges. *IEEE Vehicular Technology Magazine*, *14*(3), 92-99.

[18] B. Li, Z. Fei, and Y. Zhang, “UAV communications for 5G and beyond: recent advances and future trends,” IEEE Internet of Things Journal, vol. 6, no. 2, pp. 2241-2263, April 2019.

[19] Chowdhury, M. Z., Shahjalal, M., Ahmed, S., & Jang, Y. M. (2019). 6G Wireless Communication Systems: Applications, Requirements, Technologies, Challenges, and Research Directions. *arXiv preprint arXiv:1909.11315*.

[20] (2019). 6G. [Online]. Available: http://mmwave.dei.unipd.it/research/6g/

[21] Dang, S., Amin, O., Shihada, B., & Alouini, M. S. (2019). From a Human-Centric Perspective: What Might 6G Be?. *arXiv preprint arXiv:1906.00741*.

[22] Strinati, E. C., Barbarossa, S., Gonzalez-Jimenez, J. L., Ktenas, D., Cassiau, N., Maret, L., & Dehos, C. (2019). 6G: the next frontier: from holographic messaging to artificial intelligence using subterahertz and visible light communication. *IEEE Vehicular Technology Magazine*, *14*(3), 42-50.

[23] Zaidi, A. A., Baldemair, R., Moles-Cases, V., He, N., Werner, K., & Cedergren, A. (2018). OFDM numerology design for 5G new radio to support IoT, eMBB, and MBSFN. *IEEE Communications Standards Magazine*, *2*(2), 78-83.

[24] Lee, Y. L., Qin, D., Wang, L. C., & Hong, G. (2019). 6G Massive Radio Access Networks: Key Issues, Technologies, and Future Challenges. *arXiv preprint arXiv:1910.10416*.

[25] Tariq, F., Khandaker, M., Wong, K. K., Imran, M., Bennis, M., & Debbah, M. (2019). A speculative study on 6G. *arXiv preprint arXiv:1902.06700*.

[26] Xing, Y., & Rappaport, T. S. (2018, December). Propagation measurement system and approach at 140 GHz-moving to 6G and above 100 GHz. In *2018 IEEE Global Communications Conference (GLOBECOM)* (pp. 1-6). IEEE.

[27] Saad, W., Bennis, M., & Chen, M. (2019). A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. *arXiv preprint arXiv:1902.10265*.

[28] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, “5G-enabled tactile internet,” IEEE Journal on Selected Areas in Communications, vol. 34, no. 3, pp. 460–473, Mar. 2016.

[29] Kim, K. S., Kim, D. K., Chae, C. B., Choi, S., Ko, Y. C., Kim, J., & Lee, K. (2018). Ultrareliable and low-latency communication techniques for tactile internet services. *Proceedings of the IEEE*, *107*(2), 376-393.

Towards a Dynamic Scalable IoT Computing Platform Architecture

Desoky Abdelqawy[1], Amr Kamel[2], Soha Makady[3]
Faculty of Computers and Artificial Intelligence
Cairo University, Egypt

Abstract—**Internet of Things (IoT) has become an interesting topic among technology titans and different business groups. IoT platforms have been introduced to support the development of IoT applications and services. Such platforms connect the real and virtual worlds of objects, systems and people. Even though IoT platforms increasingly target various domains, they still suffer from various limitations. (1) Integrating hardware devices from different providers/vendors (thereafter referenced as heterogeneous hardware) is still a subtle task. (2) Providing a scalable solution without altering the end user privacy (e.g., through the use of cloud platforms) is hard to achieve. (3) Handling IoT Applications reliability as well as platform reliability is still not fully supported. (4) Addressing Safety-critical applications needs are still not covered by such platforms. A novel scalable dynamic computing platform architecture is proposed to address such limitations and provide simultaneous support for five non-functional requirements. The supported non-functional requirements are scalability, reliability, privacy, timing for real-time systems and safety. The proposed architecture uses a novel network topology design, virtualization and containerization concepts, along with a service-oriented architecture. We present and use a smart home case study to evaluate how traditional IoT platform architectures are compared to the proposed architecture, in terms of supporting the five non-functional requirements.**

Keywords—Interent of Things (IoT); IoT platforms; IoT architecture; edge computing

I. Introduction

The world is currently changing very fast, jogging to be Smart. Smart Cities [1], Smart Homes [2] and Smart Factories [3] and Smart Grid [4] are bright terms the world is currently looking up to. Internet Of Things [IoT] technology is considered the main player to achieve such aspiration; Gartner [5] reported that by 2020, 95% of new product designs will contain IoT Technology. IoT had been included in the list of six ”Disruptive Civil Technologies” with potential impact on US national power by the US National Intelligence Council [6]. In [7], There will be 50 billion things connected to the internet by 2020 as predicted by Cisco Internet Business Solutions Group.

IoT is defined as a network of devices/things coupled with sensors, actuators, software as well as required electronics to make them able to collect, process and share data. From another perspective an IoT is an architectural framework that permits the integration and/or data exchange between the physical world and computer systems through the underlying network infrastructure. This network is orchestrated with what so called an IoT platform. An IoT platform is the key software component that facilitates the development of scalable IoT applications and services that connect the real and virtual worlds between objects, systems and people. As described in [8], such platforms have to meet the expectation of different players in the IoT ecosystem. i.e (1) Device vendors require a standardized communication protocol for seamless integration and operation (2) Application developers need a simplified development support to focus on application development instead of integration and deployment issues. (3) The providers of platforms and related services seek a clean and simplified way to extend and support their services. (4) The end-users demand security and privacy support.

More than one hundred of such platforms have been created over previous years [9]. Such platforms come in various shapes, and sizes. Yet, there is still a lack of any defined agreement or a standard to manage such technology (e.g., a standard communication protocol, a standard architecture and deployment methodologies, a defined and dedicated market place. [10].

Therefore, various studies have been conducted in [8], [11], [12] and [10] to evaluate IoT platforms landscape, existing IoT architectures, and assess whether such platforms satisfy the IoT ecosystem needs. Such studies concluded that although existing IoT platforms cover a wide-range requirements for IoT platforms, the following four non-functional requirement still remain relatively unexplored: (1) system-wide scalable dynamic resource discovery, (2) reliability (3) Real Time support and (4) privacy.

In this paper, we propose a scalable computing platform architecture, that simultaneously satisfies the above mentioned four non-functional requirements in addition to securing the required support for safety critical IoT Applications.

The remainder of this paper is structured as follows. Section II provides a background on traditional IoT platform architectures. Section III presents a motivational scenario for an extendable temporary virtual key management system as a Smart-Home use-case. In Section IV we present our proposed architecture and apply it to the proposed smart-home use-case in Section V. We compare our proposed architecture against traditional IoT architectures in Section VI. Section VII presents the related work, whereas Section VIII concludes the paper.

II. Background

Simply IoT platform could be defined as the enabler platform addressing IoT Full stack. Such IoT stack includes devices/actions/connectivity management, analytic, developer ecosystem, orchestration and open-external interfaces [9].

A classification for IoT Platforms could be done from platform architecture point of view as defined in [8].

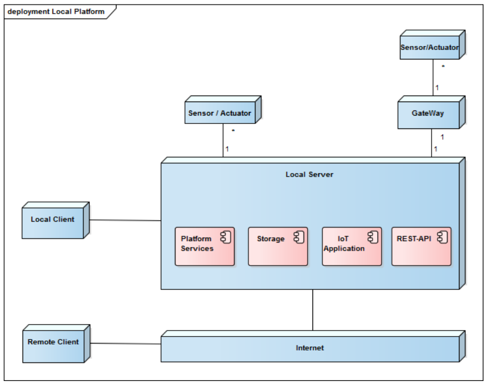


Fig. 1. UML deployment Diagram For Local Based Architecture IoT Platforms

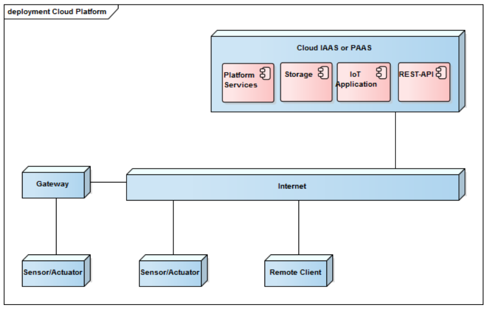


Fig. 2. UML deployment Diagram For Cloud Based Architecture IoT Platforms

- Local Platform
  Figure 1 models that type of platform, where there is a local server (i.e., a centralized node) that acts as a single computing resource. Such node contains four main components: the platform's services, the data storage, the hosted IoT applications, and the REST API to expose the platform services through Internet for remote clients. The local server connects to sensors and actuators through either a wired or a wireless connection. The connection could be done in a direct manner or through a gateway that acts as a bridge for such a group of sensors or actuators to the platform. This centralized node might expose a REST full API to Remote Client through Internet i.e mobile applications or web-app dashboards. Alternatively a local client can access the platform functionalities through a direct connection with the local server.

- Cloud Platform
  Figure 2 models that type of platform, where the sensors and/or actuators are connected directly through the internet, or through a gateway, to the cloud. The Cloud provides the required services, storage and needed computing resource for the applications in either infrastructure-as-a-service (IAAS) or platform-as-a-service (PAAS). Remote Client can access the platform functionalities through an Internet connection.

A summary for a set of non-functional requirements for IoT platforms has been presented in [13] that includes the following: (1) Scalability; an IoT platform shall support expansion with heterogeneous devices and applications diversity inside ultra large network. (2) Reliability, an IoT platform shall have the capabilities to cope with the IoT nodes/devices constrained resources and the dynamic nature of IoT hubs/networks where devices/nodes not always available all the time. (3) Timing for real-time applications, IoT platform shall be able to serve real time applications with timing requirements. (4) Safety, Where IoT platform shall provide the required support to execute safety critical applications i.e. redundancy support, and application migration to recover from hardware failures.

TABLE I. SUMMARY OF PLATFORMS ARCHITECTURE PROPERTIES

Architecture Properties	Local Platforms	Cloud Platforms
Scalability	✗	✓
Reliability	✗	✓
Privacy	✓	✗
Timing for Real-Time	✓	✗
Safety	✗	✓

Table I presents an analysis for whether existing IoT platforms architecture (mainly local and cloud based platforms) support the above mentioned non functional requirements. we further explain as follows:

- Scalability
  Cloud based platforms are scalable by nature where computing capabilities can be scaled up with pay-as-you-go model. On the other hand, local based platforms suffer from fixed computing capabilities which is defined from the beginning thus limiting their scalability. In case of un-reliable connections with the cloud, Local based platforms could be considered as more reliable.

- Reliability
  In case of a reliable connection, cloud based platforms could be considered more reliable due to high availability of resources and infrastructure reliability. On the other hand Local-based platform might not provide the needed support for reliability due it's fixed static resources available from the beginning.

- Privacy
  Local based platforms satisfy privacy through hosting all the data of an IoT application locally. Such local hosting gives the user full control on who could be authorized to access such data. On the contrary, cloud-based platforms cannot satisfy privacy as the data is hosted remotely on a cloud. Such hosting raises issues specially when there is no clear strategy for data ownership.

- Timing for Real-Time
  Local based platforms provide the needed support for Real-Time application since it does not suffer from latency related issues which is part of cloud based platforms by nature.

- Safety
  Safety Critical applications are defined as applications where failure might lead to death or serious

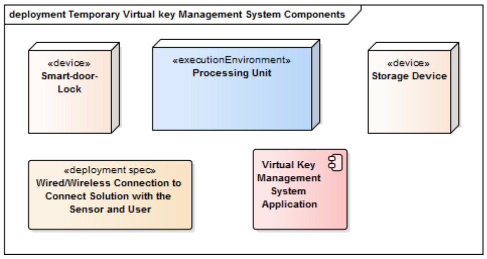

deployment Temporary Virtual key Management System Components
«device»
Smart-door-Lock
«executionEnvironment»
Processing Unit
«device»
Storage Device
«deployment spec»
Wired/Wireless Connection to Connect Solution with the Sensor and User
Virtual Key Management System Application

Fig. 3. Use-case main components

injury to people, severe damage in equipments or environment harm. This type of applications exist heavily in Smart Factories and Smart-Vehicles. Those applications might need a special type of platforms that support redundancy, cross checks, freedom from interference. Cloud based platforms have the available resources to support redundancies for such applications. On the other hand, local-based platforms cannot easily be extended to provide the required redundant copies. That is mainly because of the fixed resources defined from the beginning within local-based platforms.

III. Motivational Scenario

Smart-Home is a home with an automation system that enables electronic, electrical and technology based tasks within a home. A home Automation system might control lighting, climates, entertainment systems. A home automation system could also manage home security such as access control and alarm systems.

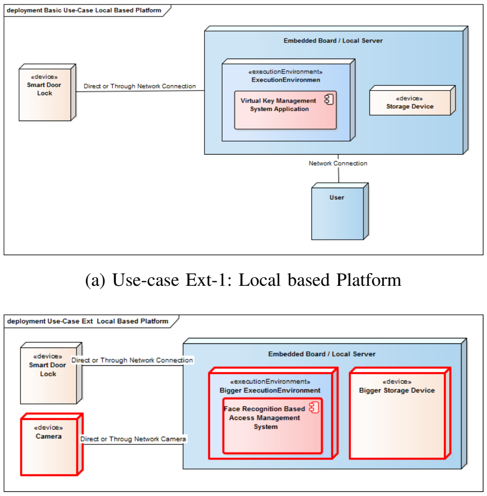
A. Temporary Virtual Key System

An owner of a smart house decides to rent his home in the summer to different tenants every month. Accordingly, different tenants could be using the house with the need to share the physical key of the house across such tenants. Any house tenant could miss closing the door of the house by mistake, hence increasing the chances of robbing such house. Accordingly, a virtual key management system would be a highly needed feature in such a smart house. Such system would allow the house owner to create and share a virtually time bounded house entry key.

Figure 3 shows the main components needed to setup and deploy such system. Those components include: a smart door access lock (a device actuator), a virtual key management software system/application (IoT Application), processing unit (Execution environment) to execute the application and storage device to persist the application.

IoT platform is responsible for managing and organizing such components together to achieve use-case requirements. Local or Cloud based IoT platforms could be utilized to deploy the above mentioned simple concrete use-case.

A local based platform architecture for the smart home virtual key management system presented in Figure 4a has a single node that consists of a storage to store virtual key management software application, processing unit to execute it and wired/wireless connection with the smart door lock actuator device. Such architecture is defined since the initial design of such Figure 5a presents an alternative cloud-based platform architecture for the same smart home virtual key management system where the smart door lock is connected directly to a cloud hosted virtual key management software application. Such application would be stored and executed inside cloud infrastructure and the control signal is pushed back to the smart lock over Internet. Also such architecture is considered a scalable one due to the ease of adding more processing power, it suffers from privacy and reliability issues in-case of un-stable Internet connection.


deployment Basic Use-Case Local Based Platform
Embedded Board / Local Server
«device»
Smart Door Lock
Direct or Through Network Connection
«executionEnvironment»
ExecutionEnvironmen
Virtual Key Management System Application
«device»
Storage Device
Network Connection
User
deployment Use-Case Ext Local Based Platform
«device»
Smart Door Lock
Direct or Through Network Connection
Embedded Board / Local Server
«executionEnvironment»
Bigger ExecutionEnvironment
Face Recognition Based Access Management System
«device»
Bigger Storage Device
«device»
Camera
Direct or Throug Network Camera

(a) Use-case Ext-1: Local based Platform

(b) Case II

Fig. 4. Local Based Platform Use-case(s) Architecture

B. Supporting Face-recognition within Smart Home Virtual Key Management System

The house/property owner decides to extend the basic use-case through the addition of a door mounted camera to enable face-recognition capabilities over the provided temporary virtual key access solution. Such extension would help the house owner to validate if the person who is trying to enter the house is from an allowed list of faces, and additionally owns a correct virtual key. Such an extension would demand installing an en-adding more computational resources to support the facial recognition application needed to identify.

For the local based platform shown in 4a, such an extension would demand more computational resources to support facial recognition application of the camera. As the architecture is a preset local one, the owner would end up replacing the platform with a totally new one that supports the needed camera setup.

Figure 4b presents architecture modification for such extension where a new node with bigger capabilities has replaced the original node used to cover basic use-case and the camera has been connected to the platform beside existing smart door lock. Such node will include a newly developed or extended software application to realize face recognition based access.

Cloud Based platform might support such extension if the mounted camera has the required capabilities to be directly connected to the cloud. Figure 5b presents the required architecture change through adding such Internet connected Camera and re-implement/extending the existing cloud hosted/stored temporary virtual key access application to receive the camera feed and support face-recognition based access. The cloud cost package used to cover the basic use-case might need to be upgraded. Since camera feed still needs to be sent for processing on the server, that solution suffers from latency issues. Camera Feed now is accessed/owned by the cloud provider thus leading to privacy issues as anyone could easily access that stream of data. Furthermore, in-case of a limited bandwidth situation this solution might not be applicable all together where the limited bandwidth might affect camera feed transmission to the cloud for recognition.

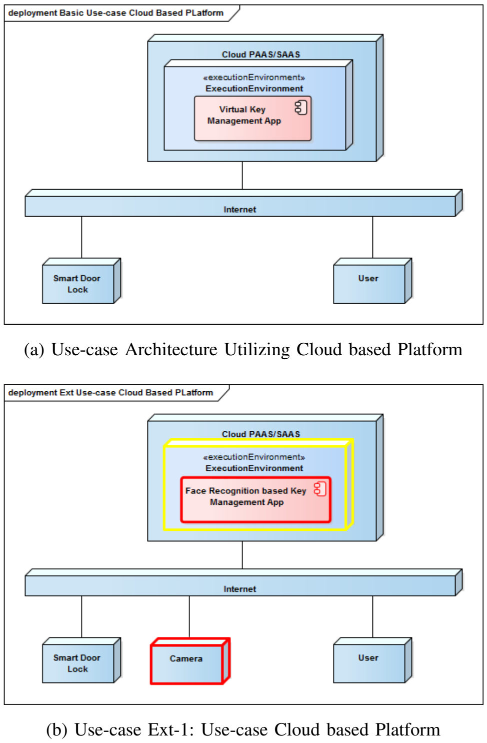
C. Supporting Window Status Monitoring within Smart Home Virtual Key Management System

On the other hand if the house/property owner decides to further extend the smart home through adding a door/window status sensors. Such sensors might be used to detect if door/window status is opened or closed, so that he can monitor/confirm his hours/property status; he will suffer from the same limitation described in first extension. Another possible use case extension is to add a door/window status monitoring application. Such application kind is considered a safety critical application assume that the smart house's door monitoring system crashes suddenly while the house owner is away from the house. With only one instance of the monitoring system, a thief would easily access the house/property. Accordingly two instances of the monitoring system need to be present. Such two instances imply un-needed extra cost in both cloud and local based solutions.

Although there exists an under-utilized processing power that not used all the time for example facial recognition door access processing resources is only used during door accessing process. Home automation systems and IoT applications in general are scalable, dynamic and heterogeneous by nature. Use-case(s) extensions are infinite. There is a high need for dynamic, scalable platform/middle-ware that absorb and support such nature. In the next section we will present a proposed dynamic scalable computing platform architecture for IoT hub(s). After that we will evaluate it against the described use-case with its extensions.

IV. Proposed Architecture

The following section will present a dynamic scalable computing IoT platform architecture used to manage IoT networks/Sub-networks resources. Main objectives for such architecture are to (i) Secure the required computing resources for different IoT applications without breaching users' privacy. Furthermore, such computing resources should not lead to unneeded latency, specialty within applications with real-time demands. (ii) Abstract the underline IoT network infrastructure for such IoT applications. Through such abstraction IoT applications will be totally decoupled from underline hardware constrains as well as sensors/actuators availability and providers. (iii) Orchestrate the network and (IV) Maximize resource utilization.



(a) Use-case Architecture Utilizing Cloud based Platform

(b) Use-case Ext-1: Use-case Cloud based Platform

Fig. 5. Cloud Based Platform Use-case(s) Architecture

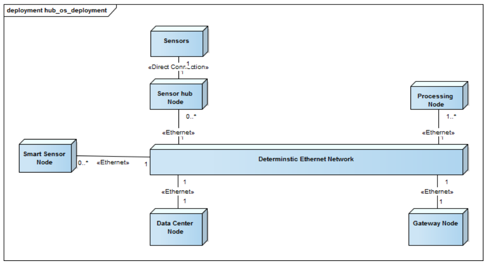


Fig. 6. Proposed Platform Architecture Network Topology.

A. Architecture Overview

To achieve the proposed architecture two main objectives will be introduced:

- Topology Change:
  Hub/Subnetwork topology need to be introduced
- Software Architecture Change:
  Extending service oriented architecture concepts to bring cloud flexibility into the local based platforms.

Topology changes will be discussed in section IV-A1, and Software architecture changes will be presented in section IV-A2.

1) Platform Hardware Network Topology: Figure 6 illustrate the proposed architecture Network topology which contains the following:

- Data Center Node:
  A standalone device/machine used to host the platform

system services and store the available IoT applications package.

- Processing Node(s):
  A one or more standalone device(s)/machine(s) used to host IoT application while executing. These nodes could be added or removed dynamically and the platform will adapt itself and the running application against such situations.

- Gateway Node:
  A standalone device/machine used to act as a translation unit that bridge the communication between IoT Hub/sub-network currently managed by the platform and external world.

- Sensors Hub Node:
  A standalone device/machine used to group number of different or similar sensors to be exposed for IoT application(s). This node is optional

- Smart Sensor(s) Node:
  A standalone sensor that directly connected to IoT Hub/Sub-network.

- Time Sensitive Networking:
  A network medium that support IEEE TSN which has clock synchronization profile 802.1AS based on 1588v2 and messages are forwarded as part of scheduled queues 802.1Qbv.

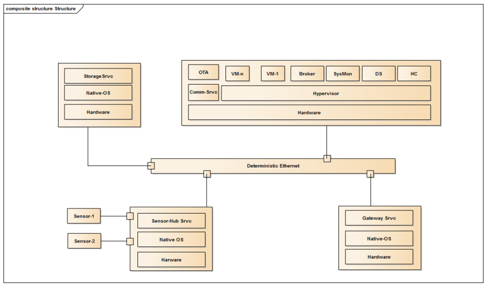


Fig. 7. Proposed Platform Software Architecture.

2) Platform Software Architecture: Figure 7 illustrate the proposed platform architecture software system services within IoT hub network which includes the following services:

- Data Storage Service:
  A service to store IoT application(s) packages in a storage efficient manner while providing them on-demand over network connection. IoT application package should contains (1) IoT application binary image that hold the application executable as well as all its dependency in a container or Virtual Machine image format. (2) Application manifest file which contains all requirement to be provided by the platform i.e. maximum needed cpu load, required specific architecture (X86 or Arm) existing of acceleration, memory needs..etc.

- System Monitor Service:
  A service used to monitor the overall platform status including which IoT application is running over which processing node, loads of each processing node as well as the availability of one or more sensor(s) or services.

- Broker Service:
  A service that manage IoT application(s) scheduling [activation and shutdown] over the clustered processing nodes. This service abstracts the scheduling algorithms to maximize the resource utilization through using all available processing nodes or minimize the power usage through packing all application VMs to as minimum as required processing nodes.

- Node-Control Service:
  A service used to manage the underneath processing node as a generic computing unit, it will dynamically configure, load and execute an IoT application sent through network by Broker Service and gather real-time statistics information about it and send it to System Monitor Service.

- Sensor-As-a-Service Service:
  A service that abstract the underneath sensors/actuator and facilitate its discovery and usage by the platform. this service might be part of Sensors/actuators Hub Node or smart sensor/actuator node.

- Gateway Service:
  A service used to bridge the platform specific Ethernet communication to other external different Networks or Internet.

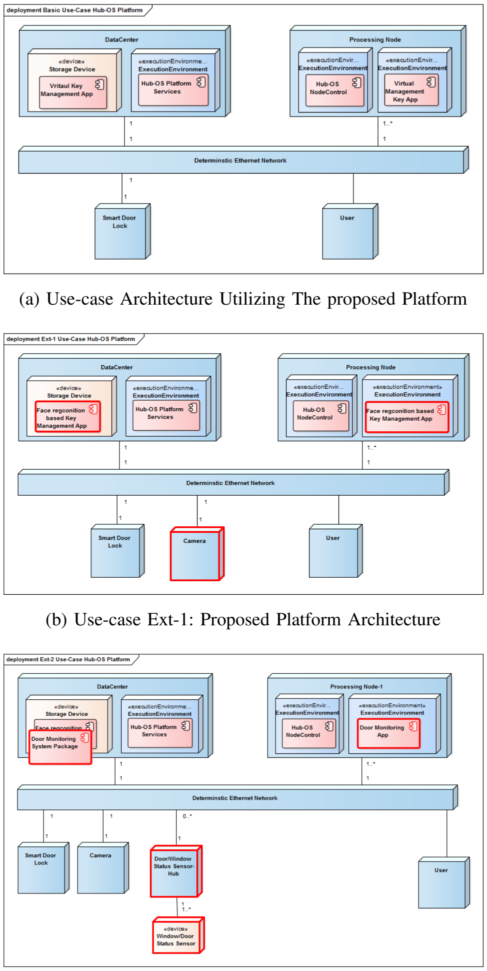
- OTA Service:
  A service used to manage IoT application packages versions and control their updates inside data center node storage.

V. Applying Proposed Architecture To Temporary Virtual Key System

Referring to our motivational scenario presented in Section III we will apply our proposed architecture to alive-ate the limitations of both local and cloud based IoT platforms.

Figure 8a presents basic use-case architecture based on the introduced platform architecture. Data Center Node will store temporary virtual key management system application package. This package contains (i) virtual key management software executable binary container image, (ii) a resource configuration file that describes the required resources what is the maximum amount of memory and CPU resource needed, the required access to certain sensor(s) or actuators and (iii) an optional Application activation binary which is a standalone executable that interact with the architecture platform Sdk to Control the application execution based on certain condition i.e Existing of a sensor output of another application. Data Center Node will host also system services described in section IV. Processing Node which contains a Node Control system service. Such service will start/terminate/monitor the temporary virtual key management software application in separate Execution environment based on need. Smart Door Lock actuator that is controlled by application executed inside processing node.

Scalability is one of the main features of the presented platform architecture. To support Use-case Extension for face



(a) Use-case Architecture Utilizing The proposed Platform

(b) Use-case Ext-1: Proposed Platform Architecture

(c) Use-case Ext-2: Proposed Platform Architecture

Fig. 8. Proposed Platform Use-case(s) Architecture

recognition based virtual key management system we will need to connect the Camera to platform network and update the software to process the camera feed and implement the required application logic. Such application will be hosted/executed on the same processing node used to host/execute the basic use-case application. Figure 8b present such required extension. Camera Feed is locally processed inside local processing node so it's not suffer from neither latency or privacy issues.

To support use-case extension-2 that include door and window monitoring system; The proposed architecture sensor hub node will be added to encapsulate/abstract different number of door/window status sensors. Monitoring Application package will be stored on Data-Center node. Processing Node will host/execute the monitoring application in a separate execution environment. Figure 8c presents such modified architecture. Since door and Window monitoring system is a consider a safety critical application which need redundant execution; The Proposed Architecture easily support such safety related requirement either by allowing for executing two different instance(s) of the application inside a single processing node or into two different nodes.

Unlike the previous platforms architecture (Local and Cloud based), the proposed one provides a scalable, reliable solution with supportive capabilities for real-time and safety critical applications in fully controlled private environment.

VI. Proposed Architecture Comparison against other Architectures

Referring to non-functional requirement discussed in Section II; we will compare our proposed approach against each one of them. Table II show a comparison between the proposed architecture approach and existing Local/Cloud based IoT platforms.

- Scalability:
  The proposed Architecture is build from the ground-up to support be scalable in both Hardware and Software. To extends processing resource capabilities of the platform a Processing node will be attached to the platform network and it will be dynamically discovered by SysMon system service and be available to Broker system service to host application(s).
- Reliability:
  The proposed Architecture separates software storage-node from execution i.e. application package is stored inside a centralized data center node and executed on another processing node based on its availability. With support of SysMon system service as a global monitoring system for the platform; a failure in an application could be easily detected and re-executed. Even in case of Hardware failure the application still could be scheduled to be executed in one of the other available processing nodes.
- Privacy:
  The proposed Architecture is a hub/sub-network manager and orchestrator where every-thing is hosted and executed locally with full data owner-ship and control.
- Timing for Real-Time Support:
  The proposed Architecture utilize Time Sensitive network (TSN) to guarantee latency between nodes so it's provide the required system level support for applications that needs a real-time feature.
- Safety Support:
  Based on full flexibility to execute multiple instance(s) from an application either on the same processing node or on a different ones, the proposed architecture secure the required redundancy at minimum cost.

VII. Related Work

Current advance in IoT researches shows a lot of platforms developments. Large number of these platforms has been surveyed in [8] [10] [11] [12] [13], [14], [15] and [16] concluded the lack of a dynamic scale-able IoT platform that helps in utilizing global system resources, support discovery and composition, reliability, security and privacy; the proposed architecture addresses such features.

Non-functional requirements of IoT platform architectures has been explored separately in the literature. Security and privacy has been addressed in [17], [18] and [19]. In [17] an access control provider (ACP) based solution for has

TABLE II. PROPOSED PLATFORM ARCHITECTURE VS LOCAL AND CLOUD BASED PLATFORMS

Architecture Properties	Local Platforms	Cloud Platforms	Proposed Platform Architecture
Scalability	✗	✓	✓
Reliability	✗	✓	✓
Privacy	✓	✗	✓
Timing for Real-Time	✓	✗	✓
Safety	✗	✓	✓

been introduced to support security and privacy requirements of interoperable IoT architecture without any pre-established secret information. In [18] a cooperative system between internet service provider (ISP) and and home-gateway has been introduced to provide efficient yet privacy-aware IoT security services. In [19] the effect of using for-oriented architecture could be used for improving the user-privacy and a mapping of privacy patterns to IoT fog/ cloud architecture has been introduced. Conceptually these work complements the proposed architecture which secure the needed resources to realize and implement such techniques.

Where, Real-Time support has been address in [20], [21] and [22]. In [20] a design for building evacuation as a real-time emergency safety critical IoT application where real-time performance and evacuation time are critical. The proposed architecture easily support such kind of use-case implementation through securing the needed resources to provide a collaborative distributed approach for such applications. In [21] a network optimization techniques has been surveyed as one of the enabler technologies to support real-time application in IoT platforms. they complements the proposed architecture. In [22] IoT Fog computing architecture has been used to leverage user-centric technologies that bring the IoT control and analytics closer to the user and cover latency and real-time support gap in cloud based IoT platform solutions. a fog sensing concepts has been introduced and their major challenges has been analyzed. an IoT-in-the-Fog controller has been introduced that used to probe local resources and manage communication directly with local fog-mediators.

Moreover, Scalability and dynamic nature supports of IoT systems has been addressed in [23] and [24]. In [23] a software defined IoT units concepts has been introduced to encapsulate a fine-grained IoT resources and capabilities. it automate the configuration and provisioning of IoT application in IoT cloud systems. In [24] UBIWARE, LinkSmart, OpenIOT and CHOReOS IoT middle-wares has been analyzed with respect to Scalability and heterogeneity of dynamic IoT environment and they concluded none of these middle-wares/platforms support fully autonomnus and scalable service registration, discovery and composition. as well as no one of them scales well in service discovery and service composition response time.

On the other hand, Service Oriented architecture (SOA) has been used by literature to address IoT challenges such as interoperability as well as Middle-ware design and implementations [25], [26] and [27].

In [25] a Service Oriented architecture for Home Area Network (SoHAN) has been introduced to facilitate and abstract a network of sensors and/or actuators for application developer over 5G networks. they based on a Sensor node that has a processing capabilities to process the sensor data and send it to a gateway hosted in Home premises which send it to the Cloud/server for processing. this architecture is not scale well still suffer from privacy, and scalability issues compared to the proposed architecture. In [26] SOA has been revisited to address the scale, dynamic and heterogeneity of IoT. they introduce probabilistic approach for service discovery to filter out redundant data and support ultra-large number of things. service composition aggregate data stream within a network to reduce the network load. advice eVolution Service Bus (VSB) to enable interconnection of things that adhere to different interaction style through utilizing set of Binding components (BC) to proxy the sensor specific communication protocol to VSB. the porposed architecture align with such refinement through the system services including for ex sensor as a service (SAS) component which is used to bridge/proxy the sensor communication protocol to the proposed architecture. In [27] Network server as a service has been implemented to enable porting Long Range device (LoRa) networks to the cloud. a LoRaWare is a service oriented architecture that allow the developers to enhance the capabilities of LoRa enabled application has been introduced as well. the main focus was interoperability and exposing the Long Range devices.

VIII. CONCLUSION

A scalable computing IoT platform architecture has been introduced in this paper. The Main objectives of such platform architecture are to (i) Secure the required computing resources for different IoT applications, (ii) Abstract the underline IoT network infrastructure for them, (iii) Orchestrate the network and (IV) Maximize resource utilization. A Smart Home application use-case has been introduced for Virtual Key management system with two extensions (I) Face recognition based authentication virtual key management system (II) Door/Window status monitoring system. An evaluation of the proposed architecture to support such application has been introduced compared to local and cloud based platforms.

REFERENCES

[1] H. Schaffers, N. Komninos, M. Pallot, B. Trousse, M. Nilsson, and A. Oliveira, "Smart cities and the future internet: Towards cooperation frameworks for open innovation," in *The Future Internet*, ser. Lecture Notes in Computer Science, J. Domingue, A. Galis, A. Gavras, T. Zahariadis, D. Lambert, F. Cleary, P. Daras, S. Krco, H. Müller, M.-S. Li, H. Schaffers, V. Lotz, F. Alvarez, B. Stiller, S. Karnouskos, S. Avessta, and M. Nilsson, Eds. Springer Berlin Heidelberg, pp. 431–446.

[2] V. Ricquebourg, D. Menga, D. Durand, B. Marhic, L. Delahoche, and C. Logé, "The smart home concept : our immediate future," pp. 23–28.

[3] C. Constantinescu, D. Lucke, and E. Westkämper, "Smart factory - a step towards the next generation of manufacturing."

[4] H. Farhangi, “The path of the smart grid,” vol. 8, no. 1, pp. 18–28. [Online]. Available: http://ieeexplore.ieee.org/document/5357331/

[5] Gartner top strategic predictions for 2018 and beyond. [Online]. Available: https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2018-and-beyond/

[6] “Six technologies with potential impacts on US interests out to 2025,” p. 48.

[7] D. Evans, “The internet of things: How the next evolution of the internet is changing everything,” *CISCO white paper*, vol. 1, no. 2011, pp. 1–11, 2011.

[8] J. Mineraud, O. Mazhelis, X. Su, and S. Tarkoma, “A gap analysis of internet-of-things platforms,” vol. 89-90, pp. 5–16. [Online]. Available: http://arxiv.org/abs/1502.01181

[9] A. Sabella, R. Irons-Mclean, and M. Yannuzzi, *Orchestrating and Automating Security for the Internet of Things: Delivering Advanced Security Capabilities from Edge to Cloud for IoT*. Cisco Press, 2018.

[10] IoT cloud platform landscape | 2019 vendor list. [Online]. Available: https://www.postscapes.com/internet-of-things-platforms/

[11] J. Guth, U. Breitenbücher, M. Falkenthal, P. Fremantle, O. Kopp, F. Leymann, and L. Reinfurt, “A detailed analysis of iot platform architectures: concepts, similarities, and differences,” in *Internet of Everything*. Springer, 2018, pp. 81–101.

[12] J. Guth, U. Breitenbucher, M. Falkenthal, F. Leymann, and L. Reinfurt, “Comparison of IoT platform architectures: A field study based on a reference architecture,” in *2016 Cloudification of the Internet of Things (CIoT)*. IEEE, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/7872918/

[13] Middleware for internet of things: A survey - IEEE journals & magazine. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7322178/

[14] H. Hejazi, H. Rajab, T. Cinkler, and L. Lengyel, “Survey of platforms for massive IoT,” in *2018 IEEE International Conference on Future IoT Technologies (Future IoT)*, pp. 1–8.

[15] L. Bittencourt, R. Immich, R. Sakellariou, N. Fonseca, E. Madeira, M. Curado, L. Villas, L. DaSilva, C. Lee, and O. Rana, “The internet of things, fog and cloud continuum: Integration and challenges,” vol. 3-4, pp. 134–155. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2542660518300635

[16] V. Bastidas, M. Helfert, and M. Bezbradica, “A requirements framework for the design of smart city reference architectures,” in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[17] N. Fotiou and G. C. Polyzos, “Authentication and authorization for interoperable iot architectures,” in *International Workshop on Emerging Technologies for Authorization and Authentication*. Springer, 2018, pp. 3–16.

[18] H. Haddadi, V. Christophides, R. Teixeira, K. Cho, S. Suzuki, and A. Perrig, “Siotome: An edge-isp collaborative architecture for iot security,” in *Proc. IoTSec*, 2018.

[19] S. Pape and K. Rannenberg, “Applying privacy patterns to the internet of things’(iot) architecture,” *Mobile Networks and Applications*, vol. 24, no. 3, pp. 925–933, 2019.

[20] C. Arbib, D. Arcelli, J. Dugdale, M. Moghaddam, and H. Muccini, “Real-time emergency response through performant iot architectures,” in *International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, 2019.

[21] N. Srinidhi, S. D. Kumar, and K. Venugopal, “Network optimizations in the internet of things: A review,” *Engineering Science and Technology, an International Journal*, vol. 22, no. 1, pp. 1–21, 2019.

[22] S. M. Oteafy and H. S. Hassanein, “Iot in the fog: A roadmap for data-centric iot development,” *IEEE Communications Magazine*, vol. 56, no. 3, pp. 157–163, 2018.

[23] S. Nastic, S. Sehic, D.-H. Le, H.-L. Truong, and S. Dustdar, “Provisioning software-defined iot cloud systems,” in *2014 international conference on future internet of things and cloud*. IEEE, 2014, pp. 288–295.

[24] A. Palade, C. Cabrera, F. Li, G. White, M. A. Razzaque, and S. Clarke, “Middleware for internet of things: an evaluation in a small-scale iot environment,” *Journal of Reliable Intelligent Environments*, vol. 4, no. 1, pp. 3–23, 2018.

[25] M. R. Abd Rahim, R. A. Rashid, A. M. Rateb, M. A. Sarijari, A. S. Abdullah, A. H. F. A. Hamid, H. Sayuti, and N. Fisal, “Service-oriented architecture for iot home area networking in 5 g,” *5G Networks: Fundamental Requirements, Enabling Technologies, and Operations Management*, pp. 577–602, 2018.

[26] V. Issarny, G. Bouloukakis, N. Georgantas, and B. Billet, “Revisiting service-oriented architecture for the iot: a middleware perspective,” in *International conference on service-oriented computing*. Springer, 2016, pp. 3–17.

[27] K. Tsakos and E. G. Petrakis, “Service oriented architecture for interconnecting lora devices with the cloud,” in *International Conference on Advanced Information Networking and Applications*. Springer, 2019, pp. 1082–1093.

KWA: A New Method of Calculation and Representation Accuracy for Speech Keyword Spotting in String Results

Nguyen Tuan Anh[1]
School of Electronic and Information Engineering,
South China University of Technology,
Guangzhou 510641, P.R.China

Hoang Thi Kim Dung[2]
Faculty of Civil and Environment,
Thai Nguyen University of Technology,
Thai Nguyen, Vietnam

***Abstract*—This study proposes a new method to measure and represent accuracy for Keyword Spotting (KWS) problem in non-aligned string results. Our approach, called Keyword Spotting Accuracy (KWA), was improved from the Levenshtein Distance algorithm, that used to evaluate the accuracy of the keywords in KWS by measuring the minimum distance between two strings. The main improved algorithm is to show the status of each keyword in training phase for predicted and true labels. In which, representing which words are correct, which ones need to be inserted, substituted or deleted when comparing the prediction labels with true ones during the training phase. In addition, a new method of presenting the multiple keywords in results was proposed to indicate the accuracy of each keyword. This method can display detailed results by keywords, from which, we can obtain the accuracy, distribution, and balance of the keywords in the training dataset by actual speech variance, not by counting keywords in true labels as usual.**

***Keywords*—Speech Keyword Spotting; KWS; keyword accuracy; Keyword Spotting Accuracy (KWA); speech recognition**

I. Introduction

The objective of this study is researching the evaluating methods of the speech keyword spotting (KWS) problem when results in string. The goal of the KWS problem is to detect predefined keywords in a stream of user utterances, usually used as device wake-up words, speech enable for smart devices or find the keywords in video or audio files.

KWS has developed for many years, with significant progress and quality of algorithms. The methods are also very diverse, using Audio only, without labels [3]. Both audio and label are used for supervised learning, from using traditional methods [8], to the basic forms of deep learning [16], and Deep Neural Network Based types are of great interest [28], [30], [2], [16], [15], with different methods of evaluating results, but all of them have not solved the KWS results as a string.

Currently in speech recognition and KWS, it is possible to classify into two categories: classification and regression. KWS is classified into binary and multiple classes in the classification.

The first type, binary classification, is usually a type of wake-up word, applied in electronic products such as smartphones and smart devices. Some companies are using this type such as Apple with “Hey Siri", Google with “Hello Google", Xiaomi with “Xiao Ai Tong Xue". In this type, it usually only has one keyword, the length of the keyword has little variation in speech data. The KWS’s mission is to find out in a utterance that contains or not a keyword, so it is classified into binary classification problem. For example, with google, a user said “OK Google, open gmap", after the phrase “OK google" is detected, a connection will be opened so that the device can communicate directly to a server, and then the server will do the task in the end of the command that converts “open gmap" into text, understand the semantics and transfer the command to the device to serve the user.

The second type, multiple-class classification, the goal of this type is to classify utterances into groups. Such as in game applications, keywords are forward, backward, left, right, up, down, etc. each keyword is a utterance in the data set with the same length. In 2017, Google has created a dataset with a list of these keywords, called Google Speech Command. This dataset contains 35 keywords, each of them has one second long, classified into 36 separate groups [33].

With regression type, a data set consists of utterances, with different lengths, in each utterance that can be contained or not one or more keywords in a given keyword list. True labels are strings, they are not classified into groups, and the position of each word in speech data also unknown. KWS’s task is to check if the keywords are in utterances, if they are, then which keywords. In essence, this problem is similar to the Speech Recognition problem, but with a much smaller set of word as keywords, the remaining words are garbage [6].

To measure results, in the classification type, there are some methods to do, like confusion matrix, including true positive (TP), true negative (TN), false positive (FP), false negative (FN) and measures based on those values [34], in article [32], they used this method to to present the results. Based on these methods, a model based on parameters is evaluated such as true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), false negative rate (FNR), accuracy (ACC), F1 score. With these methods, it is easy to calculate the confusion matrix, but this method can not apply to string results, because when only one character changes, the comparison result is no longer accurate. In the regression type, there are some system assessment measures such as: word error rate (WER), token error rate (TER), character error rate (CER), word accuracy (WACC). Speech recognition (SR) accuracy measurement is mainly based on word error rate (WER) [36], It is calculated based on the Minimum Edit Distance algorithm, and calculations based on unit of word. WER is an effective tool to compare and evaluate the accuracy of different systems as well as the improvement of a system. In KWS, the concept of Token Error Rate (TER) is also used, instead of using WER, it uses each keyword (possibly containing multiple words) as a unit of calculation. Character Error Rate (CER) is used similarly to WER, but the unit of measurement is based on characters. These methods can evaluate the system accuracy, but if a systems with zeros-resource is developed, we will need more information, such as the number of utterances of each keyword, the accuracy of each keyword, the ratio between accuracy and the number of utterances (because of some languages, like Chinese, there are variation, changing the pronunciation according to the words standing next to each other), if using WER only, it is impossible to know exactly.

There are several methods to evaluate the system, based on the calculation of the correct and incorrect prediction of the predictive labels with the true labels such as Term Weighted Value (TWV), Maximum Term Weighted Value (MTWV) [10]. In paper [4], they used Actual TWV (ATWV), they only consider whether or not the keyword is in the predictive label. In the article [17], they used P@n method to present results of top n keywords. In the article [22], they introduced the DR/FA evaluation method for telephone speech, these methods can evaluate the models, but still evaluate the accuracy of entire keyword set, so the problem of estimating the accuracy of each keyword is still unresolved. it is hard to know how many keywords have correctly predicted, not predicted or missed, when the output of KWS model is a string and when training, only accuracy of entire data set is calculated, by calculating the minimum string distance of predicted labels by true labels. When studying the evaluation method of KWS problem, we found that it is difficult to measure the accuracy of each keyword on predicted results. Because KWS model returns the results as strings, so it is difficult to determine the accuracy in percent of each word. But this analysis is necessary, allowing us to know 11the distribution of each keyword in the data-set, especially with words that have multiple pronouncement ways, mutations and modifications as in Chinese or dialect in other languages, for example, see Fig. 1. The more variation, the more data is needed for a keyword during training. Evaluating a KWS model is to evaluate the accuracy of predicted outputs compared to the true labels in the form of string. This study focuses on solving this problem.

Different from the existing assessment methods, the objective of this study is to provide a method for calculating the accuracy of each keyword in the output sequence of the Regression problem. Proposing a method to display the results on a new chart type so that we can observe the number of keywords in the data set, the number of correct predictive keywords, false predictions and unpredictable, that's also the reason because the name Keyword Accuracy is selected.

II. Theory

Making it easier to compare methods, some theory of representing results for the KWS problem is reviewed. As mentioned above, the existing results representations method can be classified into two categories, classification and regression.

Write	pingyin	Read/say
你好	nǐ hǎo	→ ní hǎo”
我很好	Wǒ hěn hǎo	→ “Wǒ hén hǎo” / “wó hén hǎo”
不爱	Bù ài	→ Bú ài
不变	Bù biàn	→ bú biàn
一共	Yīgòng	→ yí gòng

Figure 1. Chinese characters, when reading and writing differently

TABLE I. TYPICALLY USED ERROR RATES AND THEIR SYNONYMS

Name	Acronym	Formular	Synonyms
False Positive Rate	FPR	$\frac{FP}{FP + TN}$	False Accept Rate (FAR), Fall-out
False Negative Rate	FNR	$\frac{FN}{FN + TP}$	False Reject Rate (FRR), False Alarm Rate
True Positive Rate	TPR	$\frac{TP}{TP + FN}$	True Accept Rate, Sensitivity, recall, Hit Rate
True Negative Rate	TNR	$\frac{TN}{TN + FP}$	True Reject Rate, Detection, Rate, Specificity, Selectivity
Positive Predictive Value	PPV	$\frac{TP}{TP + FP}$	Precision
Accuracy	ACC	$\frac{TP + TN}{TP+TN+FP+FN}$	
F1 score	F_1	$\frac{2TP}{2TP + FP + FN}$	

Classification type is easily calculating results into confusion matrix parameters such as true positive, false positives, false negatives, true negatives. The second type, regression, is a comparison between the predicted string labels and the true labels that currently applied by WER and the result is accuracy over the entire data set. In this study, the regression model is focused for strings predicted results.

The first method, the Confusion matrix and related formulas, aims to evaluate accuracy in binary and multiple class classification. To classify results, with binary classifiers, predictive results is classified into one of the two classes that are real positive cases and real negative cases; With multi-keywords, the results are classified into n*n matrices with n being the number of keywords. In a dataset, the number of real positive cases is called condition positive (P), the number of real negative cases is called condition negative (N). Since then, the predicted results are classified into one of four categories, accurate predictions include true positive (TP) and true negative (TN), incorrect predictions include false positives (FP) and false negatives (FN). From the predicted results, the relevant results is calculated as in Table I, equations obtained from [25], [9], [34], [20]. Finally, we have methods to evaluate results based on those formulas via ROC curves, e.g TPR/FPR [21], Precision/Recall [26], [14], False reject Rate/ False Alarm Rate [7], [29], False Negative Rate/Hourly False Positives [1]

The second method, P@k. In the article [27], the accuracy algorithm was used the formula (1) for evaluating method. The returned result is the accuracy of top k keywords in the system.

$$P@k = \frac{|\{W_r\} \cap \{kW_p\}|}{|\{kW_p\}|} \tag{1}$$

where W_r is relevant words, kW_p is retrieved words, $P@k$ is a precision measurement. The result returns a number, representing the system's accuracy, for example, $P@6 = 0.617$

The third method, TWV. Term Weighted Value (TWV) is a measurement method of KWS system evaluation, introduced in [10], illustrated by the formula (2-5).

$$P_{miss}(\theta) = 1 - \frac{N_{correct}(\theta)}{N_{true}} \tag{2}$$

$$P_{fa}(\theta) = 1 - \frac{N_{incorrect}(\theta)}{N_{Ninc}} \tag{3}$$

$$TWV(\theta) = 1 - (P_{miss}(\theta) + \beta P_{fa}(\theta)) \tag{4}$$

with:

$$\beta = \frac{E}{V}.(Pr^{-1} - 1) \tag{5}$$

where θ refers to detection threshold, $N_{correct}$, $N_{incorrect}$ refer to the number of keyword correct and incorrect detections, respectively. N_{true} refers to the number of occurrences of keywords in that utterance, N_{Ninc} refers to the number of incorrectly detected keywords in that utterance, $P_{miss}(\theta)$ and $P_{fa}(\theta)$ denote the probability of miss and false alarm, respectively. The cost/value ratio, C/V, is 0.1, thus the value lost by a false alarm is a tenth of the value lost for a miss. The prior probability of a term, Pr, is 10^{-4} [10]. Detection score is greater than or equal to θ, The result of this method returns a number to evaluate the system, such as TWV = 0.1962. Recently some articles, such as [11], also use this measure method to represent their results, and the value also returns a number to evaluate the accuracy of their model. In order to evaluate the number of keywords and their correlations, it is necessary to do more in another way. This method can evaluate the accuracy of the model, but in speech, it does not only simply consider that true label and predicted label contain which keywords but also consider the order in which these words appear. So the WER method is based on the Minimum Edit Distance, which is still used in many speech recognition systems. There are two other methods to calculate accuracy based on TWV method of Actual TWV (ATWV) and Maximum TWV (MTWV). ATWV uses actual decisions to represent the system's ability to predict the optimal operating point given by the TWV scoring metric. MTWV is a TWV value of θ yields the maximum TWV [10]. This method is used by some studies such as [5], [12],

The fourth method, MED. The Levenshtein algorithm [18], [35] used to calculate the Minimum Edit Distance (MED) between two strings. Suppose the two strings given for comparison are s and t, the length of the strings is $|s|$ and $|t|$, minimum edit distance is calculated according to the formula (6) ([18], [35]):

$$MED_{s,t}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0 \\ min \begin{cases} MED_{s,t}(i-1,j) + 1 \\ MED_{s,t}(i,j-1) + 1 \\ MED_{s,t}(i-1,j-1) + 1_{s \neq t} \end{cases} & otherwise \end{cases} \tag{6}$$

If $s_i \neq t_j$ then $1_{(s_i \neq t_j)} = 1$ and 0 otherwise, $\mathtt{MED}_{s,t}(i,j)$ is the smallest distance of the first i characters of s compared to the first j characters of t To measure the accuracy of a model, Word Error Rate (WER) is used, calculated according to the formula formula (7) [36].

$$\mathtt{WER}_{s,t} = \frac{S + I + D}{N} = \frac{\mathtt{MED}_{(s,t)}}{N} \tag{7}$$

Where S, I and D represent the number of substitutions, insertions and deletions, N is the number of words in the reference.

In order to evaluate a KWS problem, we have four main methods as mentioned above, but in all of them, there is no one strong enough to calculate the accuracy of each keyword that one or more keywords are inside a string; Displays the balance distribution of each keyword in the data set. That is the motivation for us to carry out this research. Moreover, this study has provided a new way of displaying graphics, thereby fully demonstrating simultaneous information. That is the motivation for this research to be done

III. Propose Method

In this study, we propose an algorithm that calculates the accuracy of the model according to the keyword, with the model output being a string of characters that can have keywords or not, and proposes a new method of representing the results. This one is improved from the Minimum Edit Distance algorithm of Levenshtein for the KWS problem. The output of regression model is a string, to match the multi-lingual problem (like Chinese and Vietnamese, completely different from the structure of words). We introduce an algorithm in equation (8) so called Speech Keyword Accuracy (KWA), to determine the exactly editing position of each keyword, based on the minimum edit distance. To be compatible in multiple languages, each label will be separated into a list of words, in Chinese, separated by each character, in Vietnamese separated by space between words.

$$MED_{s,t}(i,j) = \begin{cases} \begin{cases} i \\ TOC_{1..i,j} = M_{ins} \end{cases} & if\ j = 0 \\ \begin{cases} j \\ TOC_{i,1..j} = M_{del} \end{cases} & if\ i = 0 \\ min \begin{cases} \begin{cases} MED_{s,t}(i-1,j) + 1 \\ TOC_{i,i} = M_{del} \end{cases} \\ \begin{cases} MED_{s,t}(i,j-1) + 1 \\ TOC_{i,j} = M_{inc} \end{cases} \\ \begin{cases} MED_{s,t}(i-1,j-1) + 1 \\ TOC_{i,j} = M_{sub} \end{cases} & if\ s_i \neq t_j \\ \begin{cases} MED_{s,t}(i-1,j-1) + 1 \\ TOC_{i,j} = M_{eq} \end{cases} & if\ s_i = t_j \end{cases} & otherwise \end{cases} \tag{8}$$

In the KWA algorithm in equation (8), the input is provided by two lists s, t and a list output `TOC` (abbreviation of type of changes), in which each element is equal, substitution, insertion or deletion, denoted by M_{eq}, M_{sub}, M_{inc} and M_{del}, respectively, each of them is a constant number. The result is updated to a global variable, from there, accuracy of each keyword is obtained as in equation (12), the accuracy of the whole model across the dataset as definition in equation (13).

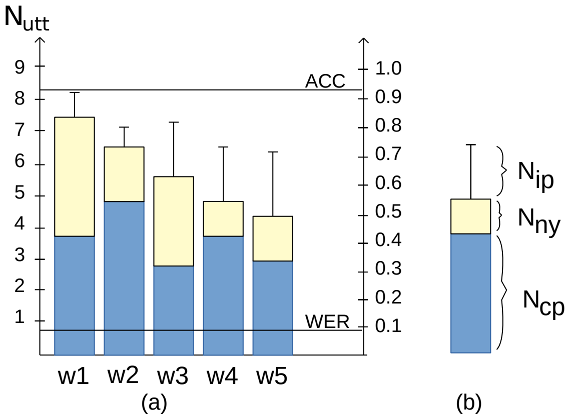


Figure 2. Example of presentation of Speech Keyword Accuracy (KWA) algorithm
N_{utt}: Number of utterances,
w_i (i=1,2...): predefined keywords,
ACC: Model's accuracy,
WER: keyword error rate of model,
N_{ip}: Number of keywords incorrectly predicted (not in true label),
N_{ny}: The number of keywords not yet predicted,
N_{cp}: Number of keywords correctly predicted.

WER based on TOC also observed as in equation (7), where, in each utterance, parameters is calculated as in equation (9-11).

$$S_i = \sum_j (TOC_{i,j} == M_{sub}) \tag{9}$$

$$I_i = \sum_j (TOC_{i,j} == M_{inc}) \tag{10}$$

$$D_i = \sum_j (TOC_{i,j} == M_{del}) \tag{11}$$

or $WER = MED_{s,t}/N$.

This study also propose a method to presenting results in a graph to easily observe the accuracy of each keyword in the keywords set. In Fig. 2, The total number of each keyword occurrences denote as N_{kw}: $N_{kw} = N_{ny} + N_{cp}$. This representation method tells us the overall WER of that system, the number of keywords, the status of each keyword, how many percent each keyword predicted correctly, correlation in terms of number of keywords included in dataset and the number of incorrectly predicted words and not yet predicted. That information can be read along the vertical axis on the left. According to the vertical axis on the right, the results in accuracy as a percentage and WER can be observed, either of which may be missing. During training, incorrectly predicted words can have many reasons, which may be due to lack of data, imbalance in the data set (in classification of images dataset or isolated speech dataset maybe easier to identify than speech recognition dataset). From here, in training process, we will be known that which keywords is needed to prepare more training data so each keyword can be balanced on WER with others. The formula for calculating ACC [23] for each keyword (acc_i) is given in equation (12), and global ACC can be calculate as in (13).

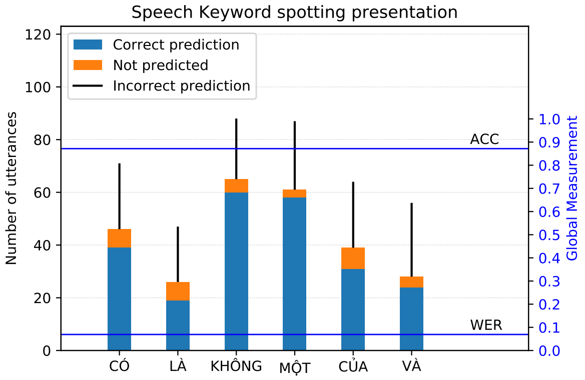


Figure 3. The graph shows the correlation of results between keywords of ViVos dataset

$$acc_i = \frac{N_{cp} - N_{ip}}{N_{cp} + N_{ny}} \tag{12}$$

$$ACC = \frac{1}{N} \sum_{i=0}^{N-1} acc_i \tag{13}$$

where $N_{cp}, N_{ip} N_{ny}$ refer to number of correctly predicted, incorrectly predicted and not predictable, respectively. N denotes as the number of utterances in the dataset. Here, parameters is calculated as equation (14-16)

$$N_{cp}(i) = \sum_j (TOC_{i,j} == M_{eq}) \tag{14}$$

$$N_{ip}(i) = \sum_j (TOC_{i,j} == \{M_{del} | M_{sub}\}) \tag{15}$$

$$N_{ny}(i) = \sum_j (TOC_{i,j} == M_{inc}) \tag{16}$$

IV. Experiments and Results

To do the experiment, we selected two small database sets, representing the low-resources languages, ViVos and THCH-30.

A. Dataset

THCHS-30 corpus. THCHS-30 corpus is an open speech Chinese database [31], publicized in openslr [24], for a total of up to 30 hours for free of reading audios with labels, recorded in a quiet room. This corpus has the characteristics as shown in Table II. To get results for the KWS problem, 10 keywords are selected and implemented by taking 10 words with the highest occurrence frequency in the entire data set to perform the test. After selecting, we have the following keyword list:

KW=[的, 一, 有, 人,了, 不, 为, 在, 用, 是]
(De, yī, yǒu, rén, le, bù, wèi, zài, yòng, shì)

ViVos corpus. ViVos corpus is a open speech Vietnamese data set [19]. It includes 15 hours of voice recording for Automatic Speech Recognition (ASR) purposes. published by

TABLE II. STATISTICS OF THCHS-30 DATABASE

Dataset	Speaker	Male	Female	Age	Utterance	Duration(hour)
Train	30	8	22	20-50	10893	27:23
Test	10	1	9	19-50	24	6:24

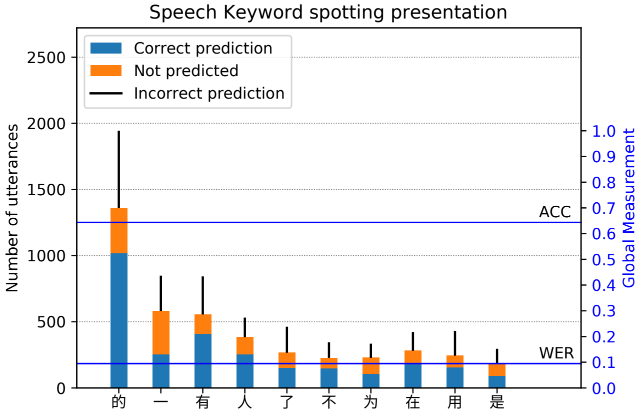


Figure 4. The graph shows the correlation of results between keywords of THCH-30 dataset

AILAB, VNU's computer science laboratory - Hanoi University of Technology. Descriptive characteristics are shown in Table III. The method of selecting keywords is the same as on THCH-30 dataset, and the keyword list has been selected including 6 keywords as:

KW= [Có, Là, Không, Một, Của, Và]

These two sets of data will be used to train with LSTM-CTC model based on [13], outputs of the model and true labels are saved to calculate KWA and display results.

B. Presentation Method

Both ViVos and THCH-30 data sets are trained by LSTM-CTC model, during training, the model is evaluated by CTC loss, based on [13]. CTC loss does not show us how much the accuracy of the model is, but it is possible to evaluate the same model, the same data set, which training session has lower loss, the weight is better. From there the training system can be optimized, to give out the predicted results of the model and combine it with true labels, calculate accuracy according to each keyword and overall accuracy. The formula (12) and (13) are used. The result of this step, is shown on the graphic.

In Figure 3, we can observe, firstly, the number of each keyword is small, and therefore, the difference between the keywords is small, but the percentage is large. Secondly, although the model of accuracy results is quite high, but the percentage of incorrect prediction is also high, and finally, observing WER and accuracy of the system visually, giving us an overview of the model.

TABLE III. STATISTICS OF VIVO CORPUS

Dataset	Speaker	Male	Female	Utterance	Duration(hour)	Unique Syllables
Train	46	22	24	11660	14:55	4617
Test	19	12	7	760	00:45	1692

In Figure 4, it can easily be observed that a huge difference in the number of keywords, the first keyword has approximately twice to sixth times the number of remaining keywords, this leads to difficult for training model to get higher accuracy for the entire set of keywords in the dataset. On the other hand, it is observed that in the second keyword bar, ACC of this keyword has not reached about 50%, while other keywords having higher ACC, thereby giving us a clue to understanding the cause of global ACC is not high.

V. CONCLUSION

We have just described an improved speech keyword spotting accuracy measurement method (KWA) and a new way of presenting results, providing useful information for the training deep learning model. With the KWA method, the accuracy, WER, keyword accuracy, hit/miss in two string label sets, predicted label and true label sets, were evaluated based on improved minimum edit distance algorithm. Besides, the KWA presenting method also provides a figure that we can observe how many keywords correctly, incorrectly predicted and not yet predicted out,the accuracy and WER of the model in a figure. This method helps us understand the balance of keywords in the data set instead of WER or accuracy only. Despite many advantages, KWA still cannot avoid such complex drawbacks. Only string data should be used. In many cases it is not necessary to use an accuracy rating to each keyword. This method can be applied to Speech Recognition problem for almost zero-resource languages and semi-supervised ASR, which will be our future research work.

REFERENCES

[1] A. Abdulkader, K. Nassar, M. Mahmoud, D. Galvez, and C. Patil. Multiple-instance, cascaded classification for keyword spotting in narrow-band audio. *ArXiv*, abs/1711.08058, 2017.

[2] M. A. Al- Rababah, A. Al-Marghilani, and A. A. Hamarshi. Automatic detection technique for speech recognition based on neural networks inter-disciplinary. *International Journal of Advanced Computer Science and Applications*, 9(3):179–184, 2018.

[3] M. Awaid, A. H., and S. A. Audio Search Based on Keyword Spotting in Arabic Language. *International Journal of Advanced Computer Science and Applications*, 5(2):128–133, 2014.

[4] Y. Bai, J. Yi, H. Ni, Z. Wen, B. Liu, Y. Li, and J. Tao. End-to-end keywords spotting based on connectionist temporal classification for mandarin. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE, 2016.

[5] Y. Bai, J. Yi, H. Ni, Z. Wen, B. Liu, Y. Li, and J. Tao. End-to-end keywords spotting based on connectionist temporal classification for Mandarin. In *Proceedings of 2016 10th International Symposium on Chinese Spoken Language Processing, ISCSLP 2016*, 2017.

[6] E. Chandra and K. A. Senthildevi. Keyword Spotting: An Audio Mining Technique in Speech Processing – A Survey. *IOSR Journal of VLSI and Signal Processing Ver. II*, 5(4):22–27, 2016.

[7] G. Chen, C. Parada, and G. Heigold. Small-footprint keyword spotting using deep neural networks. *Acoustics, Speech and Signal ...*, i:1–5, 2014.

[8] H. F. C. Chuctaya, R. N. M. Mercado, and J. J. G. Gaona. Isolated Automatic Speech recognition of Quechua numbers using MFCC, DTW and KNN. *International Journal of Advanced Computer Science and Applications*, 9(10):24–29, 2018.

[9] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[10] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddingtion. Results of the 2006 spoken term detection evaluation. In *Proc. sigir*, volume 7, pages 51–57, 2007.

[11] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[12] M. J. F. Gales, K. M. . Knill, A. Ragni, and S. P. . Rath. Speech recognition and keyword spotting for low resource languages: Babel project research at CUED. In *Spoken Language Technologies for Under-Resourced Languages (SLTU)*, number May, pages 14–16, 2014.

[13] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772, 2014.

[14] Y. Huang and W. Y. Wang. Deep Residual Learning for Weakly-Supervised Relation Extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1803–1807, 2017.

[15] P. D. Hung, T. M. Giang, L. H. Nam, and P. M. Duong. Vietnamese speech command recognition using Recurrent Neural Networks. *International Journal of Advanced Computer Science and Applications*, 10(7):194–201, 2019.

[16] M. K, I. A., and G. Onwodi. Neural Network Based Hausa Language Speech Recognition. *International Journal of Advanced Research in Artificial Intelligence*, 1(2):39–44, 2012.

[17] H. Kamper, G. Shakhnarovich, and K. Livescu. Semantic keyword spotting by learning from images and speech. *arXiv preprint arXiv:1710.01949*, 2017.

[18] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[19] H.-T. Luong and H.-Q. Vu. A non-expert kaldi recipe for vietnamese speech recognition system. Technical report, 2016.

[20] S. Marcel, M. S. Nixon, and S. Z. Li, editors. *Handbook of Biometric Anti-Spoofing*. Advances in Computer Vision and Pattern Recognition. Springer London, London, 2014.

[21] R. Menon, H. Kamper, J. Quinn, and T. Niesler. Fast ASR-free and almost zero-resource keyword spotting using DTW and CNNs for humanitarian monitoring. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, volume 2018-Septe, pages 2608–2612, 2018.

[22] J. Nouza and J. Silovsky. Fast keyword spotting in telephone speech. *Radioengineering*, 18(4):665–670, 2009.

[23] A. Ogawa, T. Hori, A. Nakamura, A. Ogawa, T. Hori, and A. Nakamura. Estimating speech recognition accuracy based on error type classification. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(12):2400–2413, 2016.

[24] Open Speech and Language Resources. Thchs-30. http://www.openslr.org/18. [Online; accessed 31-March-2019].

[25] D. M. W. Powers. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2007.

[26] I. Pratikakis, K. Zagoris, B. Gatos, G. Louloudis, and N. Stamatopoulos. ICFHR 2014 Competition on Handwritten Keyword Spotting (H-KWS 2014). In *Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR*, volume 2014-Decem, pages 814–819, 2014.

[27] I. Pratikakis, K. Zagoris, B. Gatos, G. Louloudis, and N. Stamatopoulos. Icfhr 2014 competition on handwritten keyword spotting (h-kws 2014). In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 814–819. IEEE, 2014.

[28] J. Ren and M. Liu. An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks. *International Journal of Advanced Computer Science and Applications*, 8(12):48–52, 2017.

[29] T. N. Sainath and C. Parada. Convolutional Neural Networks for Small-footprint Keyword Spotting. *Proceedings INTERSPEECH*, pages 1478–1482, 2015.

[30] M. Walid, B. Souha, and C. Adnen. Speech recognition system based on discrete wave atoms transform partial noisy environment. *International Journal of Advanced Computer Science and Applications*, 10(5):466–472, 2019.

[31] D. Wang and X. Zhang. Thchs-30: A free chinese speech corpus. *arXiv preprint arXiv:1512.01882*, 2015.

[32] Z. Wang, X. Li, and J. Zhou. Small-footprint keyword spotting using deep neural network and connectionist temporal classifier. *arXiv preprint arXiv:1709.03665*, 2017.

[33] P. Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[34] Wikipedia contributors. Confusion matrix — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=881721342, 2019. [Online; accessed 31-March-2019].

[35] Wikipedia contributors. Levenshtein distance — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Levenshtein_distance&oldid=887999285, 2019. [Online; accessed 28-March-2019].

[36] Wikipedia contributors. Word error rate — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Word_error_rate&oldid=888037079, 2019. [Online; accessed 31-March-2019].

A Visual Analytics System for Route Planning and Emergency Crowd Evacuation

Saleh Basalamah
Umm Al-Qura University
Makkah, Saudi Arabia

***Abstract*—Emergency evacuation from crowded public spaces is of great importance to all authorities around the world. Many systems have been developed by researchers to address the optimization of emergency evacuation routing and planning. This paper presents a visual analytics system for route planning and emergency crowd evacuation; a web-based visualization and simulation system that allows stakeholders to develop and assess route planning and evacuation procedures for emergency scenarios. The system takes advantage of the available OpenStreetMap comprehensive spatial database to enable users to implement evacuation scenarios in almost anywhere OpenStreetMap dataset is available. Using multiple infrastructure-specific varying scenarios, such as adjusting capacities of roads/pathways or their closure, the tool can identify bottleneck areas thus allowing the assessment of potential improvements to pedestrian and transportation network to relieve the bottleneck and improve evacuation time. As a case study, we use this system for the city of Makkah in Saudi Arabia and the city of Minnesota in the United States.**

Keywords*—Emergency evacuation; crowd management; visualization; intelligent systems

I. Introduction

Emergency evacuation is the movement of a large number of people from one place to another in a short period of time in response to an emergency. This is due to different types of emergencies; natural disasters, terrorism, stampede or war. Throughout the years, many emergency evacuation incidents occurred which needed mass evacuation. Planning of evacuation scenarios plays an important factor in the success of a mass emergency evacuation and many evacuations failed for the lack of good planning. Authorities in different parts of the world provide guidelines and help materials for people to use in case of an emergency evacuation; e.g. www.ready.gov

Integrated decision support systems are used to assist authorities to respond to emergencies in an optimized way [1]. They support strategic and operational decision making in terms of using resources and planning scenarios.

In this regards, we present a visual analytics system for evacuation route planning. Most of the small and large-scale events require an emergency evacuation tool to handle any disaster conditions including flood, fire, or any terrorist attack. The visual analytics system helps to identify the evacuation routes and time based on number of evacuees. This system helps authorities to make a plan to handle any emergency scenarios by changing number of evacuees, shelter points and by closing or expanding some paths. This will help decision-makers in the selection of optimized number of shelters, maximum number of evacuees that can be evacuated in a given time, impact on time and flow of the evacuation plan by closing or expanding some roads. Unlike existing systems, our system is generic and easily applicable to any outdoor events. The system takes advantage of the available OpenStreetMap (OSM) comprehensive spatial database to enable users to implement evacuation scenarios in almost anywhere OSM dataset is available. The system converts OSM dataset into spatial relational database. Moreover, our system allows end user to enhance the spatial relational database interactively such as assigning the width of each road to calculate the capacity, source point, number of evacuees in each source, shelter points and capacity of each shelter point. The visual output of the evacuation routes and time graph makes the system more convenient for analytics. To prove our system capabilities, we demonstrate the system for the city of Makkah, Saudi Arabia and Minnesota, United States of America.

II. Literature Review

Emergency situations like fires, floods, earthquakes and terrorist attacks pose threat to the people living at the vicinity. A system is therefore needed to help the stakeholders and decision-makers to be in the state of readiness to efficiently and effectively evacuate large crowd from such areas of threat. The state-of-the-art [2] of routing algorithms for static and continuous-time network produces paths and plans, however, these approaches are confined to building and small areas. Similarly, the routing algorithms for single-source shortest path, all-pair shortest path, time dependent network provide nearest and shortest destination and time but neglects road capacity constraints [3]–[5]. Likewise, work done in [6]–[8] are restricted to evacuating buildings. Moreover, evacuation of evacuees in building has been done using wireless sensor networks with the demonstration of simulation on static data [9]. Evacuating crowds on a large scale was studied using the Capacity-Constrained Route Planning (CCRP) [10], [11], nearest exist or shelters (NES) [12], and crowd-separated allotment of routes and shelters (CARES) [13].

CCRP heuristic approach algorithms [10], [11] contemplates on shortest path algorithms with capacity constraints of each node and edge of the road. Nearest terminus along with the edge capacity is considered by NES [12] algorithm but it violates the destination capacity. CCRP and NES can result in movement turbulences and stampede in case of evacuating huge crowd. The algorithm, CARES [13], overcomes the drawbacks of above mentioned algorithms and takes into consideration the capacity constraint along with spatial anomaly to avoid movement turbulence and stampede. Haj, results in huge gathering every year in the city of Makkah, and the

CARES algorithm demonstrated the evacuation of large crowd in pedestrian mode very well.

Our proposed visual analytics system for route planning and evacuation uses CARES [13] algorithm. The system, can work on any part of the world where OSM data set is available. Instead of developing from the scratch for different venue, the proposed system allows end-user to do little tuning by providing details about the roads along with the number of evacuees, source and destination points.

III. System Overview

The architecture of the proposed system is depicted in Figure 1, which consists of two systems; a backend-system and a front-end web-based visualization system. Towards the development of these two systems, two tasks have been pursued.

A. Back-end Database and Computational System

We have developed and tested an integrated architecture of the back-end system of the proposed tool that consists of various databases maintaining data from numerous channels, depicted in Figure 1. Using the OpenStreet map, the tool generates a graph modeling the road network for the area under consideration for evacuation. The travel time of each segment of a road/street is computed based on the end points of the segment. Roads are classified and tagged according to their width. Subsequently, the capacity of each road is stored in a table based on its type. Such type-capacity table is provided by the user. A key component of the backend system is the CARES algorithm which accesses the backend databases and generates the evacuation plan for the selected area by identifying routes to a set of pre-selected shelters. Evacuation scenarios are orchestrated through the front-end visual interface which allows a user to select various sites from the digital map database requiring evacuation. In addition, it allows the user to select the destination sites for the evacuees. Such interaction between the user and the proposed system is supported by the front-end system. The user can interactively specify constraints specific to route planning during emergency evacuation, prior to the execution of the CARES algorithm. Constraints can include partial/full closure of some roads/streets from the map which may have blockage, due to construction or impact of on-going emergency, such as growing level of flooding. The pre-computed values for capacities of routes and the travel times based on the physical environment due to the changing context, are utilized by the system. The physical environment depicts the geo-spatial characterization of a road/pathway in terms of the degree of its curvature, its uphill/downhill slope, the width, and blockages etc. For this purpose, the system uses the route map of the city under consideration, the City of Makkah and the Hajj premises, in our case. A high-level pseudo-code for CARES algorithm is given in Algorithm. 1.

Algorithm 1: High-level Pseudocode

```
Data: $dataInput$:
Undirected graph, travel time and capacity of all edges, list
of shelters along with their capacities
Result: Route allotment and total evacuation time
begin
    Allocate all sources to nearest shelters
    Check all shelters alloted occupancies and calcuate their
      weight
    Balance violated shelters to remove violation and
      re-route sources to the new shelter and avoid
      criss-cross
end
```

B. Front-end System

In this task have developed the front-end visual analytics system to analyze the various emergency scenarios and pre-planning exercise for emergency evacuation, as well as decision support capability for route planning, as a part of the pre-planning exercise. The system provides multi-level linked views and interactive displays that allow interactive analysis, as shown in Figure 2. User can provide input to generate various evacuation scenarios. These scenarios comprise of the point(s) of incident and number of evacuees at each point of incident. Potential destination sites for the evacuees can be identified in the city of Makkah based on Metro land road connectivity including Jamaraat, Arafat, and Kudai Parking. Based on crowd modeling given in [15], [16], we have identified the capacity and travel speed as shown in the table in Figure 3.

The tool allows users to assess potential increase or decrease in risks associated with selecting evacuation sites in terms of performance factors including total evacuation time, potential number of people evacuated, and management of routes. In essence, the tool can provide a thorough assessment of all evacuation operations conducted by the Ministry of Hajj. In addition, the tool's functionality can be extended to provide officials an automated capability to make decisions about optimal allocation of evacuation sites under various emergency scenarios.

IV. Scalability

The scalability is one of the important aspect of the system. It becomes more challenging when the system is related to evacuation. In evacuation cases, there are too many variations starting from small area or stadium to an entire city. To make the system scalable, the evacuation system should be intelligent enough to handle all such scenarios. The proposed system assuming n is the number of nodes (road intersections), m is the number of edges (roads), s is the number of shelters, and p is the number of evacuees, the computational complexity of the tool is

$$O(p * s * (n * log(n))). \tag{1}$$

Since the complexity is sub-quadratic in network size and the number of evacuees, the system provides a scalable strategy for evacuation planning. The extraction of nodes and edges of the data from OSM are discussed in next section.

V. Graph Generation and Scenario Selection

In this section, we demonstrate several functionalities of the tool. These functionalities illustrate both the intelligent component (the back-end system) and the visual analytics capability. Following are the database functions that have been implemented in the backend system and are required to run the tool.

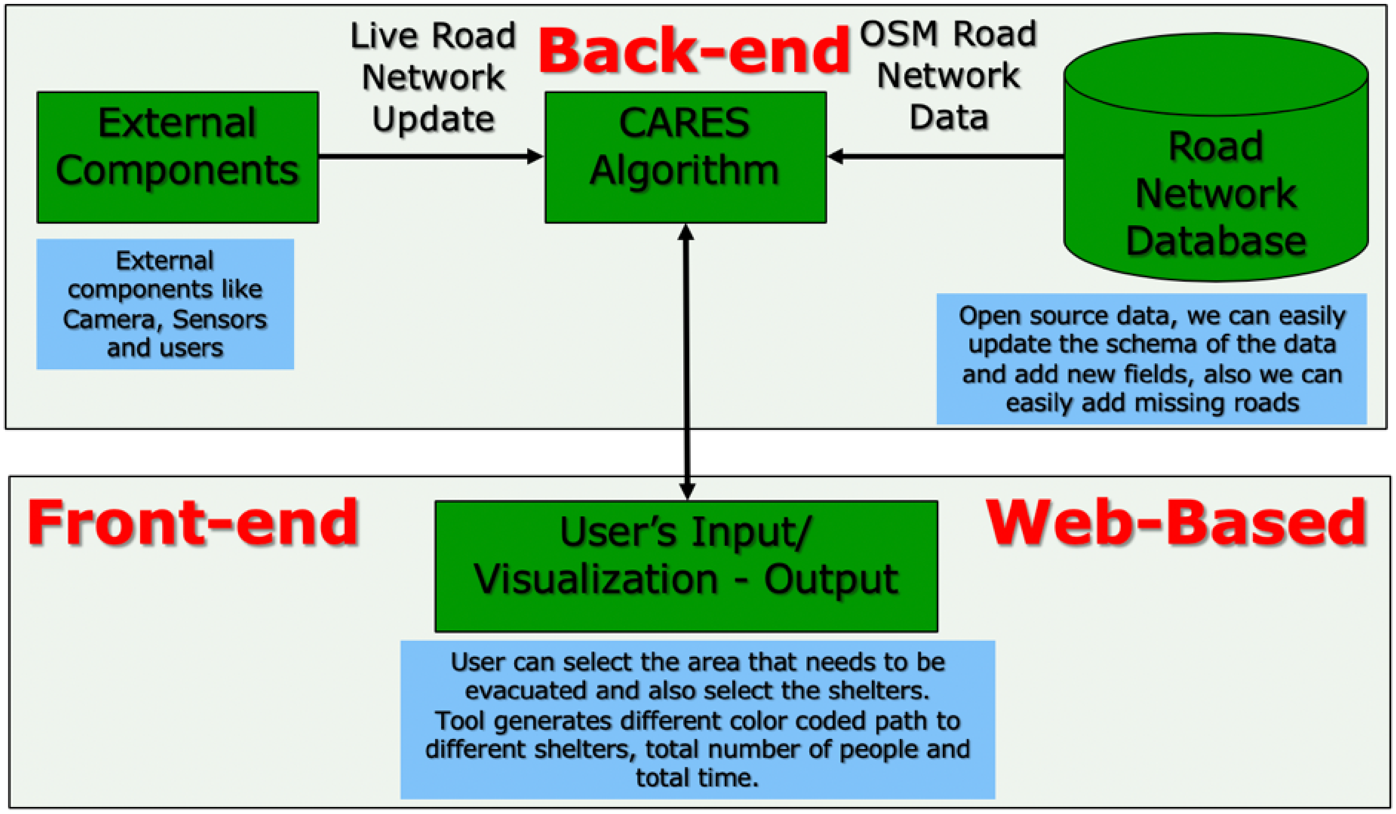

Back-end
External Components
Live Road Network Update
CARES Algorithm
OSM Road Network Data
Road Network Database
External components like Camera, Sensors and users
Open source data, we can easily update the schema of the data and add new fields, also we can easily add missing roads
Front-end
User's Input/ Visualization - Output
Web-Based
User can select the area that needs to be evacuated and also select the shelters. Tool generates different color coded path to different shelters, total number of people and total time.

Fig. 1. Overall architecture of the proposed system

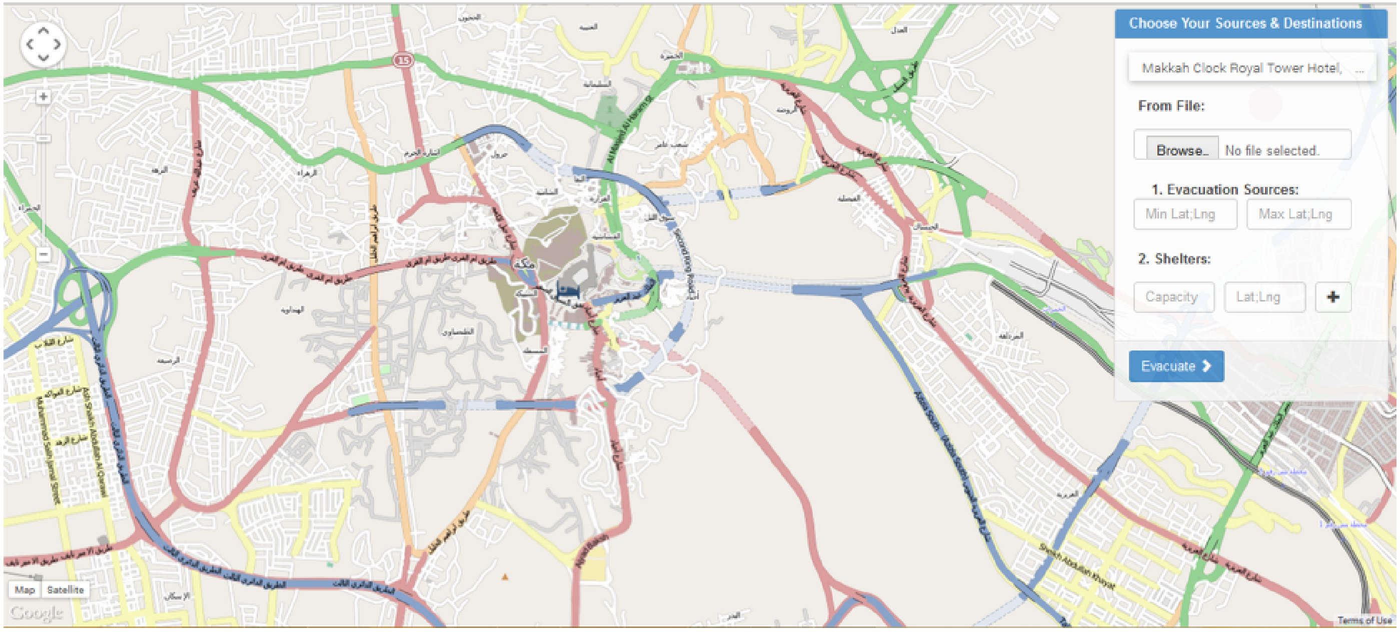
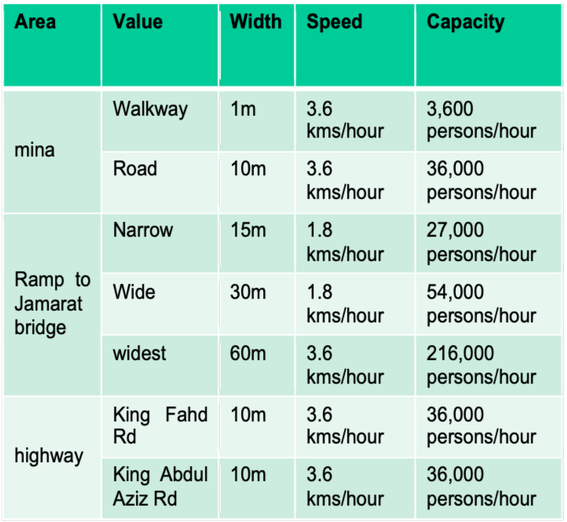


Fig. 2. Visual analytics interface of the system.



Area	Value	Width	Speed	Capacity
mina	Walkway	1m	3.6 kms/hour	3,600 persons/hour
	Road	10m	3.6 kms/hour	36,000 persons/hour
Ramp to Jamarat bridge	Narrow	15m	1.8 kms/hour	27,000 persons/hour
	Wide	30m	1.8 kms/hour	54,000 persons/hour
	widest	60m	3.6 kms/hour	216,000 persons/hour
highway	King Fahd Rd	10m	3.6 kms/hour	36,000 persons/hour
	King Abdul Aziz Rd	10m	3.6 kms/hour	36,000 persons/hour

Fig. 3. Speed and Capacity.

A. Database Generation

1) Fetching Data from OSM: The system fetches the data of any city under consideration (city of Makkah, Saudi Arabia for the case study) from Open Street Map (OSM). The OSM provides the data in a PBF/XML format that requires preprocessing to convert the data into required format along with the cleaning. TAREEG [14] uses OpenStreetMap data set and develops a web-service that extracts spatial data from any part of the world easily using MapReduce-based techniques. TAREEG is scalable and removes the irrelevant noisy data and allows the user to download the data in multiple format. In the proposed system, the same approach of TAREEG has been used to extract the data of OSM dataset in edges and nodes relational format by selecting the area on the map as shown in Figure 4.

2) Converging of Data into Road Network: OSM provides the data in XML/PBF format. By using OSM2pgSQL tool, the backend system converts the OSM into a graph model, G = (V,E), where V represent the set of vertices and E is the set of edges. The tool is helpful to extract road relation database of the selected area.

3) Assignment of Average Travel Time and Capacity: Subsequently tool runs a time and capacity module that provides the time and capacity values for each edge of the graph model. Table I shows different types of roads that are retrieved from Open Street Map. We assume an average speed (for all types of pedestrians on all types of roads) of 1 m/s [13]. The default capacity is assigned based on the type of the road. The system also allows user to change capacity of the road by editing the width of the road. The width of the road differ from city to city and its an important parameter that is taken into consideration while making the evacuation plan.

B. Evacuation Scenario Generation

We have successfully extracted and loaded the complete road network with all the nodes and edges from Open Street map. As mentioned earlier, Open Street map is an open source and can be easily modified by adding new column in the table such as capacity and travel time. In some cases, we need to add more nodes and create new edges to provide further detail about the routes, for example, for the area of Mina. Figure 5 shows various nodes and edges for Mina. The corresponding graph model, in terms of nodes and edges are displayed through the front-end visual interface as depicted in Table II and Table III. Capacity and travel times are added to nodes and edges, respectively based on the approach provided in [15], [16].

TABLE I. Different Types of Roads

Types of Road
Motorway
Trunk
Tertiary
Primary
Secondary
Residential
Unclassified

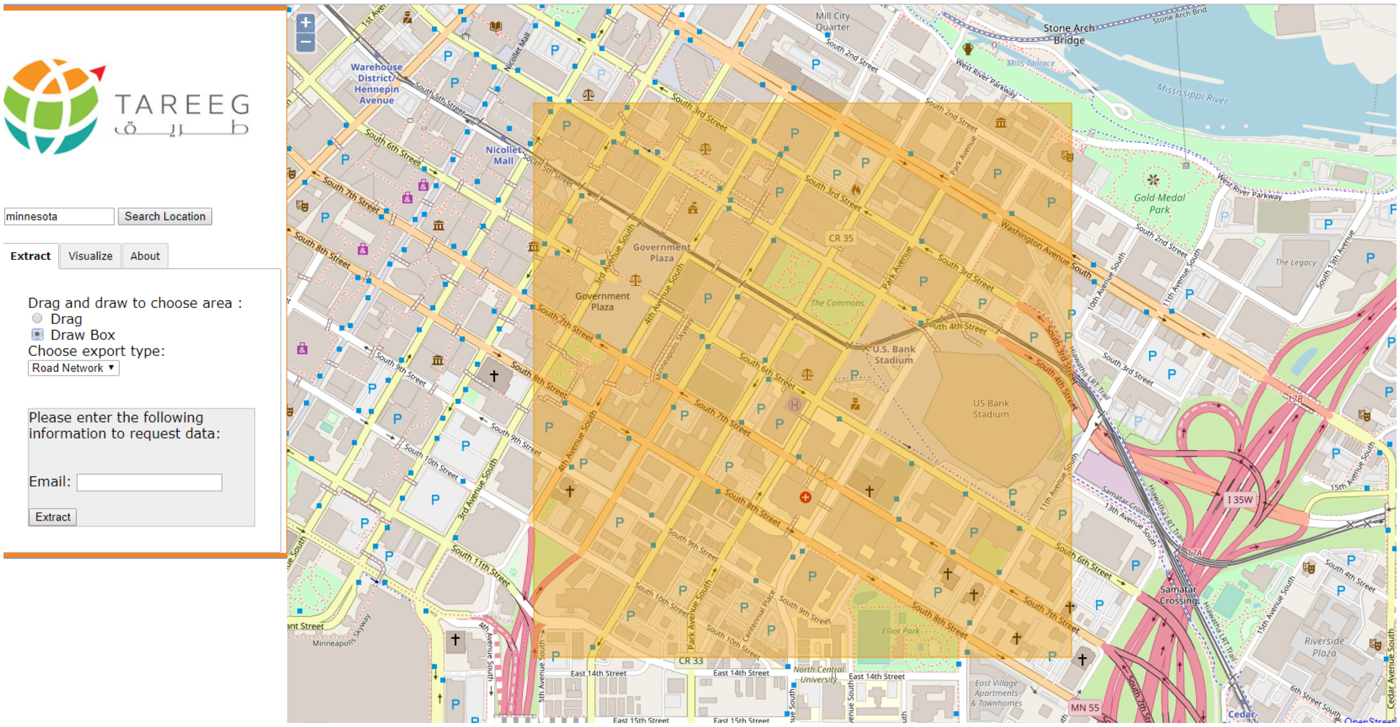
As mentioned earlier, the tool provides a robust visual interface that allows the user to orchestrate an evacuation scenario encompassing several options depicting scale and the location of evacuation area, along with various other parameters. Through the following demo illustrations, we provide the detail of this important visual capability of the tool.

C. Selection of Evacuation Area

The user can identify the area to be evacuated in form of grid by visual demarking the area as depicted in Figure 6.

D. Interactively Changing Road Conditions

The tool also provides an option to the user (or emergency management authorities acting as users) to change the road plan by closing or expanding a road (or set of arbitrarily selected roads) on the map, as an additional capability for orchestrating a scenario.



minnesota [Search Location]

Extract | Visualize | About

Drag and draw to choose area :
○ Drag
● Draw Box
Choose export type:
Road Network ▾

Please enter the following information to request data:

Email:

[Extract]

Fig. 4. Generic Tool to extract OpenStreetMap [14].

TABLE II. Example of the Data for Node Table

id	name	lat	lon	isShelter	capacity
1	1114915162	21.414340	39.862888	FALSE	0
2	1123952383	21.414178	39.858879	FALSE	0
3	1123952396	21.417453	39.859701	FALSE	0
4	1123952398	21.415552	39.859227	TRUE	2000
5	1123952400	21.415031	39.858813	FALSE	0

TABLE III. Example of the Data for Edge Table

id	name	sid	did	roadtype	time	capacity
19627	304911020	1471778539	1471741582	”highway”=”residential”	72	144
19630	380891295	1471741582	1471741580	”highway”=”residential”	195	390
19631	380908949	1471741581	1471741579	”highway”=”residential”	74	148
19744	380891308	1471741600	1471733686	”highway”=”residential”	185	370
19750	380874956	1471741587	1471733682	”highway”=”residential”	18	36

VI. Visualizing and Assessing Route Planning and Evacuation Performance

The back-end algorithm in the tool runs in such a way that it avoids criss-crossings, which in turn helps the crowd to avoid another disaster or congestion. For example: if the evacuation output is from source ‘s’ to destination ‘d’ passing through multiple nodes N = n1, n2, n3 .... n then all the evacuees on intermediate nodes ‘N’ will move for evacuation towards destination ‘d’ only [13]. The results presented in this section illustrate how the performance of the tool, measured in terms of rate of evacuation, i.e. the number of evacuates arriving at the designated sheet(s) as a function of time. In particular, the effect of the following user’s selected scenarios on the performance is presented. These visual outputs can be simultaneously and synchronously shared among distributed end-users.

This system gives the ability to decision-makers to assess different scenarios of evacuation by visualizing the change in routes and evacuation time. The decision-makers leverages the ability of the interactive system to visualise the effect of distinct scenarios such as effect of blocking or expanding the roads, effect of spreading and clustering the given number of shelters, effect of increasing the number of shelters and evacuees.

A. Effect of Blocking Roads or Expanding the Capacity of Roads

To identify the impact of closing or expanding the road is very critical. The tool is helpful to identify the impact of the road while evacuating the people with respect to time. Results displayed in Figure 7 show the effect on the overall evacuation time after closing the road. In particular, graph on the right

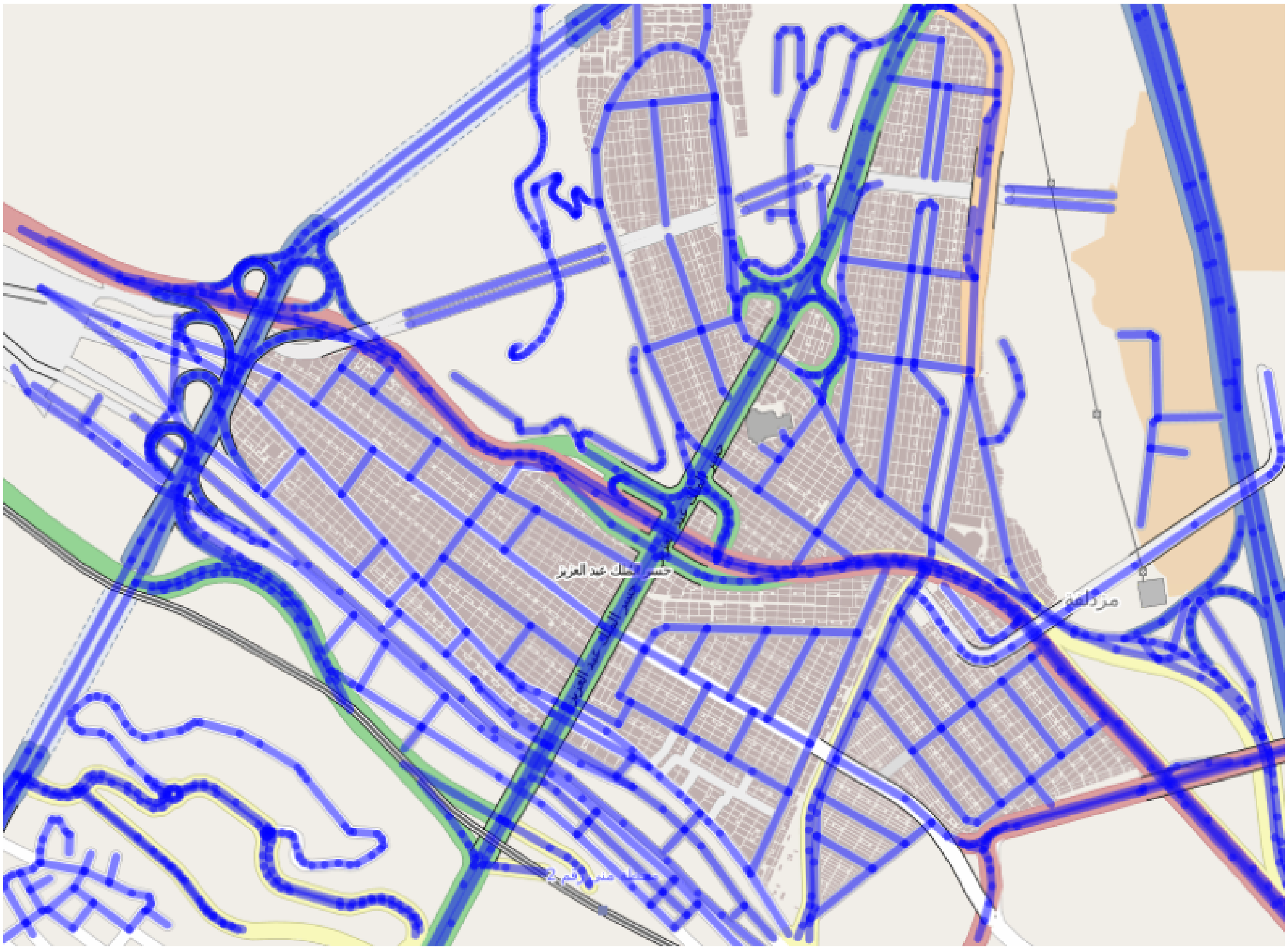
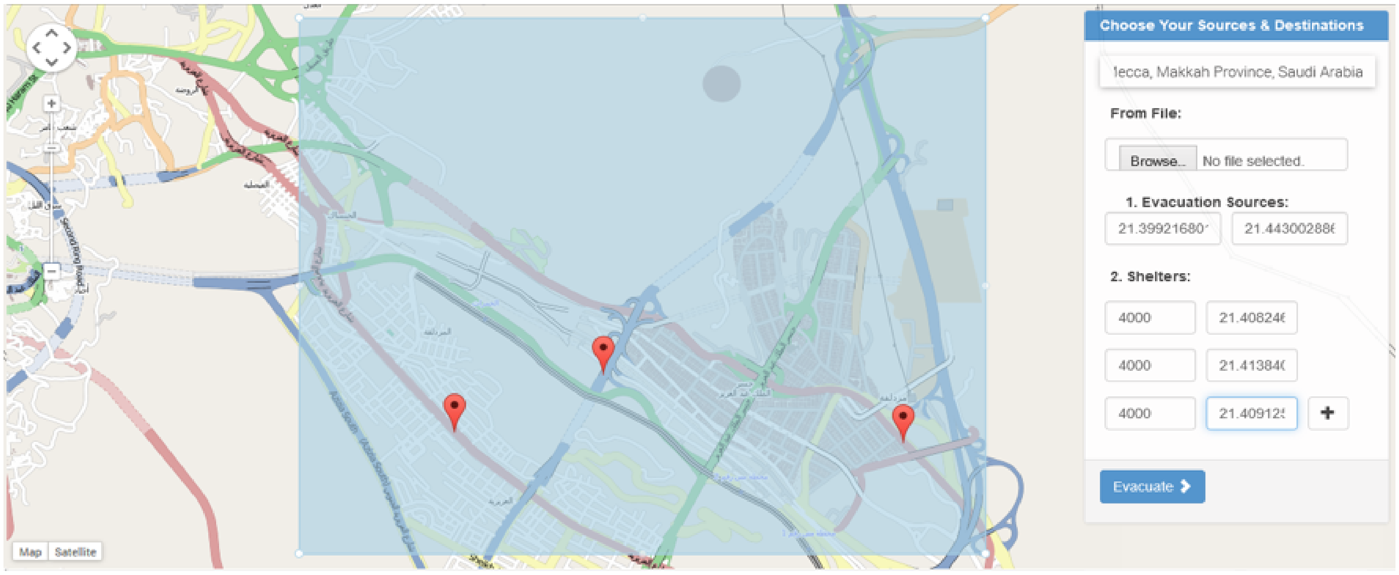


Fig. 5. Detailed Route Map for Mina.



Choose Your Sources & Destinations

ǀecca, Makkah Province, Saudi Arabia

From File:

Browse... No file selected.

1. Evacuation Sources:

21.39921680' 21.44300288€

2. Shelters:

4000 21.40824€

4000 21.41384(

4000 21.40912! +

Evacuate ❯

Fig. 6. Grid selection to demark the evacuation area with shelter points shown in red markers. Capacity is assigned to each shelter.

shows the impact on the evacuation time after closing the road. The increase in the evacuation is noticeable, as expected under this scenario. By closing the major highway the evacuation time almost doubled from 4435 to 9149 seconds. The system allows decision-makers to take the decision by blocking or expanding the roads and visualize the effect on the evacuation time and routes.

B. Effect of Increasing the Number of Evacuees

The tool can be used by the end-user for evacuating different size of population. The output of the tool (exemplified in Figure 8) can provide an assessment whether or not the given population can be evacuated within a certain time frame, with the given size and capacity of the road network and shelters. Through such analysis, decision can be made about the limit on the size of the population that can be accommodated in the region on interest.

C. Effect of Spreading/Clustering a Given Number of Shelters

By running multiple scenarios, the tool can depict the difference in the evacuation time as the shelters are clustered or spread. Figure 9 shows the output with clustered shelters. Figure 9 and Figure 10 show the timing output for the case when the shelters are clustered close together vs. spreading them out. This allows decision-makers to identify the optimum location of shelter points to evacuate desired number of evacuees in a quick and an efficient manner.

D. Effect of Increasing the Number of Shelters

Figure 11 shows the impact of increasing the number of shelters on the evacuation time. Graph on the right shows the output with 5 shelters and graph on the left with 3 shelters shows the evacuation time 722 seconds and 943 seconds respectively to evacuate 23810 evacuees. Such type of assessment can enable the end-user to take the decision on the number of shelters to be allotted for a given scenario.

VII. Conclusion

In this paper we propose a system for route planning and emergency crowd evacuation. We describe in detail the various capabilities of the tool, such as: scalability of route generation, visual assessment of various scenarios, along with demo outputs of Makkah, Saudi Arabia and Minnesota, USA. The proposed system utilizes Spatial database of the OpenStreetMaps and allow users to execute evacuation wherever dataset of the OpenStreetMaps are available. The capabilities of this tool can be expanded in various dimensions. For example, for real-time applications, it can be integrated with multi-modality data streams, such as live camera feeds, mobile cell data etc. In addition, the tool can be integrated with existing crowd and traffic management tools currently available.

VIII. Acknowledgment

This research was funded by the Center of Research Excellence in Transportation and Crowd Management, Umm Al-Qura University, Saudi Arabia.

References

[1] Christian Artigues, Emmanuel Hébrard, Yannick Pencolé, Andreas Schutt, Peter Stuckey. A Study of Evacuation Planning for Wildfires. The Seventeenth International Workshop on Constraint Modelling and Reformulation (ModRef 2018), Aug 2018, Lille, France. 17p.

[2] H.W. Hamacher and S.A. Tjandra. Mathematical modelling of evacuation problems - a state of the art. Pedestrian and Evacuation Dynamics, pages 227-266, 2002. Springer Verlag.

[3] Madkour, Amgad, Walid G. Aref, Faizan Ur Rehman, Mohamed Abdur Rahman and Saleh M. Basalamah. A Survey of Shortest-Path Algorithms. CoRR, abs/1705.02044, 2017

[4] Zwick, Uri. ”Exact and approximate distances in graphs - a survey.” In European Symposium on Algorithms, pp. 33-48. Springer, Berlin, Heidelberg, 2001.

[5] Sommer, Christian. ”Shortest-path queries in static networks.” ACM Computing Surveys (CSUR) 46, no. 4 (2014): 45.

[6] Chang Liu, Zhan-li Mao, Zhi-min Fu, Emergency Evacuation Model and Algorithm in the Building with Several Exits, Procedia Engineering, Volume 135, 2016, Pages 12-18, ISSN 1877-7058.

[7] Pu, Shi, and Sisi Zlatanova. ”Evacuation route calculation of inner buildings.” In Geo-information for disaster management, pp. 1143-1161. Springer, Berlin, Heidelberg, 2005.

[8] Pursals, Salvador Casadesús, and Federico Garriga Garzón. ”Optimal building evacuation time considering evacuation routes.” European Journal of Operational Research 192, no. 2 (2009): 692-699.

[9] Barnes, Matthew and Leather, Hugh and Arvind, D. (2007). Emergency Evacuation using Wireless Sensor Networks. Proceedings - Conference on Local Computer Networks, LCN. 851-857. 10.1109/LCN.2007.48.

[10] Shekhar, Shashi, KwangSoo Yang, Venkata MV Gunturi, Lydia Manikonda, Dev Oliver, Xun Zhou, Betsy George, Sangho Kim, Jeffrey MR Wolff, and Qingsong Lu. ”Experiences with evacuation route planning algorithms.” International Journal of Geographical Information Science 26, no. 12 (2012): 2253-2265.

[11] Zhou, Xun, Betsy George, Sangho Kim, Jeffrey MR Wolff, Qingsong Lu, Shashi Shekhar, and O. Nashua. ”Evacuation Planning: A Spatial Network Database Approach.” IEEE Data Eng. Bull. 33, no. 2 (2010): 26-31.

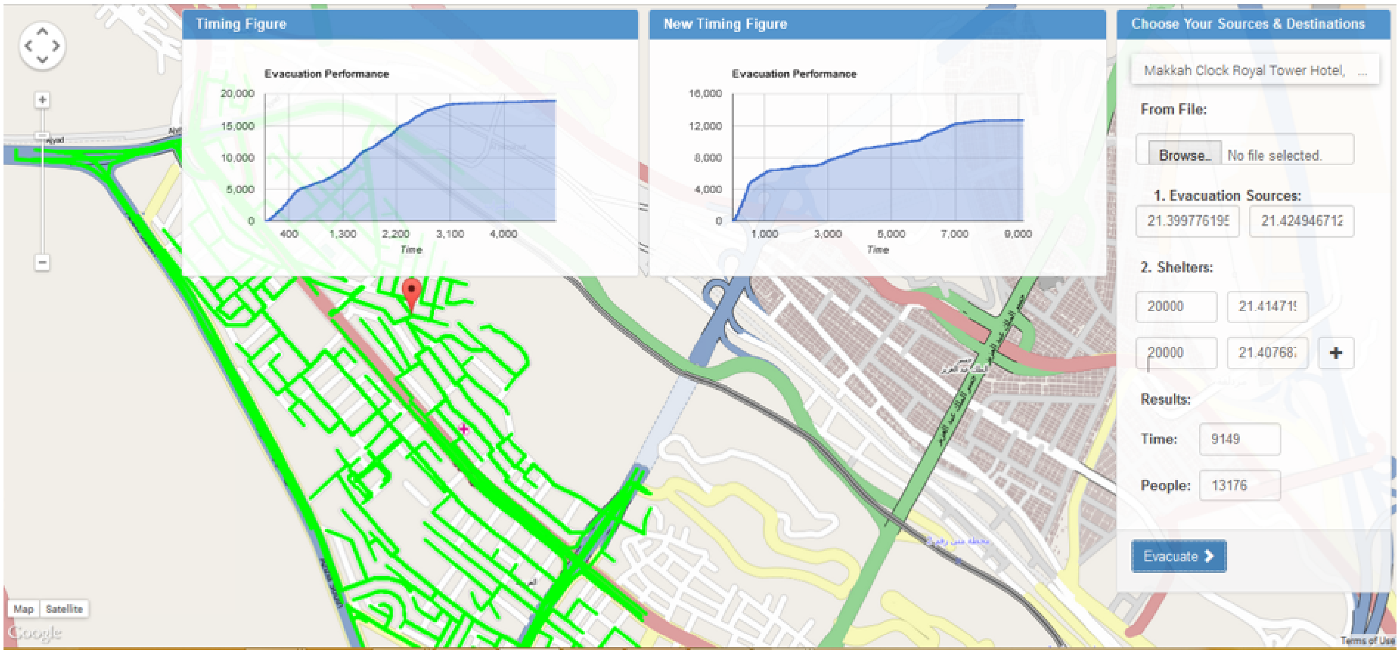
[12] K. Yang et al., ”Intelligent Shelter Allotment for Emergency Evacuation Planning: A Case Study of Makkah.” HajjCore tech. report P1104-T1, 2012; http://docs.lib.purdue.edu/cctech/9

[13] Yang, KwangSoo, Apurv Hirsh Shekhar, Faizan Ur Rehman, Hatim Lahza, Saleh Basalamah, Shashi Shekhar, Imtiaz Ahmed, and Arif Ghafoor. ”Intelligent shelter allotment for emergency evacuation planning: A case study of makkah.” IEEE Intelligent Systems 30, no. 5 (2015): 66-76.

[14] Louai Alarabi, Ahmed Eldawy, Rami Alghamdi, and Mohamed F. Mokbel. 2014. TAREEG: a MapReduce-based system for extracting spatial data from OpenStreetMap. In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL ’14). ACM, New York, NY, USA, 83-92. DOI: http://dx.doi.org/10.1145/2666310.2666403

[15] D. Helbing, A. Johansson and H. Z. Al-Abideen, The Dynamics of Crowd Disasters: An Empirical Study. Physical Review E 75: 046109, 2007. (www.trafficforum.ethz.ch/crowdturbulence/)

[16] R.L. Hughes, ”The Flow of Human Crowds”, Annual Review of Fluid Mechanics, 35, 169-182 (2003)



Choose Your Sources & Destinations

Makkah Clock Royal Tower Hotel, ...

From File:

Browse... No file selected.

1. Evacuation Sources:

21.399776195 | 21.424946712

2. Shelters:

20000 | 21.41471

20000 | 21.40768 | +

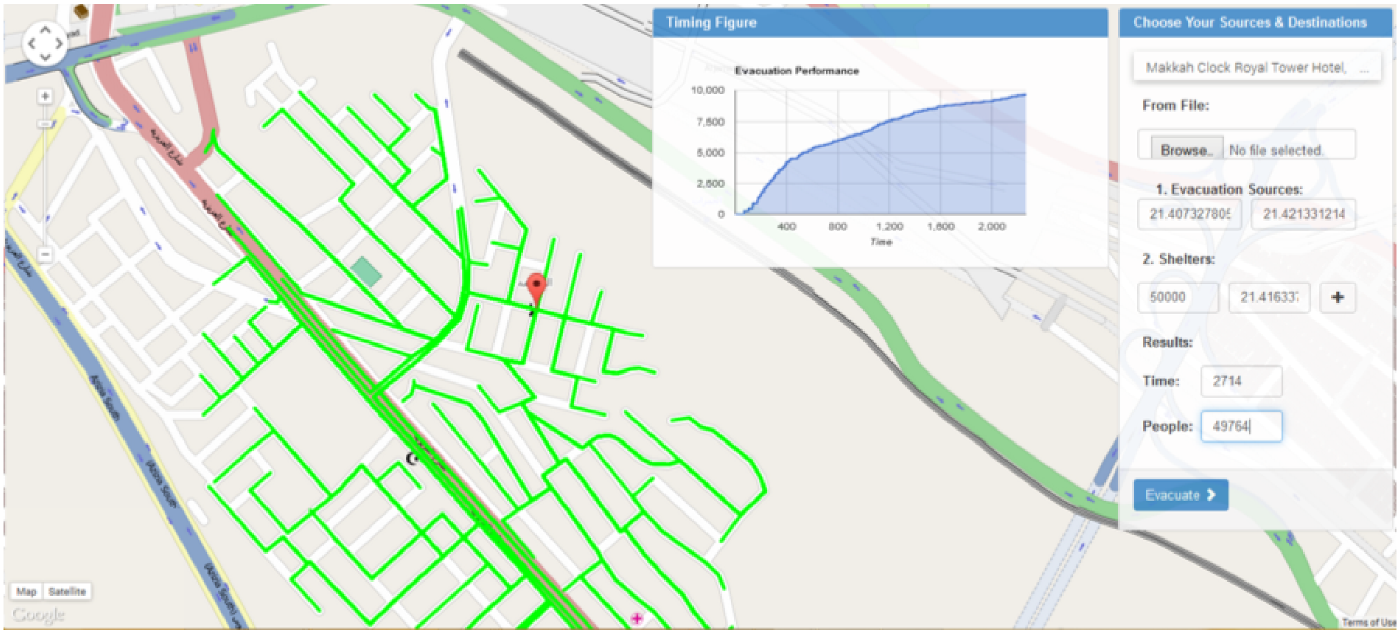
Results:

Time: 9149

People: 13176

Evacuate ❯

Fig. 7. Effect of blocking/expanding roads on evacuation time. Roads with red color are marked as closed.



Choose Your Sources & Destinations

Makkah Clock Royal Tower Hotel, ...

From File:

Browse... No file selected.

1. Evacuation Sources:

21.407327805 | 21.421331214

2. Shelters:

50000 | 21.41633 | +

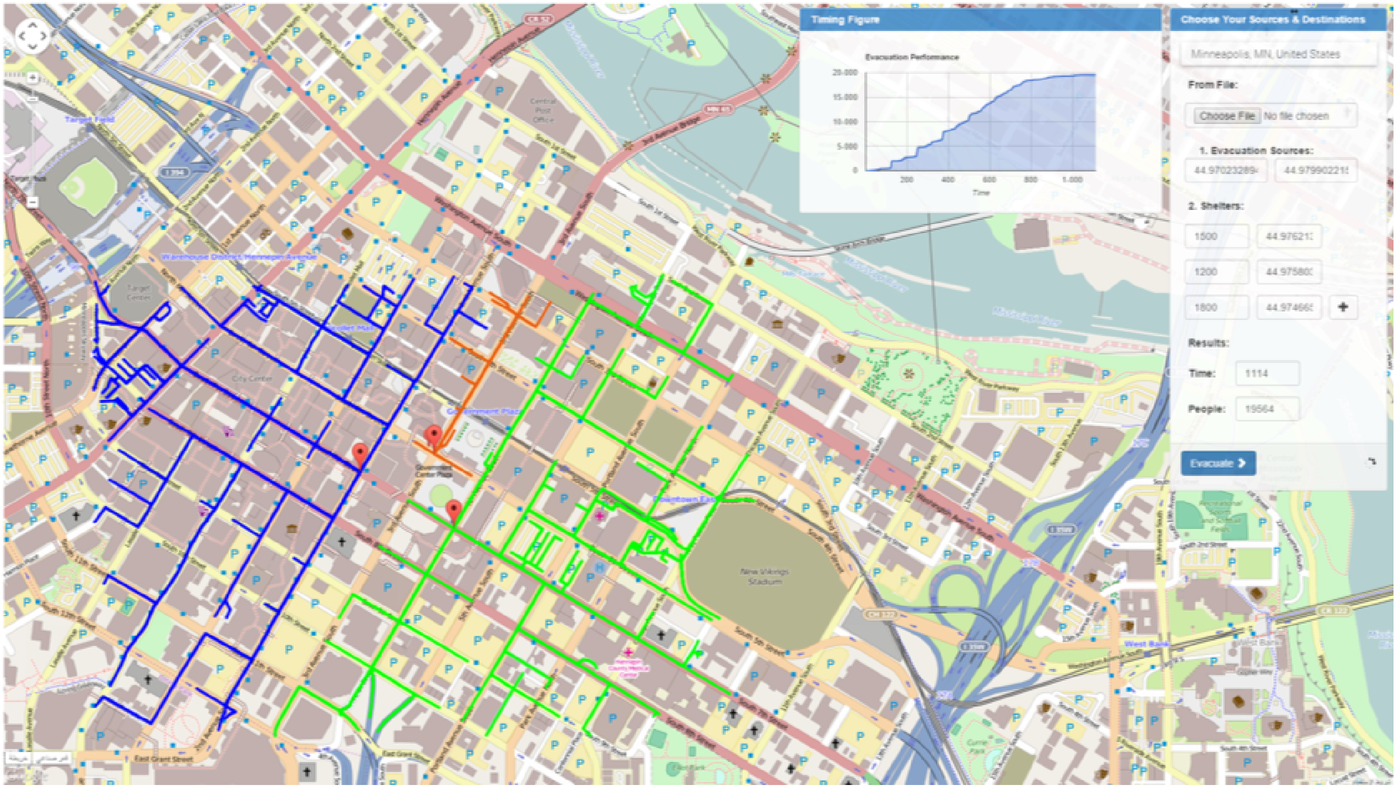
Results:

Time: 2714

People: 49764

Evacuate ❯

Fig. 8. Effect of increasing the number of evacuees.



Minneapolis, MN, United States

From File: No file chosen

1. Evacuation Sources: 44.97023289 44.97990221

2. Shelters:

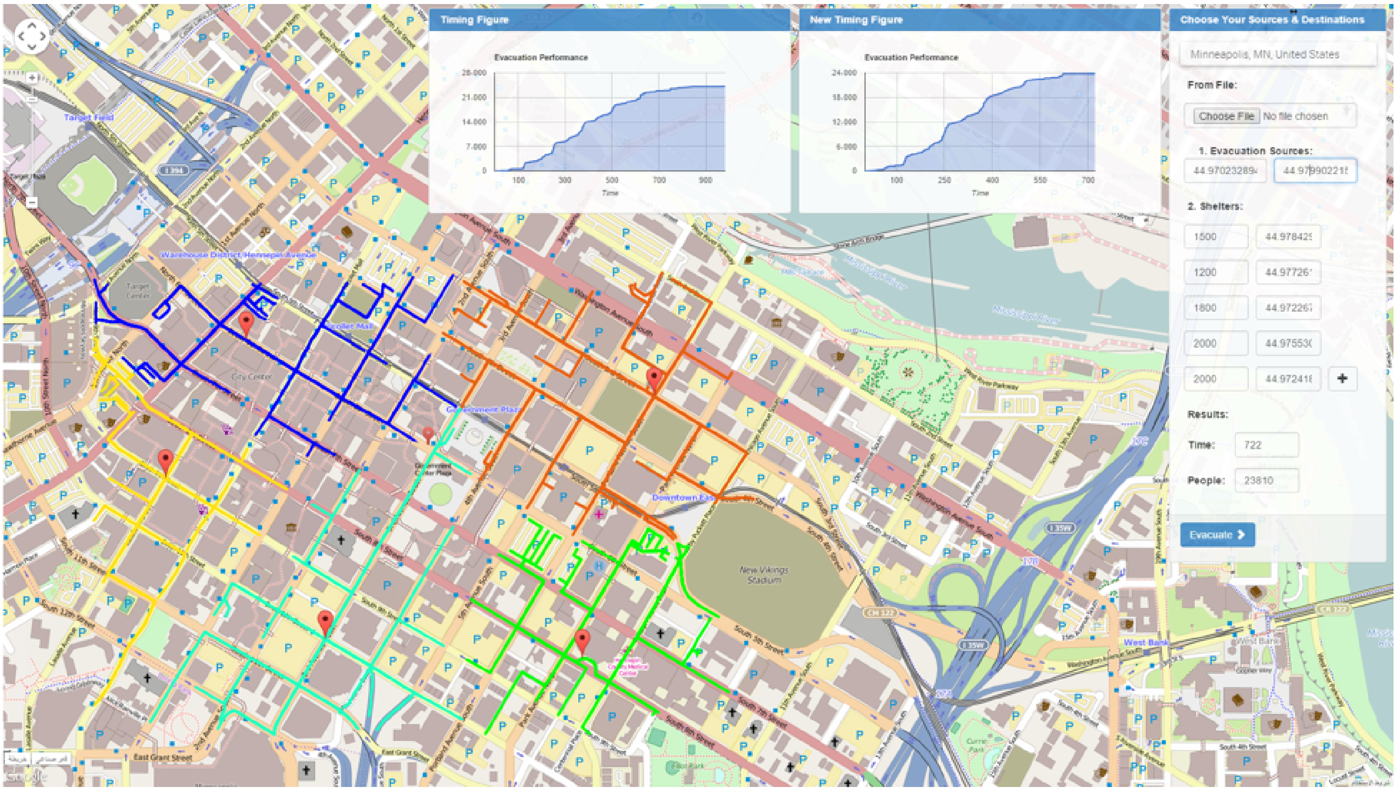
1500	44.97621
1200	44.97580
1800	44.97466

Results:

Time: 1114

People: 19564

Fig. 9. Effect of clustering shelters.



Minneapolis, MN, United States

From File: No file chosen

1. Evacuation Sources: 44.97023289 44.97990221

2. Shelters:

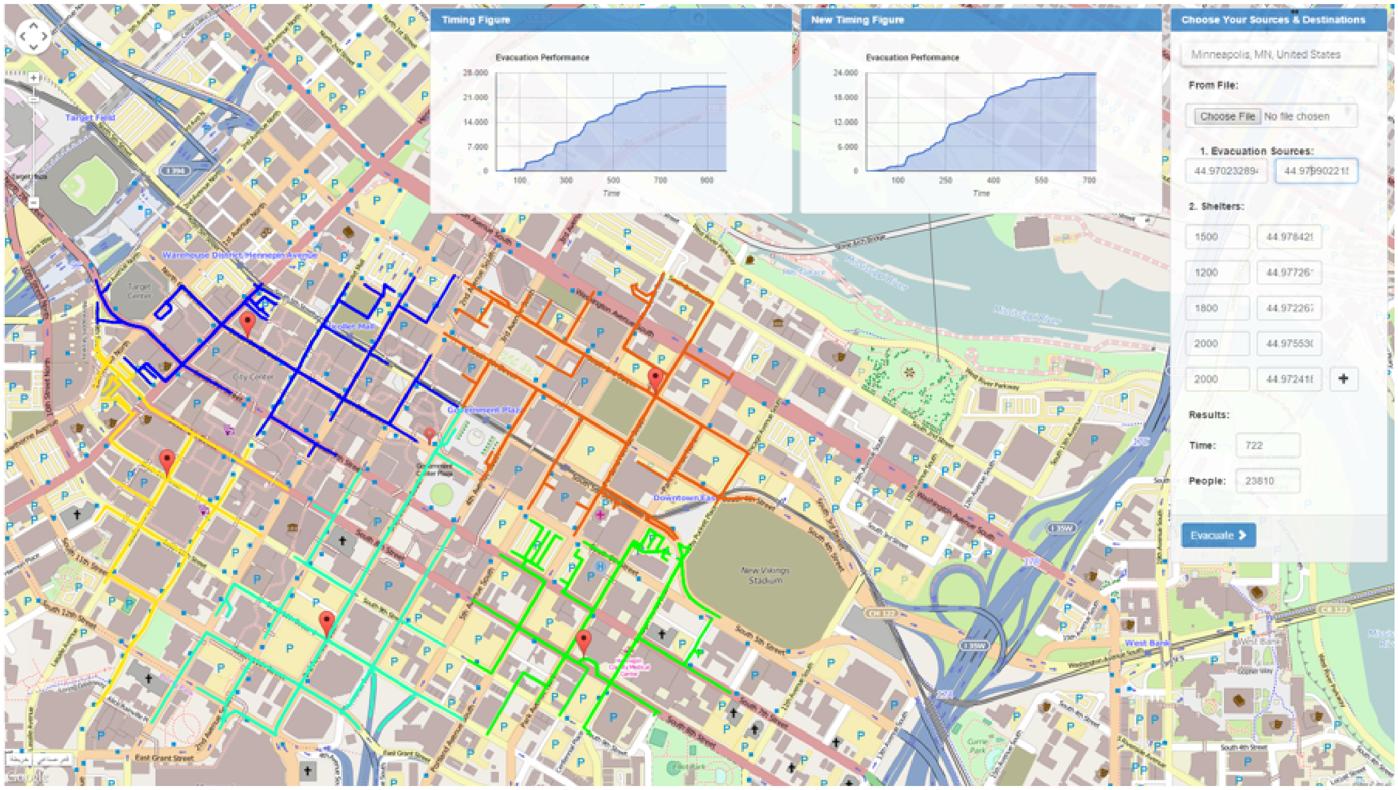
1500	44.97842
1200	44.97726
1800	44.97226
2000	44.97553
2000	44.97241

Results:

Time: 722

People: 23810

Fig. 10. Effect of spreading shelters.



Choose Your Sources & Destinations

Minneapolis, MN, United States

From File:

No file chosen

1. Evacuation Sources:

44.97023289 | 44.97990221

2. Shelters:

1500	44.97842
1200	44.97726
1800	44.97226
2000	44.97553
2000	44.97241

Results:

Time: 722

People: 23810

Fig. 11. Effect of increasing the number of shelters.

Detecting Video Surveillance Using VGG19 Convolutional Neural Networks

Umair Muneer butt[1]
Department of Computer Sciences
University Sains Malysia
University of Lahore

Sukumar Letchmunan[2]
Department of Computer Sciences
University Sains Malysia

Fadratul Hafinaz Hassan[3]
Department of Computer Sciences
University Sains Malysia

Dr. Sultan Zia[4]
Department of Computer Sciences
University of Lahore
Chenab Campus, Pakistan

Anees Baqir[5]
Faculty of Computing & IT
University of Sialkot
Sialkot, Pakistan

***Abstract*—The meteoric growth of data over the internet from the last few years has created a challenge of mining and extracting useful patterns from a large dataset. In recent years, the growth of digital libraries and video databases makes it more challenging and important to extract useful information from raw data to prevent and detect the crimes from the database automatically. Street crime snatching and theft detection is the major challenge in video mining. The main target is to select features/objects which usually occurs at the time of snatching. The number of moving targets imitates the performance, speed and amount of motion in the anomalous video. The dataset used in this paper is Snatch 101; the videos in the dataset are further divided into frames. The frames are labelled and segmented for training. We applied the VGG19 Convolutional Neural Network architecture algorithm and extracted the features of objects and compared them with original video features and objects. The main contribution of our research is to create frames from the videos and then label the objects. The objects are selected from frames where we can detect anomalous activities. The proposed system is never used before for crime prediction, and it is computationally efficient and effective as compared to state-of-the-art systems. The proposed system outperformed with 81 % accuracy as compared to state-of-the-art systems.**

Keywords*—*Anomalous detection; surveillance video; VGG16; VGG19; ConvoNet; AlexNet

I. Introduction

As the technology is growing rapidly, the crime ratio and strategies are also advancing. One of the major crimes faced by almost all over the world is street and theft crime [1]. One of the basic countermeasures is to do surveillance, i.e. monitoring the area, which is done by the CCTV cameras, it allows the user to watch what is going on in different places, and their footage can also be accessed remotely by the number of authenticated users and agencies. However, there is no intelligent method to identify or detect a specific object or person. The manual and common approach are to watch the lengthy videos carefully from one CCTV recording to another. It is quite difficult to detect abnormal activities through this CCTV footage. The picture quality, the motion and objects were identified through CCTV cameras [2] [3].

Here, the important question arises is that how we can detect the abnormal activities before it occurs. The basic challenge is to automatically and intelligently watch the video surveillance and detect the abnormal and anomalous events in rushy areas and protect the individual(s) at the spot. There are huge limitations which makes it challenging and tough to detect an anomalous event at the spot [4]. The selection of features is very tough because, through these features, we can detect anomalous activities. The features selected are responsible for detecting the moving object and has a significant impact on the analysis of behaviour and the performance of the system [5]. Figure 1 depicts the basic structure of a surveillance network.

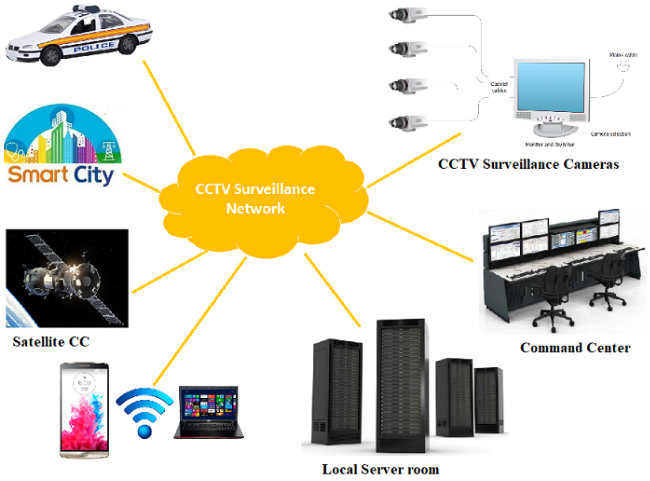


Fig. 1. CCTV Surveillance Network

Nowadays, data mining is considered the most vigorous research field. By data mining, we mean the process of mining knowledge from the raw data and discovering fascinating pattern from a huge set of data. In data mining, most work is done on heterogeneous and unstructured data, i.e. videos, images, etc. [4]. A variety of technical tools are available for detecting video surveillance. LI Yi et al. [6] uses neural network architecture for segmentation and shape estimation. To achieve optimal performance, their architecture alternate between correspondence, deformation flow and segmentation in an Inductively Coupled Plasma (ICP) like fashion. The important part is the induction algorithm, which successfully generalizes to new and unseen objects.

The most popular among all video surveillance system is a traffic surveillance system. Because the surveillance sensor and processors are available in the market at a very cheap rate and their decision-making capability is very much effective [7]. Usually, the majority of the system provides the facility of detecting motions and record the video when motions are detected; it reduces the processing and storage time of the video. It allows the users to remotely access the cameras from multiple devices and store the recorded videos in various formats. Figure 2 represents the basic structure of a traffic surveillance system.

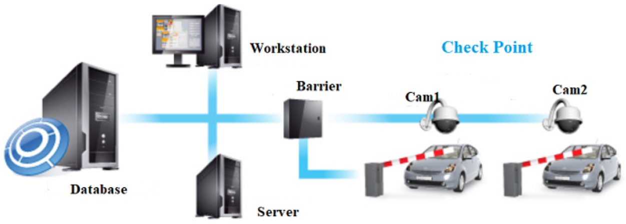


Fig. 2. Traffic Surveillance System

The main focus of this paper is detecting street crime i.e. snatching and theft via video surveillance. We mainly focuses on the object which usually occurs at the time of snatching [8] [9]. The amount of moving targets imitates the performance, speed and amount of motion in the anomalous video. The dataset used in this paper is Snatch 1.0[1], the video in the datasets are further divided into frames and from these frames, features are extracted. For this purpose, VGG19 algorithm has been employed in this paper and the results were found which compares the features from their original video [7]. The understand that the proposed method outperforms the state-of-the-art techniques, comparative analysis is performed and it is concluded based on the results that the proposed method outperformed them.

Moreover, the paper is organized as follows. Section II emphasis on the literature, while section III of the paper explained the methodology. In section IV, we present the results and outcomes of the research. Section V concludes the paper with future work.

II. Literature Review

Video surveillance is an essential part of our society to foresee criminal activities. Numerous efforts have been made in this area, but efficiency is still a big challenge as shwon in Table I. In [5] Appearance and Motion DeepNet (AMDN) based method is used to find out more Stacked Denoising Auto Encoders (SDAE) active video scene appearance and presentation of the motion. This new method of unsupervised learning is based on the depth study of anomalies architecture of video detection. This method examines the appearance, features and joint representation. There is an extensive experimental evaluation, taking into account three complex sets of social video data anomaly detection of the train, UC San Diego(UCSD) and subway, and demonstrated that the proposed method is reliable and effective. To detect additional unusual events using co-occurrence of more than one pattern, the AMDN method is beneficial.

The sparse method of representation is widely used in abnormal population detection; specifically, they represent dimensional movements. In [11], they proposed a method for detecting abnormal crowds. The proposed method includes two deep replacement processes, each of which uses a dynamic daily updated dictionary. Dynamic data update dictionary process dynamically adds normal test procedures to the dictionary while other dictionaries with exception samples are also resolved. The proposed model offers a wider vocabulary about normal and anomalous events. Additionally, abnormal events are more accurate than prior art methods. The results calculated from the experimental datasets depicted higher accuracy achieved by the proposed method as compared with the latest methods in local and global anomaly detection.

Dinesh et al. [12] use deep learning through bidirectional Long Short Term Memory (LSTM) for real-time violence detection in football stadiums. They used real-time video streams processed by the Spark framework along with Histogram of Oriented Gradients (HOG) to separate frames and features of each frame. Features are then labelled based on violence model and used for training to Bidirectional Long short term memory networks (BDLSTM) to recognize violent activities in the scenes. They validated this model with 94% accuracy in the detection of violent actions.

The video surveillance camera is increasingly challenging the video control system. The monitoring center needs monitoring tools to drive. Intelligent video perimeter protection solutions have to select and display cameras with evidence of these events, but background-based modelling systems only focus on the problem, whether or not an intrusion occurs. In [13], the authors recommended that you add a module based on machine learning and global functionality to adapt the video surveillance solution to identify problematic situations and provide the best priority. Instead of improving the robustness of a virtually impossible environment, the authors propose a way to solve problematic events based on global features.

The employment of surveillance cameras is increasing indoors and outdoors; it requires system intelligent enough to detect anomalous activities. Authors in this paper [14] investigated the most popular methods of extraction and description methods and presented an overview of the behaviour modelling classification methods and frames. Additionally, the authors presented a dataset and metric evaluation challenges for video systems. Finally, they introduced some intelligent real-world video systems.

Anpei et al. [15] introduced a novel algorithm based on the neural network called ”Deep Surface Light Field” or DSLF for moderate sampling. Leveraging different patterns of sampling, DSLF fills in the missing data. They also addressed image registration. Aniqa et al. [16] proposed a framework which works by extracting the visual-based features from the frames of video by employing ”Convolutional Neural Networks” (CNN). Furthermore, the framework passed the derived representations to the LSTM model. For the natural language description of

[1] https://sites.google.com/view/debadityaroy/datasets

TABLE I. A BRIEF COMPARISON OF THE TECHNIQUES TO DETECT VIDEO SURVEILLANCE

Reference	Techniques	Dataset	Description
[5]	Appearance and Motion DeepNet, SVM models	UCSD pedestrian dataset, consists of two subsets: Ped1 and Ped2	Depicted better performance as compared to existing methods. The basic advantage their approach was that it did not depend on any piror knowledge to design the features
[4]	FCN architecture	UCSD (Ucsd anomaly detection dataset, 2017) and Subway Benchmarks	The proposed method helped run a deep learning-based method at a speed of about 370 fps
[2]	SVM Model, Random Forest	Violent Flows- 246 videos, half of them being violent and other half being non violent collected from Youtube with 320*240 pixels resolution	To overcome the abnormal behaviors and limitations mentioned in the paper, the system proposed by the authors used huge amount of training data including all possible scenarios
[10]	Locality Sensitive Hashing Filters,Particle Swarm Optimization	UMN dataset consists of three different crowded scenes with a frame number of 1453, 4143, and 2143.	The abnormality degree of a new test sample is estimated by calculating the filter response of the test sample to its nearest filter. It was concluded that the proposed method is effective and robust
[8]	kernel support vector machine, binary support vector machine with graph kernel	UCSDped2 dataset, which contained that scenes of movements by pedestrian parallel to the plane of camera	Graph is used to represent the interaction and the co-relation of the motions of objects/entities. By using the graph kernel, to measure the similarity between two graphs provides robustness to slight deformations to the topological structures due to presence of noise in data

video frames, they used a fine-tuned CNN model.

For this reason, the scene transformations are sensitive to the characteristics that are robust and change the appearance of the object to attract the correlation. The selection of functions used to characterize floating objects is a lightweight job because it has a great impact on the description and analysis of behavior. In this review [14], different levels of the system of video surveillance were analyzed by the authors, which resulted in a behavioural representation and behavioural pattern.

In [17], the authors suggested another technique to classify suspicious events in video observation based on locality-sensitive hashing filters. Training tests are hashed into a rundown of pails, and each bucket's middle and radius is found to create location-sensitive hashing filters. The deviation from the normal level of another test is measured by the test's filter reaction to its closest filter. With new expectations, the locality-sensitive hashing filter is refreshed online. Test results demonstrate the adequacy and power of the proposed methodology on three datasets.

In this paper [18], they proposed a methodology for anomalous movement acknowledgment dependent on chart definition of video activists and graph kernel support vector machine. As a graph of geometric relationships between space-time intrigue points, the connection of the substances in a video is detailed. The graph vertices are spatio-worldly intrigue graph, and an edge represents the connection between the appearance and the elements surrounding the points of intrigue. For this, the chart details the improvements in video activities over the issue of detecting anomalies into a graph classification problem. Using the chart kernel to approximate the resemblance between two graphs gives the topological structure's power to minor mishappenings due to the nature of information noise.

An anomalous event detection technique has been proposed in [19], which was dependent on the unsupervised deep neural network. In particular, successful highlights of video events are thus omitted from 3D slopes to reflect both the appearance and the hint of movement. They use a deep Gaussian mixture model to lean ordinary event designs, which typically perform violent execution using a few parameters. Examinations on two open datasets indicate particular upgrades when compared with state of the art algorithms.

In this study [10], they introduced a structure for snatch theft detection in unconstrained videos utilizing activity credit demonstrating to take in all the activity traits in the snatch robberies, a huge Gaussian mixture model (GMM) called all universal attribute model (UAM) was prepared to utilize existing video datasets of human activities. For development, the authors presented a dataset called snatch 1.0 that contains snatch robberies in surveillance videos. It was demonstrated that activity vector pro video better discriminate portrayal for snatch robbery.

Wenqing et al. [20] presented a novel unsupervised deep feature learning algorithm for anomalous event detection. To fully utilize the Spatio-temporal information, the proposed system used a deep three-dimensional convolutional neural network for feature extraction. To train the C3D network without any category labels, they used a sparse coding result of handcrafted features. The proposed system outperforms the state-of-the-art systems. Schuchao et al. [21] proposed a deep learning-based technique for tracking visual objects. They used CNN to rank the patches of the target objects based on how well it is centred. The promising patch is selected by the AlexNet framework using his matching function based on deep features.

Asghar et al. [22] introduced a novel algorithm for high-level feature extraction and used those features for classification and re-identification. Their proposed method is a two-tier approach. Firstly, they extract low-level features for identification and later use high-level features for classification and re-identification. In the end, they used a deep belief network to build a model based on the low and high-level features. Yonglong et al. [23] proposed a radio frequency-based fall monitoring system based on CNN. They introduce Aryokee, which is based on radio frequency to detect fall using CNN. The key idea behind this is to separate different sources of motion, which resulted in increased robustness. They achieved 94% recall and 92% precision in detecting falls.

Umair et al. [24] used a combination of HoG, and LBP features to extract features from the American Sign Language dataset (ASL). They used those features in an auto model feature of Rapid Miner and Weka software to train and test. Rapid Miner auto model performed with 99 percent accuracy. Debaditya et al. [25] proposed employing a GMM model on snatch thefts with a large number of attribute mixtures known as the universal attribute model. They used large human action data set UCF101 and HMDB51 to train the proposed model. They used factor analysis for low-level feature representation and evaluation; they used Snatch 1.0 data set. The proposed system performed well as compare to state-of-the-art systems.

III. Proposed Methodology

In this section, we present the detail of the dataset and pre-processing applied to it to enable it for further calculations. Moreover, proposed VGG19 convolutional neural network is discussed which is used for retrieving results as shown in figure 3.

A. Dataset Description

There is a large volume of datasets available for human activity recognition and object detection. But in the existing literature, there are no proper databases for street theft and motorbike theft. Hence, we have found a dataset named Snatch 1.0, consists of normal videos and snatching videos [10]. This dataset shows the normal behaviour of objects in a different place, e.g. roads, streets, markets etc. As well as the dataset depicts abnormal and anomalous behaviour of objects while snatching, for example, the position of the vehicle, body movement, facial expressions of snatcher and victim behaviour during the snatching. Few glimpses of the dataset with the aforementioned features are depicted in figure 4.

B. Convert Videos into Image Frames

The first step in detecting anomalous activities is to divide our videos into images. In this paper, we have used the data from more than 21 videos and generate the frames from these videos, which are more than 1000 images. The frames are generated by using MATLAB R2018b. These frames show different cases and behaviours before, during and after snatching, as shown in Figure 4. The following are the steps we performed while generating frames from videos.

1) Create a directory in Matlab R2018b and copied all the videos in the directory.
2) Write the commands mentioned in algorithm 1 in the Matlab R2018b and run to generate frames.

Algorithm 1: Frame generation from Videos

```
shuttleVideo = VideoReader('17.mp4');
workingDirectory = tempname;
mkdir(workingDirectory) ;
mkdir(workingDirectory,'images') ;
ii = 1;
while hasFrame(shuttleVideo) do
    image = readFrame(shuttleVideo);
    file_name = [sprintf('03d',ii) '.jpg'];
    full_name =
      (workingDirectory,'images',file_name);
    imwrite(image,full_name) ii = ii+1;
end
```

C. Snatching Scenarios

After generating frames, the next step is to select features and object on whose bases we will detect surveillance video. Few cases, i.e. Case1 in figure 5 and Case2 in figure 6, are described in the following set of figures and explain how the snatcher snatch the chains/ wallets and what are the victim's response. After getting through these scenarios, we can select the object on which we have to mainly focus on how we will further implement our algorithm to detect the surveillance video. In the scenario depicted in Case1, there are two people on the bike who came closer to the victim and snatched her chain and ran away. From this scenario, we extract different objects, e.g. motorbike, snatcher, women, empty roads etc. These objects may be further used in classifying and labelling the frames.

D. Image Labeling

The next step is to label the object and features which we have extracted from various surveillance videos. The objects and features we selected in our paper are snatchers, the vehicle used by the snatcher, the environment in which the anomalous event occurred, the victim's behaviour before and after the snatching. To label the images, we have used MatLab R2018b Image Labeler app to label the images. The process of labelling an image in MatLab is described as follow:

1) Load all the images from the given folder as depicted in figure 7
2) Define Region of Interest (ROI) and Scene Label definitions; in Matlab, we have to label the images either in pixel region or rectangular format, which is defined in the section of ROI Label as shown in figure 8. It consists of two basic parts, one is the name of

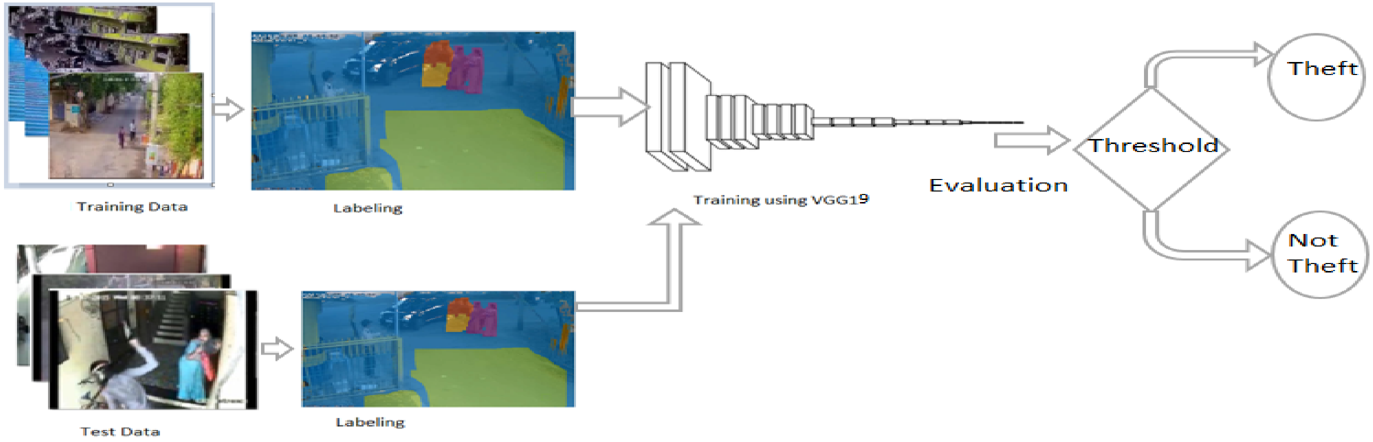


Fig. 3. Proposed Methodology

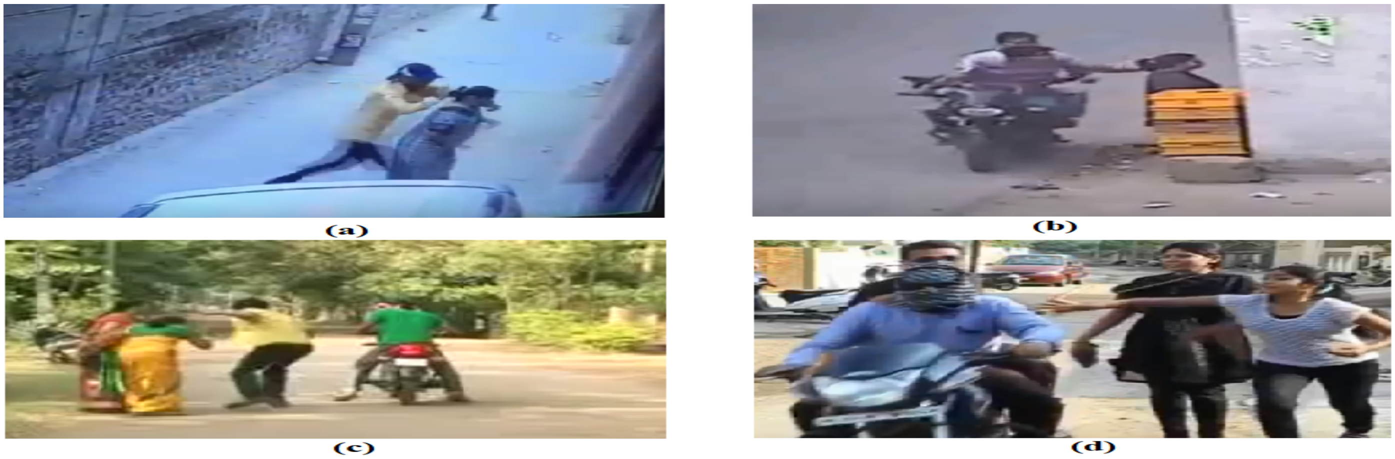


Fig. 4. Few Glimpse of Snatching from Data Set

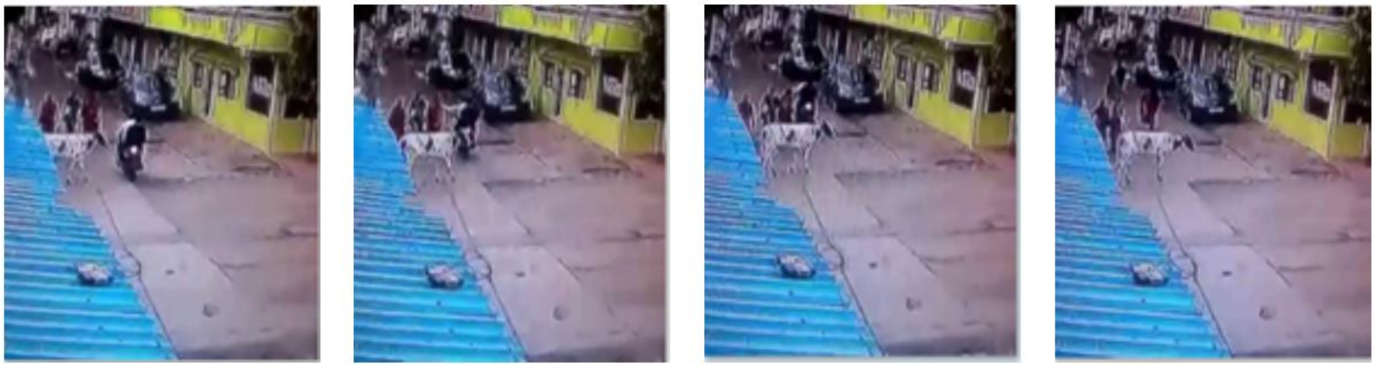
the label, and the other one is the nature of the label (rectangular and pixel region). For example, our label is "snatcher" and the region you selected (rectangular or pixel region). The nature of the object is described in the Scene label format, such as the "background" and "environment". We also relate this label to our specific, defined frame.

3) Label the image objects either in rectangular or pixel format. We labeled all the images in pixel labeler format as show in figure 9.
4) The "green" color indicates "road" object, "pink" color defines women(victim), 'orange' color defines "snatcher", "yellow" color is for "vehicle (bike)" identification and "blue" color is for "background". We further creates classes of the mentioned objects and assign indexes to all the classes. In the end, we export label to the file or in the work space to save the labeled images using the **Export Labels** option shown in figure 10.

E. VGG19 Convolutional Neural Network

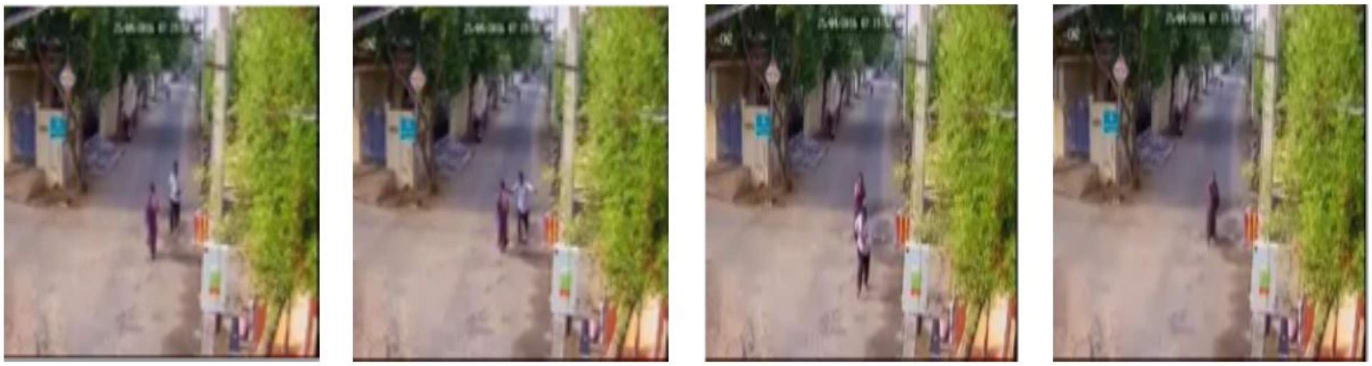
Convoluted networks (ConvNets) have been highly successful in recognition of large-scale images and videos, due to large-scale public storage depots (fast processing system such as GPU based operating system used and ImageNet or segmented the image into large clusters) [26]. In ConvNet depth measurements in fairy settings, our ConvNet Layer structure is designed with the same codes. In the training process, our fixed ConvNets input size is 112*112*128 RGB. Our only pre-processing is to reduce the average RGB value calculated for each pixel training group. The image is passed through a pile of convolutional layers, as shown in figure 11, and we use a very small cloud filter: 14*14*512 that is left to right, up to down, part of the concept [10].

We also use the max-pooling layer of 112*56*28*14*1, its nature is like a linear transformation input, but it is not linear. The first step of convolution is that 1 pixel is fixed, and the conversion space is filled. Layer input saves space resolution after convection, that is, for the conversion of 3x3, the padding is 1 pixel. Convolutional layers are stacked with different depths in different architectures following three layers Fully Connected (FC) layers: several channels in the first layer are 4096, and the last layer has 1000 channels. The last layer is a soft-max layer. We implement the same completed connection configuration for the overall system. All hidden layers are line-aligned, and our network is not standardization for Local



(a) Two persons on the bike (b) Snatching chain (c) Ran after snatching (d) Victim ran to catch him

Fig. 5. Case1: Snatching sequence



(a) Snatcher pretends (b) Snatching chain (c) Ran after snatching (d) Victim ran to catch him

Fig. 6. Case2: Snatching sequence without using the bike

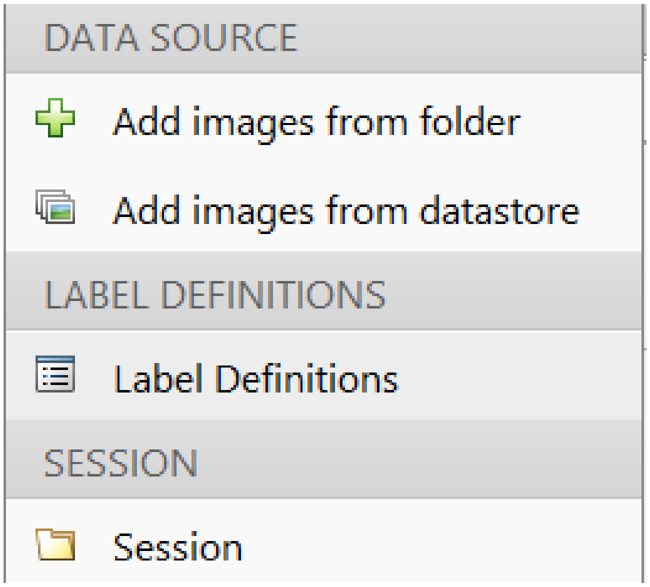


Fig. 7. Load (Video, Image Sequence or Custom Reader)

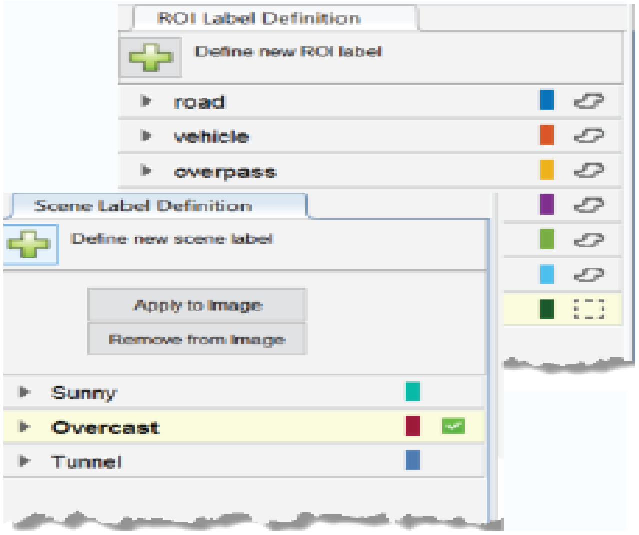


Fig. 8. Define ROI and Scene Labels

Response Normalization (LRN), which improves ILSVRC datasets but increases memory consumption and computing time [19]. Despite the great depth, the number of weights in our networks are not larger than the shallow net weight, with greater convolutional layer width and acceptable fields.

The training is done as [17] using the low-volume gradient to perform more impulse optimization for logistic regression .

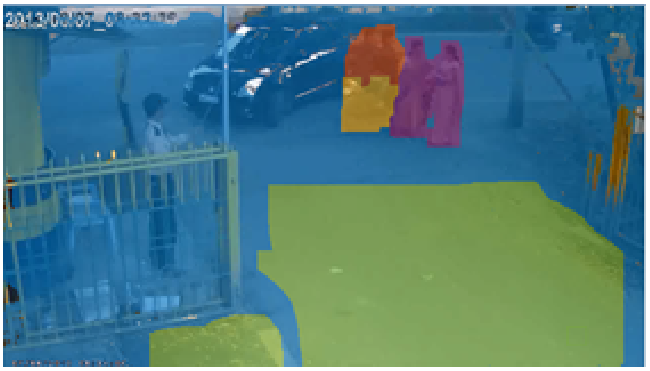
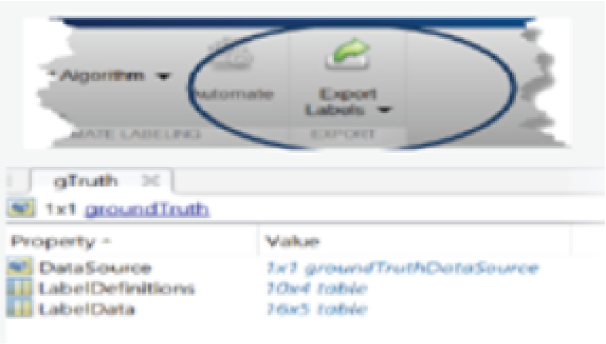


Fig. 9. Labeled objects



1x1 groundTruth

Property	Value
DataSource	1x1 groundTruthDataSource
LabelDefinitions	10x4 table
LabelData	16x5 table

Fig. 10. Export Labeled

The batch size goes to 256, and the torque is 0.9. Among the first two connecting layers, there are two ways to configure the training scale. The first one will change S, which one-scale training corresponds to the contents of the image can be represented by crop samples still multi-scale statistical image. In our experiments, we evaluated the two-scale model of training: S = 256 widely used before art, and S = 384 ConvNet setup, we used the first S = 256 Training network S = 384 networks, S = 256 pre-trained. We start with heavy weights, we have used a lesser initial 10-3-degree instruction, the second method is to configure a multi-scale S training, a specific range across each training, smax random samples, images individually readjusted S (stxikena = 256 and we use smax = 512) because the image may have different sizes for objects, so it is considered beneficial during the training period.

For many reasons, we prepare a multi-scale model with all scale layouts in the same configuration as [26] and then apply a fully uncropped image of the entire convolution network . The resulting category is graphical score; the size of the image depends on the numbers of classes as the number of channels and their variable resolution of space. In the end, to get a fixed size of the image scoring point, the category score is the average score. Since the entire computational network is applied to the whole image, multiple crops must be tested in different tests, which are more efficient, because it requires computation for each crop.

Our implementation from the public is derived from available C ++ Caffe tool [27], but it contains many important modifications that allow us to train. For training and evaluation of multiscale images we need to install multiple GPUs in one system as the authors did in this paper [10]. Multi-GPU training takes advantage of the data parallelism and is done by dividing each series of images for training on several GPU series and processing them parallel to each GPU. After calculating the GPU series of gradients, they concentrate to obtain a gradient of the complete series. Gradient calculations are synchronized between the GPU, so the results must be the same as the training model result on one GPU.

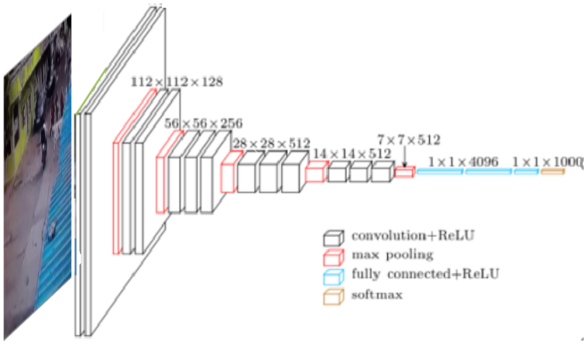


Fig. 11. VGG19 Architecture

ILSVRC has been used for many years by the algorithm of one or more of the following tasks for image classification problem has algorithms generate a list of the categories of objects that are present in the image, localization algorithm explain the scale of the image and the axis of the bounding areas of the image [19], [26], [28]. Object detection has algorithms that create a list of object classes in the image along with a border-oriented box that indicates the position and size of each copy of each object class. For checking the accuracy of ILSVRC, we have to find precision and recall.

The major contribution of our paper, the methodology used in this paper VGG19 (Convent Neural Network), has never been implemented on this type of unstructured dataset, e.g. videos. As we have studied the previous research literature, human recognition is done on only images. In this paper, we proposed a method on a video dataset, generate frames and further label the objects which detect surveillance before it. The proposed method outperformed the state-of-the-art method with 81% accuracy.

IV. Result and Discussion

Video surveillance is an important area of research for the researchers and law enforcement agencies due to the widespread usage of cameras for abnormalities detection. Several techniques have been proposed to make a robust Video surveillance system for anomaly detection, but efficiency is still a big challenge for the researchers.

In this study, we used the VGG19 architecture of a convolutional neural network to predict video surveillance in snatching videos. The proposed architecture of VGG19 is particularly altered for video surveillance and detecting anomalies in the video. The combination of ConvNets, Convolutional layer and max-pooling in a proposed order found to be efficient and effective, particularly in snatching detection. To the best of author knowledge, the proposed architecture has never been used for this purpose on an unstructured dataset.

We also compare the proposed system with the state-of-the-art system fined tuned models VGG16 and AlexNet. The earlier used models are not robust, not completely infallible and have false detections. To evaluate this process efficiently, a series of tests were carried out with the snatching theft videos that were not used in the training set. Another important aspect of our system is the processing time, which is far much better than the state-of-the-art systems with the same experimental setup. The processing time is a very important aspect of these kinds of real-world crime scenarios. The experiment was carried out on 300 video frames of the same data set. The results are shown in Table II.

TABLE II. PERFORMANCE COMPARISON OF STATE-OF-THE-ART AND THE PROPOSED METHOD

Performance Measure	AlexNet	VGG16	VGG19
Positive Detection	219	231	239
Fails	81	69	61
Accuracy (%)	73	77	81
Frames Per Second (FPS)	0.4	0.04	0.025

V. CONCLUSION

A surveillance system is to detect and identify abnormal and anomalous events. This will only be possible when we select objects and features from the anomalous events. Certain weaknesses make it harder and more difficult. The amount of moving targets imitates the performance, speed and amount of motion in the anomalous video. The dataset used in this paper is Snatch 1.0; the video in the datasets are further divided into categories of normal and snatching videos. Then the videos are converted into image frames. The frames are labelled to identify the objects which we have selected for video surveillance detection.

For the implementation of the proposed method, we used VGG19 deep neural network and performed experiments on GPU based system. Later on, a comparison of experimental results was performed with the original video, and the accuracy and performance of the model were evaluated using the evaluation, as mentioned in the above techniques. The proposed system outperformed as compared to state-of-the-art systems with 81 % accuracy and 0.025 frames per second detection time.

In the future, we aim to work on further improving its accuracy and time efficiency by using ensemble methods along with Bidirectional Long Short Term Memory (BLSTM). We will also consider the demographic factors and crime statistics of the region to predict crime so that law enforcement agencies can take precautionary measures.

REFERENCES

[1] T. Manjunath, R. S. Hegadi, and G. Ravikumar, "A survey on multimedia data mining and its relevance today," *IJCSNS*, vol. 10, no. 11, pp. 165–170, 2010.

[2] A. B. Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," *Expert Systems with Applications*, vol. 91, pp. 480–491, 2018.

[3] X. Chen and C. Zhang, "An interactive semantic video mining and retrieval platform–application in transportation surveillance video for incident detection," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 2006, pp. 129–138.

[4] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Computer Vision and Image Understanding*, 2018.

[5] D. Xu, Y. Yan, E. Ricci, and N. Sebe, "Detecting anomalous events in videos by learning deep representations of appearance and motion," *Computer Vision and Image Understanding*, vol. 156, pp. 117–127, 2017.

[6] L. Yi, H. Huang, D. Liu, E. Kalogerakis, H. Su, and L. Guibas, "Deep part induction from articulated object pairs," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, p. 209, 2019.

[7] P. Thirumurugan and S. H. Hussain, "Event detection in videos using data mining techniques," *International Journal of Computer Science and Information Technologies*, vol. 3, no. 2, pp. 3473–3475, 2012.

[8] D. Singh and C. K. Mohan, "Graph formulation of video activities for abnormal activity recognition," *Pattern Recognition*, vol. 65, pp. 265–272, 2017.

[9] E. Cermeño, A. Pérez, and J. A. Sigüenza, "Intelligent video surveillance beyond robust background modeling," *Expert Systems with Applications*, vol. 91, pp. 138–149, 2018.

[10] Y. Zhang, H. Lu, L. Zhang, X. Ruan, and S. Sakai, "Video anomaly detection based on locality sensitive hashing filters," *Pattern Recognition*, vol. 59, pp. 302–311, 2016.

[11] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, 2017.

[12] E. Fenil, G. Manogaran, G. Vivekananda, T. Thanjaivadivel, S. Jeeva, A. Ahilan *et al.*, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional lstm," *Computer Networks*, vol. 151, pp. 191–200, 2019.

[13] A. V. Kate, P. Nikilav, S. Giriesh, R. Hari Prasath, and J. Naren, "Multimedia data mining-a survey," *International Journal Of Engineering And Computer Science*, vol. 3, no. 12, 2014.

[14] J. Oh, J. Lee, and S. Kote, "Real time video data mining for surveillance video streams," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2003, pp. 222–233.

[15] A. Chen, M. Wu, Y. Zhang, N. Li, J. Lu, S. Gao, and J. Yu, "Deep surface light fields," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 1, no. 1, p. 14, 2018.

[16] A. Dilawari, M. U. G. Khan, A. Farooq, Z.-U. Rehman, S. Rho, and I. Mehmood, "Natural language description of video streams using task-specific feature encoding," *IEEE Access*, vol. 6, pp. 16 639–16 645, 2018.

[17] T. Karthikeyan, B. Ragavan, and N. Poornima, "A comparative study of algorithms used for leukemia detection," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume*, vol. 5.

[18] D. Roy *et al.*, "Snatch theft detection in unconstrained surveillance videos using action attribute modelling," *Pattern Recognition Letters*, vol. 108, pp. 56–61, 2018.

[19] H. Xu, M. Fang, L. Li, Y. Tian, and Y. Li, "The value of data mining for surveillance video in big data era," in *Big Data Analysis (ICBDA), 2017 IEEE 2nd International Conference on*. IEEE, 2017, pp. 202–206.

[20] W. Chu, H. Xue, C. Yao, and D. Cai, "Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 246–255, 2018.

[21] S. Pang, J. J. del Coz, Z. Yu, O. Luaces, and J. Díez, "Deep learning to frame objects for visual target tracking," *Engineering Applications of Artificial Intelligence*, vol. 65, pp. 406–420, 2017.

[22] A. Feizi, "High-level feature extraction for classification and person re-identification," *IEEE Sensors Journal*, vol. 17, no. 21, pp. 7064–7073, 2017.

[23] Y. Tian, G.-H. Lee, H. He, C.-Y. Hsu, and D. Katabi, "Rf-based fall monitoring using convolutional neural networks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, p. 137, 2018.

[24] U. M. Butt, B. Husnain, U. Ahmed, A. Tariq, I. Tariq, M. A. Butt, and M. S. Zia, “Feature based algorithmic analysis on american sign language dataset.”

[25] D. Roy *et al.*, “Snatch theft detection in unconstrained surveillance videos using action attribute modelling,” *Pattern Recognition Letters*, vol. 108, pp. 56–61, 2018.

[26] J. Oh, J. Lee, and S. Hwang, “Video data mining: Current status and challenges. encyclopedia of data warehousing and mining.(a book edited by dr. john wang),” *Idea Group Inc. and IRM Press*, 2005.

[27] A. Divakaran, K. Miyahara, K. A. Peker, R. Radhakrishnan, and Z. Xiong, “Video mining using combinations of unsupervised and supervised learning techniques,” in *Storage and Retrieval Methods and Applications for Multimedia 2004*, vol. 5307. International Society for Optics and Photonics, 2003, pp. 235–244.

[28] D. Saravanan and S. Srinivasan, “Data mining framework for video data,” in *Recent Advances in Space Technology Services and Climate Change (RSTSCC), 2010*. IEEE, 2010, pp. 167–170.

Understanding Attribute-based Access Control for Modelling and Analysing Healthcare Professionals' Security Practices

Livinus Obiora Nweke[1]
Information Security and Communication Technology
Norwegian University of Science and Technology (NTNU)
Gjøvik, Norway

Prosper Yeng[2]
Information Security and Communication Technology
Norwegian University of Science and Technology (NTNU)
Gjøvik, Norway

Stephen D. Wolthusen[3]
School of Mathematics and Information Security
Royal Holloway, University of London
Egham, United Kingdom
Information Security and Communication and Technology
Norwegain University of Science and Technology (NTNU)
Gjøvik, Norway

Bian Yang[4]
Information Security and Communication Technology
Norwegian University of Science and Technology (NTNU)
Gjøvik, Norway

***Abstract*—In recent years, there has been an increase in the application of attribute-based access control (ABAC) in electronic health (e-health) systems. E-health systems are used to store a patient's electronic version of medical records. These records are usually classified according to their usage i.e., electronic health record (EHR) and personal health record (PHR). EHRs are electronic medical records held by the healthcare providers, while PHRs are electronic medical records held by the patients themselves. Both EHRs and PHRs are critical assets that require access control mechanism to regulate the manner in which they are accessed. ABAC has demonstrated to be an efficient and effective approach for providing fine grained access control to these critical assets. In this paper, we conduct a survey of the existing literature on the application of ABAC in e-health systems to understand the suitability of ABAC for e-health systems and the possibility of using ABAC access logs for observing, modelling and analysing security practices of healthcare professionals. We categorize the existing works according to the application of ABAC in PHR and EHR. We then present a discussion on the lessons learned and outline future challenges. This can serve as a basis for selecting and further advancing the use of ABAC in e-health systems.**

Keywords*—Attribute-Based Access Control (ABAC); e-health systems; Personal Health Record (PHR); Electronic Health Record (EHR)

I. Introduction

There has been a growing interest in the application of ABAC in e-health systems. This is evident by the increasing number of publications and on-going research activities in that direction. According to Gartner report [1] it is predicted that 70% of enterprises will adopt ABAC mechanism as the most dominant access control mechanism for the protection of critical assets. In the healthcare industry, e-health systems interact with critical assets like electronic medical records, and ABAC has been shown to offer a promising approach to securing these critical assets.

Traditionally, medical records are paper-based but tremendous progresses in information and communication technology have led to a shift from paper-based medical records to electronic version of the medical records. Like the traditional paper-based medical record, electronic version of the medical record is a collection of medical history of an individual. However, unlike the traditional paper-based medical records, the electronic version is stored in electronic format following the required standards.

The electronic version of medical records is usually classified according to their usage i.e., electronic health record (EHR) and personal health record (PHR). Whilst EHRs are electronic medical records of an individual held by the healthcare providers; PHRs are referred to as electronic medical records of an individual held by the individual themselves. Although EHRs can be shared across different healthcare providers, PHRs have shown to be an effective approach for individuals to share their electronic medical records with different healthcare providers, family and friends.

Sharing of electronic medical records raises security and privacy concerns for both EHR and PHR. For EHR, healthcare providers are required by regulatory bodies to ensure that the security and privacy of the electronic medical records are maintained. In the case of PHR, an individual would want to ensure that only authorized entities have access to their electronic medical records. Several approaches have been proposed to address the security and privacy concerns raised by EHR and PHR. The approach that have received wide-spread acceptance is ABAC.

ABAC aims to provide fine-grained access to a resource or an object based on the attributes of the subject and that of the object; in addition to the environmental conditions. A subject refers to an entity such as a person, process or device that wishes to access a resource or an object. A resource or an object is a system-related entity containing information such as records, that a subject desires to access. The environmental conditions are the operational contexts such as the time and location of access. Hence, in ABAC, the attributes of the subject and the requested object as well as the environmental condition determines the set of operations that can be executed on the requested object.

A wide range of applications of ABAC in e-health systems have been proposed in the literature and examined in individual studies. However, a comprehensive survey of these techniques that can serve as a basis for selecting and further advancing the use of ABAC in e-health systems is still missing in the literature. Abbbas and Khan in [2] presented a review on the state of the art in privacy preserving techniques for e-health cloud based systems. The authors in [3], [4] provided a survey on the security and privacy issues in e-health cloud based systems. To the best of our knowledge, there is no survey on the application of ABAC in e-health systems.

In this paper, we present a survey on the application of ABAC in e-health systems. We categorize the different applications of ABAC in e-health systems according to those use in PHR and those apply in EHR. We present a comparison of the different approaches employ in the existing works. Then, using some of the key features of the existing approaches, we present a discussion on their differences. Also, we describe the lessons learned from the survey and outline future challenge. Lastly, the concept of modelling and analysing healthcare professionals' security practices is discussed.

The rest of this paper is organised as follows. Section II presents an overview of the security and privacy requirements for e-health systems. Also, the dominant access control mechanisms deploy in e-health systems are explored, and the justification for wide-spread acceptance of ABAC in e-health systems is described. Section III presents a literature survey of the existing works on the application of ABAC in e-health systems. Section IV discusses the lessons learned from the survey and outline future challenge. In addition a discussion on modelling and analysing healthcare professionals' security practices is presented. Section V concludes the paper.

II. Background

In this section, we provide an overview of the security and privacy requirements for e-health systems. We also examine the commonly used access control measures for e-health systems and why ABAC mechanism is the most preferred access control mechanism for e-health systems.

A. Requirements of E-Health Systems

Several standards and laws have been proposed to specify the security and privacy requirements for e-health systems. The most popular of these standards and laws is the American standard health insurance portability and accountability act (HIPAA) [5]. HIPAA is mainly concern about the privacy and security of patient health information (PHI). With the migration of PHI from paper-based to electronic format, HIPAA was upgraded to health information technology for economic and clinical health (HITECH) to address privacy and security concerns posed by such migration.

HIPAA is applicable to all types of Covered Entity or Business Associate that processes PHI. Covered Entity is a health care provider, a health plan or a health care clearing house who, in its normal activities, creates, maintains or transmits PHI [5]. Business Associate is a person or business that provide a service - or performs certain function or activity for - a covered entity when that service, function or activity involves the business associate having access to PHI maintained by the covered entity [5]. Usually, a business associate is required to sign business associate agreement with the Covered Entity stating what PHI they can access, how it would be used and that it will be returned or destroyed once the task it is needed for is completed [5]. Also, while the PHI is in the custody of the business associate, the business associate has the same HIPAA compliance obligations as a Covered Entity.

The two types of rules specified by HIPAA are the privacy rule and security rule. The privacy rule protects all PHI held or transmitted by a covered entity or its business associate, in any form or media, whether electronic, paper or oral [5]. Under the security rule, covered entities are required to evaluate risks and vulnerabilities in their environments and to implement security controls to address those risks and vulnerabilities [6]. There are three parts to the security rule: administrative safeguards, which is in the form of policies and procedures that brings the privacy rule and security rule together; technical safeguards refer to the technology that is used to protect PHI and provide access to the data; and physical safeguards, which has to do with physical access to PHI regardless of its location [6].

An international standard that defines the requirements for e-health systems is the ISO/IEC 27799 [7]. The ISO/IEC 27799 provides special recommendations on security needs in the healthcare sector, taking into account the unique nature of its operating environment. It applies ISO/IEC 27002 to the healthcare domain with appropriate security controls towards enhancing the protection of PHI. The development of ISO/IEC 27799 took into consideration, personal data protection legislations, privacy and security best practices, individual and organizational accountability, meeting the security needs identified in common healthcare situations, and operating electronic health information systems in an adequately secured healthcare environment. Also, ISO/IEC 27799 aims to protect information such as PHI, pseudonymized data derived from PHI, clinical or medical knowledge related or not related to any patient, data on health professionals, staff and volunteers, audit trail data produced by health information systems, including access control data and other security related system configuration data, for health information systems.

Other important standards for e-health systems include OpenEHR [8], the health level 7 clinical document architecture (CDA) [9], and the continuity of care document (CCD) [9]. The OpenEHR is an open standard that specifies the management and storage, retrieval and exchange of health data in EHRs. Also, openEHR defines specifications for clinical information models, EHR Extracts, demographics, data types and various kinds of service interfaces [8]. The HL7 CDA is a document markup standard that specifies the structure and semantics of clinical documents for the purpose of facilitating exchange between healthcare providers and patients [9]. A clinical document is defined by HL7 CDA as having the following features: persistence, stewardship, potential for authentication, context, wholeness, and human readability [9]. And CCD is a joint effort of HL7 International and American society for testing and materials (ASTM) to enable interoperability of clinical data [9]. It allows physicians to send electronic medical information to other providers without loss of meaning and as such, improves the overall patient care.

In general, the requirements that are of interest to this survey are the recommended technical safeguards for e-health systems. These technical safeguards aim to provide secure, reliable, access to PHR or EHR; where and when it is requested. The requirements include the following [5]:

- Implement a means of access control
- Introduce a mechanism to authenticate PHR and EHR
- Implement tools for encryption and decryption
- Introduce activity logs and audit controls

B. Access Control Mechanisms

One of the security controls necessary to meet the security and privacy requirements for e-health systems is the implementation of access control mechanisms. These are measures that can be used to regulate access to a given resource. Earlier implementation of access control mechanisms in e-health systems employ role-based access control (RBAC) [2]. RBAC restricts access to a resource based on the user's role. The use of a role based access control suffers some drawbacks as the definition of roles is static and it lacks flexibility and responsiveness. Every user needs to be enrolled in advance in the system. For example, in an emergency situation where the patient is outside the local domain where the patient health information held, a doctor not registered within the local domain of the patient will not be able to access the patient's health information. Therefore, the efficacy of role-based access control is limited because it cannot handle situations where unregistered personnel requires access to the system as in the case of emergency that we described.

Emergency access such as self-authorization and break the glass (BTG) are basic requirements in healthcare systems. Self-authorization is a provision in the access control mechanism that allows healthcare professionals to access the minimum and necessary healthcare records for therapeutic purposes during emergency situations. Similarly, BTG mechanism is used when conventional access control mechanisms are inadequate to access minimum and necessary healthcare information for therapeutic measures [10], [11]. Considering that RBAC policies rely on permissions that does not often change [12], installing emergency access mechanisms on static roles may pose a high security threat. For instance, an adversary who might have unlawfully acquired health professionals' credentials under RBAC, could easily compromise healthcare records by using the emergency access control windows since there are no other control variables to authentic the accesses of the malicious user.

A flexible access control mechanism that provides fine grained access control to a resource is ABAC. Like RBAC, ABAC employs a policy driven approach. However, in ABAC, access to a resource is granted based on the attributes of the subjects and the objects together with the environmental attributes. This eliminates the need of having to register a user into the system before providing access; instead, access is granted based on the attributes of the user and that of the requested resource. Thus, ABAC mechanisms would provide appropriate level of access to healthcare records even for any extraordinary actions that need to be taken during emergency situations.

For emergency situations, ABAC ensures that the authentication mechanism of emergency accesses can be configured to include more control variables such as attributes of the user, environment and resources to reduce risk of privacy and security breaches. For instance, the resource and environmental attributes such as the patient status and location could indicate emergency care or intensive-care services. Hence, any accesses other than the specified attributes would be restricted, to reduce the risk of exploitation. Therefore, ABAC policies enables flexible configurations for users to override their conventional access restrictions in a controlled and justifiable manner in emergency access scenarios.

ABAC have shown to be an effective and efficient mechanism for providing fine-grained access to PHRs and EHRs given the dynamic nature of today's e-health environment. Also, it can be combined with different cryptographic schemes to provide secure and anonymous sharing of PHRs and EHRs among healthcare providers and patients. So many research efforts are on-going in developing appropriate ABAC model for e-health systems. The next section provides a survey of some of these efforts to further support the assertion that ABAC is a much better access control mechanism for e-health systems.

III. Literature Survey

In this section, we present a survey of the existing literature on the application of ABAC in e-health systems. We categorize the existing work according to the type of patient's electronic version of medical records considered. Already we have observed that the electronic version of a patient health record is usually classified according to those held by the patient themselves (PHR) and those held by the healthcare providers (EHR). We use this understanding to present the different applications of ABAC in e-health systems.

A. Application of ABAC in Personal Health Record (PHR)

PHR offers a flexible and convenient way for storing and sharing a patient's electronic version of medical records. It empowers the patients by giving them control over their medical record and deciding with whom to share those records. However, the current trend in the storage of PHR has shown that cloud platforms are very popular way of storing PHR. This raises questions of security and privacy of PHR as there have been wide spread concerns that PHR stored in the cloud may be exposed to unauthorized parties. Several approaches that use ABAC in PHR have been proposed in the literature to address these concerns.

A typical use case scenario of the application of ABAC in PHR is shown in Figure 1. Li et al [13] describe a unified fine-grained access control for PHR in cloud computing. In this system, the patient utilizes the cloud storage platform for storing the encrypted version their PHRs. The policy manager facilitates the encryption of the patient's PHRs. Also, the medical staff is able to download the encrypted PHRs from the cloud and use their private keys to decrypt the PHRs. A trusted attribute authority is used for all patients and medical staff to authenticate and verify their attributes.

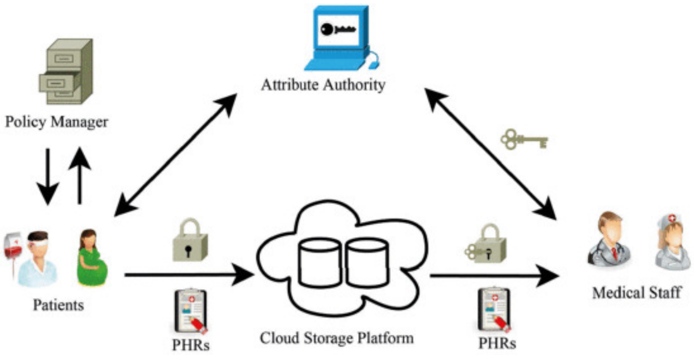


Fig. 1. Use Case Scenario of ABAC in PHR

[13]

One of the earliest approaches in the use of ABAC to provide security and privacy for PHR stored in the cloud is presented in [14]. The authors used a variant of attribute-based encryption (ABE) referred to as broadcast ciphertext policy ABE (bABE) which extends the functionality of ABE to include user revocation. An ABE uses a public key encryption system, where each user's key is labelled with a set of attributes, and the ciphertext is linked with an access policy. The private key of the user can decrypt the ciphertext only if the attribute set of the user's key matches the access policy associated with the ciphertext. Furthermore, the approach presented assumes trusted cloud provider and the use of a trusted authority to issue the relevant private keys.

Li et al in [15] propose a patient-centric framework and approach which exploits ABE techniques to provide fine-grained access control to PHR in cloud environment. In the proposed model, the system is divided into several security domains according to the different users' data access requirements. ABE is deployed to cryptographically enforce patient centric PHR access. In additional, the PHR is assumed to be stored on a semi-trusted service provider and the proposed framework supports access revocation. Another patient-centric cloud-based secured PHR system is presented in [16]. The proposed system enables secure storage of PHR data on a semi-trusted cloud service provider and allows the patient to selectively share their PHR data with wide range of users. The authors reduced key management complexity for both owners and users by dividing the users into two security domains, namely: public domain and personal domain. Also, they show that PHR owners can encrypt PHR data for the public domain using ciphertext-policy ABE scheme, while the PHR data for the personal domain can be encrypted using anonymous multi-receiver identity encryption scheme.

A fine-grained access of interactive, PHR, that extends a secure composite document format i.e., Publicly Posted Composite Documents (PPCD) is described in [17]. PPCD is a SQLite-based serialization which is developed for business workflows and is able to contain multiple documents of different sensitivity and formatting. The method proposed in this work includes both the original PPCD-type and an additional new entry table to provide for password-based and private key access. The authors employ Password Key Derivation function as the privacy preserving technique and the method also supports access revocation. Ray et al in [18] apply attribute based access control for preserving the privacy of PHR. The authors show how the privacy of PHR can be expressed and enforced through the use of an attribute based access control supported by extensible access control markup language (XACML). In this paper, the XACML is used to model the different types of policies and expressing the patient's privacy preference for subsequent enforcement by the attribute based access policies.

There are constraints imposed on cloud based PHR schemes that use ABE. An approach to address these constraints is proposed in [19]. The method adopted in this work involves the use of multi-authority system architecture, unlike existing methods that utilize single trusted authority. In addition, a proxy re-encryption scheme is deployed to ensure that only authorized users are able to decrypt the required PHR files. A more recent work by Li et al [13] present a unified fine-grained access control for PHR in cloud environment. The proposed approach is able to store PHR for multiple patients. It consists of ABE layer and symmetric layer. Whilst the ABE layer facilitates a multi-privilege access control for PHR from multiple patients; in the symmetric layer, symmetric keys that match medical workers' access privileges and the keys with higher privilege can override keys with lower privilege but not the other way around. Also, the authors use ciphertext policy ABE as the privacy preserving technique for the proposed method.

B. Application of ABAC in Electronic Health Record (EHR)

EHR is handled by healthcare providers and also, it provides them with the opportunity of sharing those records among different healthcare providers. EHR is usually stored on-premise under the administrative control of the healthcare provider but recent trends have shown a gradual shift from on-premise storage of EHR to cloud. This further increases the risk of exposing EHR to unauthorized parties. However, ABAC has demonstrated to be a promising approach to mitigating the risk of exposing EHR to unauthorized parties. Different methods that employ ABAC in EHR have been discussed in existing works.

The system architecture as shown in Figure 2, depicts a use case scenario of the application of ABAC in EHR. Joshi et al [20] in this work provide users access to the system using Access Broker Unit. The Access Broker Unit consists of the organizational Knowledge Base, the Rule Based Engine and the Policy Unit. The Organization Knowledge Base stores all the details of the users in the form of an ontology - the EHR Ontology. The Policy Unit stores all the access policies. And the Rule Based Engine uses the user and document attributes from the ontology for implementing the access control policies. The authors use ABE for encryption, and the Key Generation Unit generates the private keys required for the ABE. Then, the encrypted data are stored in the cloud, which hosts, the EHR Ontology.

Pussewalage and Oleshchuk in [21] propose an ABAC scheme for secure sharing of EHR. The scheme uses selective

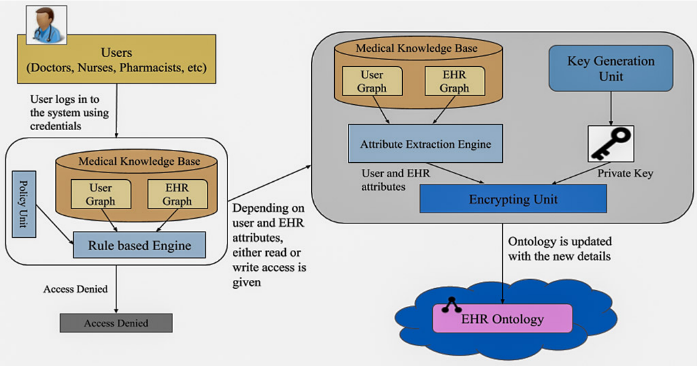


Fig. 2. Use Case Scenario of ABAC in EHR

[20]

disclosure that meets the security requirement of EHR. An access requester supplies a valid set of attributes that satisfies the underlying policy of the requested object using attribute and private key commitments. The proposed approach is said to be collision resistant; such that it is impossible to collude attributes of more than one user to gain access to EHR. This is achieved by giving a unique identifier to every user and including it to every attribute key owned by the respective users. In addition, the proposed method supports on demand user revocation and it is applicable to on-premise storage platform.

Several standards have been developed to facilitate interoperability of EHR. The most recent effort in that direction is the Fast Health Interoperability Resources (FHIR) [22], which specifies requirements for fast and efficient storage/retrieval of EHR. The authors in [23] exploit ABAC to create owner-centric methodology for granting access to EHR. They focussed on FHIR and suggested ways to allow incremental and batch release of EHR stored using FHIR to any requesting party, based on access policies defined by the resource-owners.

Cloud based storage are currently being adopted by healthcare providers for storing EHR. Joshi et al. in [20] develop an ABAC mechanism for cloud-based EHR that uses ABE to securely store EHR at field level. The developed system extracts the user and EHR filed attribute from a HIPAA complaint knowledge graph which facilitates easy querying and faster data access operation. Also, in [24] the authors propose ABAC which uses Hidden Vector Encryption system to encrypt EHR in cloud environment. The approach presented is able to protect EHR from insider attacks as EHR can only be view by those that are able to supply the appropriate attributes. Seol et al in [25] present a cloud-based EHR model that performs ABAC using XACML. The combination of XML encryption and XML digital signatures are used as security and privacy preserving technique.

There are situations where EHR is shared among different providers. It is possible for an adversary to infer the health condition of a patient by observing the frequency in which the EHR is accessed by a particular healthcare provider. This type of situation violates the privacy of the patient. The authors in [26] propose an efficient multi-show unlinkable access for collaborative e-health environment that exploits attribute-based credential scheme. They utilize anonymous attribute credentials which ensure that users can anonymously prove the ownership of a set of attributes to a verifier and by so doing, obtain access to the protected resources. The method involves randomization of the users credential along with its signature before being disclosed to a verifier. Similarly, Michalas and Weingarten in [27] describe the use of HealthShare, a secure approach for sharing EHR between multiple organizations hosting patient's data in different cloud environments. In the proposed method, a revocable key-policy ABE is used to ensure that access by a malicious or compromised user/organization can easily be revoked without generating new encryption keys.

IV. Discussion

In this section, we present a comparison of the different approaches used in the existing works. We then use some of the key features of the existing approaches to present a discussion on their differences. Also, we describe the lessons learned from the survey and outline future challenge. Lastly, the concept of modelling and analysing healthcare professionals' security practices is discussed.

A. Comparison of the Different Approaches

A detailed summary of the existing works on the application of ABAC in e-health systems that we have presented in this work is shown in Table I. Some of the key features of the existing approaches are employed to discuss the differences in the approaches. Also, we describe the lessons learned from the survey and outline future challenge.

1) Privacy Preserving Techniques: refer to approaches that may be exploited to provide confidentiality of PHR and EHR. It involves the encryption of the health data to be stored using cryptographic methodologies such that only an individual that possess the decryption key can have access to the health data. It can be observed from Table I that whilst the existing works employ different privacy preserving techniques, ABE and its variants appears to be the most popular approach.

ABE is a type of public key encryption where the private key and the ciphertext are related with a set of attributes or an access policy over the attributes of the users. There are two main variants of ABE, and they are: ciphertext-policy ABE [28] and key-policy ABE [29]. A combination of ciphertext with access policy specifying the attributes of legitimate users is employ in ciphertext-policy ABE, while key-policy ABE uses a set of attributes and private keys associated with the access policy to specify which ciphertexts the key holder can access. Li et al. in [13] argue that ciphertext-policy ABE is more flexible and appropriate for PHR than key-policy ABE in practice. This is evident from the summary in Table I as most application of ABAC in PHR use ciphertext-policy ABE for privacy protection.

Another privacy preserving technique that is used in the existing works is XACL. XAMCL defines a declarative fine-grained, ABAC control policy language which describes how to evaluate access requests according to rules stated in access policies [30]. The authors in [18] use XAMCL to show how a patient's privacy preferences could be expressed and enforced in PHR. XAMCL is deploy in [23] as the privacy preserving technique for EHR. The authors utilize XAMCL for providing

TABLE I. Summary of Existing Works on Application of ABAC in E-Health Systems

Work	Type of Health Record Considered	Privacy Preserving Technique	Access Revocation	Storage Platform Used	Adversarial Model Assumption
[15]	PHR	ABE	Supported	Cloud	Semi-trusted Service Provider
[16]	PHR	Ciphertext-Policy ABE	Not Specified	Not Specified	Semi-trusted Service Provider
[18]	PHR	XACML	Not Specified	Not Specified	Not Specified
[20]	EHR	Ciphertext-Policy ABE	Not Specified	Cloud	Not Specified
[27]	EHR	Key-Policy ABE	Supported	Cloud	Trusted Service Provider
[25]	EHR	XACML with XML Encryption and XML Digital Signatures	Not Specified	Cloud	Not Specified
[13]	PHR	Ciphertext-Policy ABE	Not Specified	Cloud	Semi-trusted Service Provider
[17]	PHR	Password Key Derivation Function	Supported	Cloud	Not Specified
[26]	EHR	U-Prove	Not Specified	On-Premise	Trusted Service Provider
[21]	EHR	Not Specified	Supported	On-Premise	Trusted Service Provider
[24]	EHR	Hidden Vector Encryption	Not Specified	Cloud	Not Specified
[19]	PHR	Proxy Re-encryption	Supported	Cloud	Semi-trusted Service Provider
[14]	PHR	Ciphertext-Policy ABE	Supported	Cloud	Trusted Service Provider
[23]	EHR	XACML	Not Specified	On-Premise	Trusted Service Provider

fine-grained authorization and access to FHIR resources. Seol et al in [25] employ XACML with XML encryption and XML digital signatures as additional measure for ensuring that the privacy and security of EHR are preserved.

Other privacy preserving techniques used in the existing works surveyed include: the use of password key derivation function, U-Prove, hidden vector encryption and proxy re-encryption. Balinsky and Mohammad [17] use password key function to provide end-to-end encryption and show that it ensures no central authority is needed when accessing plaintext data or decryption keys. Authors in [26] argue that enforcing anonymously as well as multi-session unlinkable access for users in e-health is very pertinent. They use the standard U-prove credential scheme and formally prove its multi-show unlinkability property. The paper in [24] use hidden vector encryption to encrypt and embed access control policies within the encrypted data. This approach completely removes the need for two separate security controls. Also Pussewalage and Oleshchuk [19] apply a proxy re-encryption scheme to ensure that only authorized users are able to decrypt PHR files.

2) Access Revocation: is another important feature of the existing works surveyed. Although not all the works specified the presence of access revocation, it is an essential characteristic of ABAC in e-health as it enables the disabling of a user's access to PHR or EHR. Several methods have been adopted in order to provide efficient access revocation. The authors in [15] implement access revocation by re-encrypting the ciphertexts and updating the users' private keys. For the papers in [19], [21], the attribute authority is responsible for the access revocation process.

The remaining papers surveyed in this work adopted direct access revocation. The authors in [17] present direct access revocation where the owner of PHR can revoke access by re-encrypting and signing the PHR with a set of newly generated keys. For the paper in [14], each user has a user-index which facilitates direct revocation of user access to an encrypted data. This eliminates the need for re-encrypting the data or refreshing the system parameters to implement access revocation. Also, Michalas and Weingarten [27] present an algorithm that EHR owner can use to revoke access for the unique key that is generated for a particular user. Like the approach in [14], the EHR owner does not have to decrypt and then re-encrypt file with a fresh key.

3) Storage Platform Used: refers to method used in storing the PHRs or EHRs. The traditional approach for EHRs has been on-premise, but recent trends have shown a gradual shift to cloud environment. This is due to flexibility and cost-effectiveness that cloud storage environment offers. In the case of PHRs, cloud storage has been the prevalent methodology for storage because it is infeasible for a single individual to bear the cost of setting up storage resources for storing PHRs. Hence, patients that would like to be responsible for their medical health records rely of cloud storage platforms for storing their health information.

4) Adversarial Model Assumption: has to do with the assumptions made by the different models about the nature of the storage platform used in storing PHRs and EHRs. These assumptions are necessary when developing formal proof that the proposed approach is feasible and meets all the legal and ethical requirements for storing PHRs and EHRs. The adversarial model assumption considered in most of the existing papers surveyed either assumes trusted service provider or semi-trusted provider. Although these are reasonable assumptions, it would also be insightful to consider untrusted service providers. This would guarantee that the stringent privacy and security requirements for PHR and EHR are met.

5) Lessons Learned and Future Challenge: Indeed, e-health systems require a flexible and fine-grained access control mechanism for secured access to PHRs and EHRs. ABAC has shown to be an efficient and effective approach to meeting the security and privacy requirements of e-health systems. We have presented a survey of the different applications of ABAC in e-health systems. By classifying the existing works according to the types of health records considered, we are able to investigate what have been done so far in the literature.

We observe that there has been an increasing adoption of PHR for storing patient health records. This gives the patient greater control of their health record, allowing them to share it with different healthcare providers, family and friends. Also, we notice that ciphertext-policy ABE is the predominant privacy preserving technique used for PHR as it enables the patient to revoke access easily to any user they no longer want to have access to their PHR. In addition, cloud storage platform is used in all the surveyed works for storing PHR.

The storing of EHR as observed in this survey is shifting from the traditional on-premise to cloud environment. This can be attributed to the flexibility and cost-effectiveness of the cloud storage platform. Further, there is an increasing collaboration between different healthcare providers which have led to different approaches proposed for facilitating such collaborations without compromising the privacy of the patient.

All the survey works either assumes that the service provider is trusted or semi-trusted. In the future, approaches that consider untrusted service provides needs to be examined. Recent data breaches involving cloud providers and insider threats further buttress the need to investigate ABAC mechanism for e-health systems that assumes untrusted service providers. Such stringent assumption would ensure that in the case that the third party providers are compromised, the privacy of the patient is still preserved.

B. Towards Modelling and Analysing Healthcare Professionals' Security Practices

Logging of healthcare professionals' accesses is required in the code of conduct for healthcare and care service of Norway [31] and in most international standards for healthcare service. The purpose of logging and protecting the logs includes non-repudiation and investigations [32], [33]. Access logs can be analysed to improve data quality and integrity by detecting healthcare information errors and inconsistencies [32], [33]. For this reason, the Healthcare Security Practice Analysis, Modelling and Incentivization (HSPAMI) project was initiated to determine the metrics of healthcare professional's security practices towards improving upon their conscious care behaviour [34]. One of the major tasks of HSPAMI is to analyse healthcare professionals' access logs towards improving their security behaviour [34].

Analysing RBAC logs may require a lot effort and resources to design the algorithm, for such analysis to be efficient and effective. This is because RBAC mechanisms emphasize only on the role attribute as a control variable for implementing the required protection mechanisms. Without considerable efforts and resources, a higher rate of outliers, false positives and false negative rates are likely to be recorded during the analysis. It is desirable to design the algorithm for the log analysis taking into consideration the environment attributes, the resource attributes and the attributes of the objects in emergency access scenarios. For instance, the log analysis algorithm should be able to determine if the patient status was classified under emergency within the given period. Also, the location of the patient such as the type of hospital ward could support in decision making. Thus, if the patient was admitted in the intensive-care unit (ICU) or emergency ward, the environmental attributes could provide such knowledge. Since RBAC does not include these control variables, more resources may have to be invested in designing such algorithms for efficient log analysis.

In the case of ABAC logs, analysing the logs would likely require less resource to design the algorithm for such analysis to be efficient and effective. ABAC mechanism as we already observed, contain more control variables and as such the logs of ABAC would also contain those variables. These control variables in ABAC logs are desirable variables for the design of an efficient algorithm for log analysis, unlike RBAC that uses the role attribute as the main control variable. Therefore, given that ABAC logs include the control variables needed for the design of an efficient algorithm for the analysis of access logs, fewer resources are likely to be deployed in the design such algorithms.

V. Conclusion

In summary, we have presented a survey of the existing works on the application of ABAC in e-health systems. We classified the existing works according to the application of ABAC in PHR and EHR. Our survey showed that cloud based storage of PHR and EHR is very popular and that ciphertext-policy ABE is the commonly used for providing security and privacy guarantees in the storage of PHR in the cloud environment. Moreover, we presented a comparison of the different approaches employed in the existing works and used some key characteristics of the existing approaches to present a discussion on their differences. The lessons learned from the survey are described and future challenge that needs to be investigated is outlined. Lastly, a discussion on modelling and analysing healthcare professionals' security practices is presented.

References

[1] Gartner, "Market trends: Cloud-based security services market, worldwide, 2014," 2014. [Online]. Available: https://www.gartner.com/doc/2607617

[2] A. Abbas and S. U. Khan, "A review on the state-of-the-art privacy-preserving approaches in the e-health clouds," vol. 18, pp. 1431–1441, 2014.

[3] Y. Al-Issa, M. A. Ottom, and A. Tamrawi, "ehealth cloud security challenges: A survey," *Journal of Healthcare Engineering*, vol. 2019, pp. 1–15, 2019.

[4] N. A. Azeez and C. V. der Vyver, "Security and privacy issues in e-health cloud-based system: A comprehensive content analysis," *Egyptian Informatics Journal*, vol. 20, pp. 97–108, 2019.

[5] HIPAA-Journal, "Hipaa explained." [Online]. Available: https://www.hipaajournal.com/hipaa-explained/

[6] M. Scholl, K. Stine, J. Hash, P. Bowen, A. Johnson, C. D. Smith, and D. I. Steinberg, "Nist special publication 800-66 revision 1: An introductory resource guide for implementing the health insurance portability and accountability act (hipaa) security rule," 2008.

[7] ISO, "Iso/iec 27799:2016 health informatics - information security management in health using iso/iec 27002," 2016. [Online]. Available: https://www.iso.org/standard/62777.html

[8] openEHR, "openehr – a semantically -enabled health computing platform," 2016.

[9] HL7-International, "Clinical document architcture (cda)."

[10] A. Ferreira, D. Chadwick, P. Farinha, R. Correia, G. Zao, R. Chilro, and L. Antunes, "How to securely break into rbac: The btg-rbac model," in *Proc. Annual Computer Security Applications Conf*, Dec. 2009, pp. 23–31.

[11] HIPAA, "Break glass procedure: Granting emergency access to critical ephi systems," 2004.

[12] A. D. Brucker and H. Petritsch, "Extending access control models with break-glass," in *Proceedings of the 14th ACM Symposium on Access Control Models and Technologies*, ser. SACMAT '09. New York, NY, USA: ACM, 2009, pp. 197–206. [Online]. Available: http://doi.acm.org/10.1145/1542207.1542239

[13] W. Li, B. M. Liu, D. Liu, R. P. Liu, P. Wang, S. Luo, and W. Ni, "Unified fine-grained access control for personal health records in cloud computing," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1278–1289, May 2019.

[14] S. Narayan, M. Gagné, and R. Safavi-Naini, "Privacy preserving ehr system using attribute-based infrastructure," in *Proceedings of the 2010 ACM Workshop on Cloud Computing Security Workshop*, ser. CCSW '10. New York, NY, USA: ACM, 2010, pp. 47–52. [Online]. Available: http://doi.acm.org/10.1145/1866835.1866845

[15] M. Li, S. Yu, Y. Zheng, K. Ren, and W. Lou, "Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 1, pp. 131–143, Jan. 2013.

[16] C. Wang, X. Xu, D. Shi, and W. Lin, "An efficient cloud-based personal health records system using attribute-based encryption and anonymous multi-receiver identity-based encryption," in *Proc. Cloud and Internet Computing 2014 Ninth Int. Conf. P2P, Parallel, Grid*, Nov. 2014, pp. 74–81.

[17] H. Y. Balinsky and N. Mohammad, "Fine grained access of interactive personal health records," in *Proceedings of the 2015 ACM Symposium on Document Engineering*, ser. DocEng '15. New York, NY, USA: ACM, 2015, pp. 207–210. [Online]. Available: http://doi.acm.org/10.1145/2682571.2797098

[18] I. Ray, T. C. Ong, I. Ray, and M. G. Kahn, "Applying attribute based access control for privacy preserving health data disclosure," in *Proc. IEEE-EMBS Int. Conf. Biomedical and Health Informatics (BHI)*, Feb. 2016, pp. 1–4.

[19] H. S. G. Pussewalage and V. Oleshchuk, "A patient-centric attribute based access control scheme for secure sharing of personal health records using cloud computing," in *Proc. IEEE 2nd Int. Conf. Collaboration and Internet Computing (CIC)*, Nov. 2016, pp. 46–53.

[20] M. Joshi, K. Joshi, and T. Finin, "Attribute based encryption for secure access to cloud based ehr systems," in *Proc. IEEE 11th Int. Conf. Cloud Computing (CLOUD)*, Jul. 2018, pp. 932–935.

[21] H. S. G. Pussewalage and V. A. Oleshchuk, "An attribute based access control scheme for secure sharing of electronic health records," in *Proc. Applications and Services (Healthcom) 2016 IEEE 18th Int. Conf. e-Health Networking*, Sep. 2016, pp. 1–6.

[22] HL7-International, "Fhir overview," 2019. [Online]. Available: https://www.hl7.org/fhir/overview.html

[23] S. Mukherjee, I. Ray, I. Ray, H. Shirazi, T. Ong, and M. G. Kahn, "Attribute based access control for healthcare resources," in *Proceedings of the 2Nd ACM Workshop on Attribute-Based Access Control*, ser. ABAC '17. New York, NY, USA: ACM, 2017, pp. 29–40. [Online]. Available: http://doi.acm.org/10.1145/3041048.3041055

[24] E. Mrema and V. Kumar, "Fine grained attribute based access control of healthcare data," 2018.

[25] K. Seol, Y. Kim, E. Lee, Y. Seo, and D. Baik, "Privacy-preserving attribute-based access control model for XML-based electronic health record system," *IEEE Access*, vol. 6, pp. 9114–9128, 2018.

[26] H. S. G. Pussewalage and V. A. Oleshchuk, "An efficient multi-show unlinkable attribute based credential scheme for a collaborative e-health environment," in *Proc. IEEE 3rd Int. Conf. Collaboration and Internet Computing (CIC)*, Oct. 2017, pp. 421–428.

[27] A. Michalas and N. Weingarten, "Healthshare: Using attribute-based encryption for secure data sharing between multiple clouds," in *Proc. IEEE 30th Int. Symp. Computer-Based Medical Systems (CBMS)*, Jun. 2017, pp. 811–815.

[28] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attribute-based encryption," in *Proc. IEEE Symp. Security and Privacy (SP '07)*, May 2007, pp. 321–334.

[29] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in *Proceedings of the 13th ACM Conference on Computer and Communications Security*, ser. CCS '06. New York, NY, USA: ACM, 2006, pp. 89–98. [Online]. Available: http://doi.acm.org/10.1145/1180405.1180418

[30] O. Standard, "extensible access control markup language (xacml) version 3.0," Jan. 2013. [Online]. Available: http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html

[31] D. ehelse, "Code of conduct for information security and data protection in the healthcare and care services sector," 2018. [Online]. Available: https://ehelse.no/normen/documents-in-english

[32] A. Ferreira, R. Cruz-Correia, and L. Antunes, "Usability of authentication and access control: A case study in healthcare," in *Proc. Carnahan Conf. Security Technology*, Oct. 2011, pp. 1–7.

[33] A. Ferreira, P. Farinha, C. Santos-Pereira, R. J. C. Correia, P. P. Rodrigues, A. da Costa Pereira, and V. Orvalho, "Log analysis of human computer interactions regarding break the glass accesses to genetic reports," in *ICEIS 2013 - Proceedings of the 15th International Conference on Enterprise Information Systems, Volume 3, Angers, France, 4-7 July, 2013*, S. Hammoudi, L. A. Maciaszek, J. Cordeiro, and J. L. G. Dietz, Eds. SciTePress, 2013, pp. 46–53.

[34] P. Yeng, B. Yang, and E. Snekkenes, "Observational measures for effective profiling of healthcare staffs' security practices," in *Proc. IEEE 43rd Annual Computer Software and Applications Conf. (COMPSAC)*, vol. 2, Jul. 2019, pp. 397–404.

Beyond the Horizon: A Meticulous Analysis of Clinical Decision-Making Practices

Bilal Saeed Raja[1], Sohail Asghar[2]
Department of Computer Science
COMSATS University Islamabad, Pakistan

Abstract—**Clinical advancements are one of the major outcomes of the technological phase shift of data sciences. The significance of information technology in medical sciences by utilizing the Clinical Decision Support System (CDSS) has opened the spillways of exponentially improved predictive models. Utilizing the latest norms of classification algorithms on clinical data are widely incorporated for prognostic assessments. Medical experts have to make decisions that are crucial in nature and if the research can develop a mechanism that assists them in evolving solid reasoning, infer the knowledge and clearly express their clinical decision by justifying their assertions made, it will be a win-win situation. However, this field of science is still an unknown world for clinicians despite the fact that the enormous amount of medical data cannot be exploited to its maximum without invoking the technological support. The objective of this research is to introduce the clinicians and policymakers of the medical domain with the renowned computer-based methodologies employed to construct a clinical decision support system. We expect that gaining the technical insight into the medical domain by the stakeholders will ensure commissioning the accurate and effective CDSS for improved healthcare delivery.**

Keywords*—*Decision support system; clinical decision support; classification; clustering; association rule mining; multi-objective evolutionary optimization

I. Introduction

Decision Support Systems (DSS) are the most studied areas of data sciences and their widespread adoption has earned standing in multiple domain such as education sector [48], customer relationship management [34], fraud detection in financial matters [33], detection of eavesdroppers and intruders in networks [39] and health care [6] including genetic programming [19]. Technical advancements have given a new dimension to the automated decision-making capability and health care is no exception to this due to its importance and critical nature as humans are the direct beneficiary or affectees of the outcome. Automation has given great ease to handle and query the huge volumes of data however it is an uphill task to process this data manually due to its size and non-homogeneity. Clinical Decision Support Systems (CDSS) are the most researched models as compared to any other science domain due to their effective decision-making capability. CDSS can be described as an “extraction of implicit potentially useful and novel information from medical data to improve accuracy, decrease time and cost, construct decision support system with the aim of health promotion” [15]. Due to the subjective nature decisions made by the domain experts may add inaccuracies that can be minimized by utilizing the technology to increase effective decision-making [67]. Furthermore, technological induction in health care has the strength to extract relationships within variables, identify the factors that may cause various risks and further impart fresh knowledge to yield befitting precision augmented with a convincing reasons [29]. This can only be achieved if the policymakers and clinicians have a deep insight into the technical strengths and weaknesses of computer-based methodologies being employed to construct CDSS. These decision support systems must be assessed for their performance evaluation so that quality care be provided with high precision value [36]. Data gathering and pre-processing is always a tough choice in different fields of science but it becomes much harder when it comes to health care due to its critical nature [11]. The medical data is not only numeric but may include images, temporal & combination of these which makes it a tough choice for automated decision-making. Medical data has huge volumes and its handling and organization in a manner that is understandable to the clinical decision support model is another daunting task. Therefore, the selection of appropriate classification schemes must be opted to handle multifaceted data for better throughput. In a broader scene, clinical and temporal data are the major categorizations in the health care domain [25]. This data is either quantitative (numeric), qualitative (non-numeric), temporal data (based on timestamp) or time-series based.

The major contribution of this research is to highlight the latest and most widely used decision-making models and techniques that enlightens the reader to earmark the pros and cons of these predictive models at the early stage of development. Furthermore, the research effort is to educate technical as well as non-technical domain specific audience that makes it an interesting academic resource by imparting better understanding of decision support methodologies that can be applied from analysis, design, development and deployment phases of CDSS. The SWOT analysis covers the socio-behavioural aspect of commissioning these models in medical domain.

Because of the non-homogeneous and varied characteristics of clinical data and keeping in view the importance of clinical decisions, selection of suitable and befitting methodology must be adopted by the stakeholders for assistive CDSS. An additional yet very crucial deliberation must be pursued on the selection of CDSS based on the operational characteristics that are as follows [16]:

- Trigger-based which are mainly used for drug prescription.
- Data Repository such as patient records & patient’s pathological results.
- Interventions systems that send alert messages to

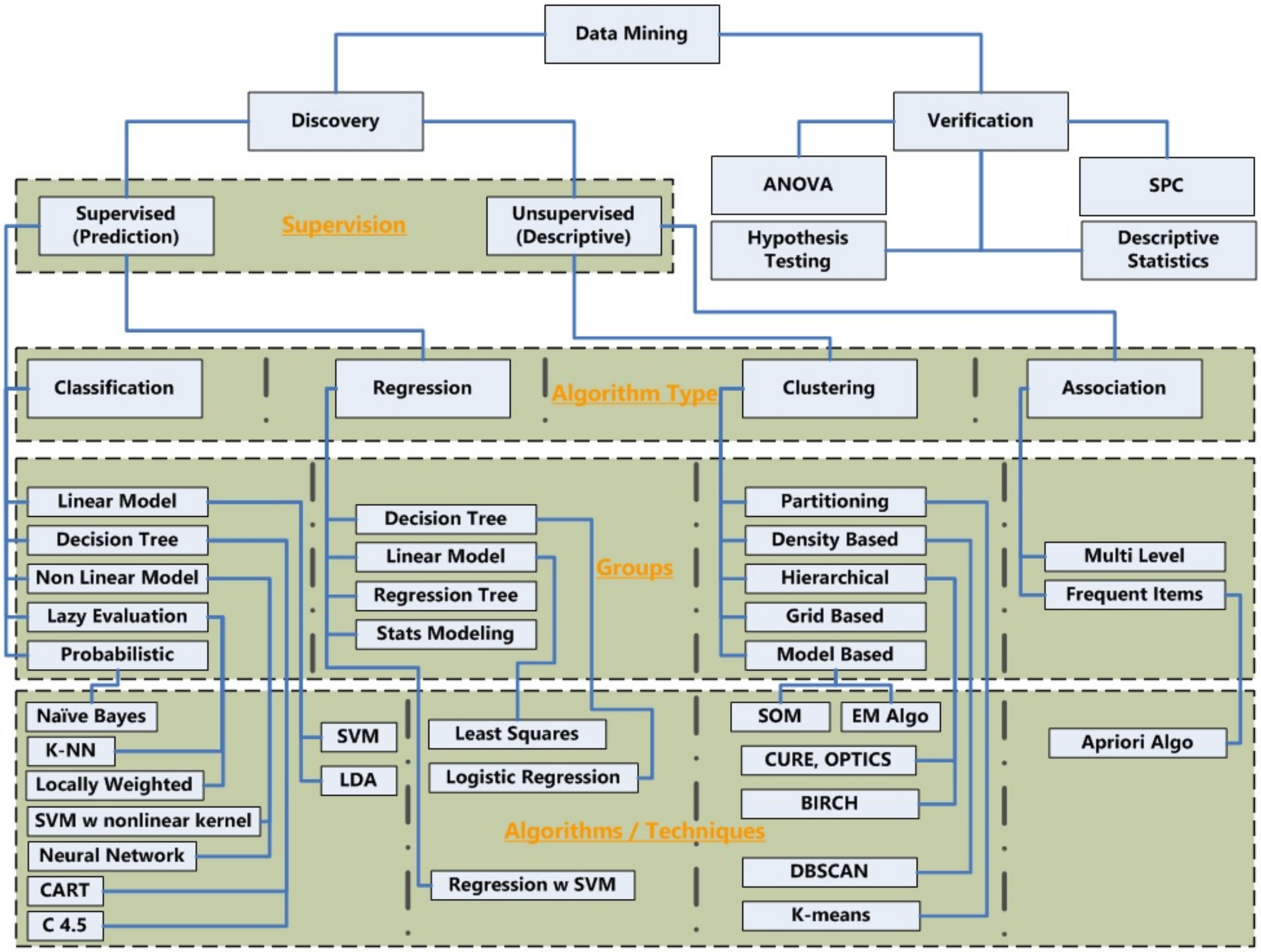


Fig. 1. Taxonomy of data mining algorithms

clinicians.

- Offered Choices in prescribing treatments and medication.

The automatic discovery of intelligent knowledge from the clinical decision support system commissions the major science of information technology that is data mining. Data mining is the branch of business intelligence that has two major branches one is the discovery and the other is verification. Disease classification and clinical decision support systems rely on data mining algorithms and these algorithms play a vital role in harnessing accurate and interpretable clinical decisions. Based on the available literature, taxonomy is presented in Figure 1.

II. Method of Literature Review

Substantial research has been carried out so far to acquire clinical decisions based on a multi-faceted web of knowledge. Most relevant and current research content was extracted from the web of knowledge. The search term used to search the literature was “clinical decision support systems, disease prediction, decision making, and classification algorithms”. The search revolved around the most relevant journals that were primarily focusing on the medical and bioinformatics domain (Artificial Intelligence in Medicine Elsevier, Computational and Structural Biotechnology Journal Elsevier. BMC Neurology Springer. Journal of Biomedical Informatics, Health Policy and HPT, Elsevier, etc.) from the period 2000 to date.

In quest of the resolution to the research questions in mind, relevant research articles on disease prediction/classification were collected from Google Scholar, IEEE Xplore, SpringerLink, ACM, DBLP ISI Web of Knowledge to name a few. Initially, we selected 209 research articles from which we studied and included the research contribution of 56 most relevant articles. The major focus was to include the latest research of the domain and with few exceptions, we confined our survey to the year 2010. This comprehensive review is carried out to shortlist and include the most recent and suitable articles of clinical sciences with high-quality research. The methodology of the literature review and count of research paper that are studied for compiling this research is presented in Figure 2.

To conduct this survey, a systematic literature review methodology is adopted. Furthermore, a brief description of the disease prediction/classification scheme is presented followed by strengths and weaknesses. This methodology is adopted by keeping one aspect in mind that the reader should have a clear understanding of the pros and cons of the classification

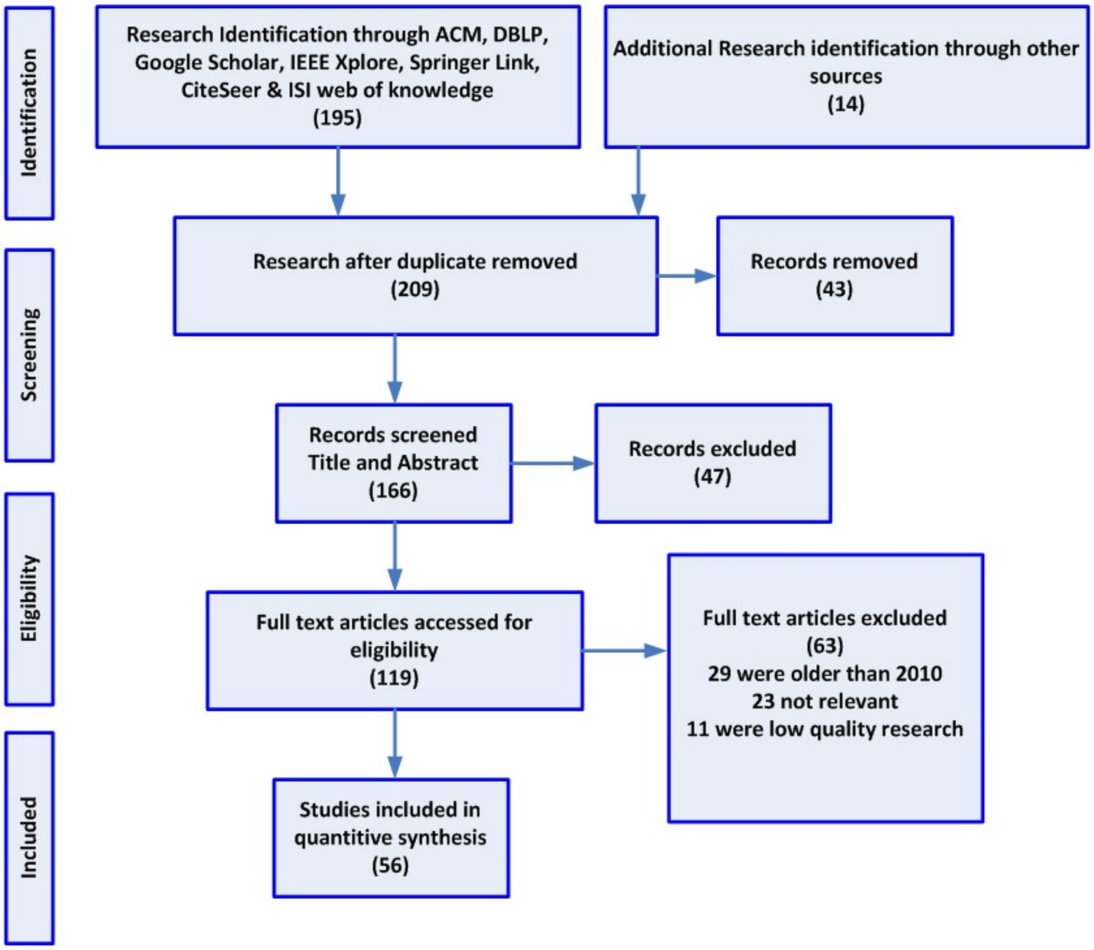

Identification
Screening
Eligibility
Included
Research Identification through ACM, DBLP, Google Scholar, IEEE Xplore, Springer Link, CiteSeer & ISI web of knowledge (195)
Additional Research identification through other sources (14)
Research after duplicate removed (209)
Records removed (43)
Records screened Title and Abstract (166)
Records excluded (47)
Full text articles accessed for eligibility (119)
Full text articles excluded (63)
29 were older than 2010
23 not relevant
11 were low quality research
Studies included in quantitive synthesis (56)

Fig. 2. Methodology of Literature Review

scheme. By doing so, policymakers can reckon very easily about the classification scheme being used during the development of CDSS.

A. *Research Questions*

Following questions were kept in mind while conducting this survey:

- What are the existing disease prediction techniques/algorithms?
- What kind of methodology is adopted in these techniques?
- What are the effect/repercussions of the findings regarding techniques is going to have while developing new CDSS?
- What are the major shortcomings/bottlenecks limitations of these techniques?
- What are the latest development methodologies in place to evolve effective CDSS?

In the upcoming sections, we will explain the various techniques and methodologies like machine learning, knowledge representation, text mining and multi-objective optimization techniques by which we can evolve an effective and efficient disease classification and clinical decision support system. It should also be considered that these methodologies can be used to evolve any type of CDSS and disease classification system elaborated in the research above. However, the selection of suitable and optimum classifier is a very important and crucial decision. Since all of the methodologies cannot be elaborated in detail so most relevant and famous are explained. The rest of the paper is organized to dovetail relevant methodologies in their categorical class and comprehensive yet most recent literature is organized in a fashion that suits almost all kinds of audience. This will not only help the technical team developing the CDSS but will also help the clinicians and decision-makers of health care facility to have prior wisdom in selecting appropriate solution at the very initial stage.

III. Machine Learning

Machine learning is primarily based on the learning curve extracted from the data [20]. The main objective applied in machine learning is data cleansing in which missing values are identified and outliers are removed. After the data cleansing process training of classifier is carried out on the dataset. The prime objective of this training is to make the classifier intelligent enough to perform requisite prediction and classification

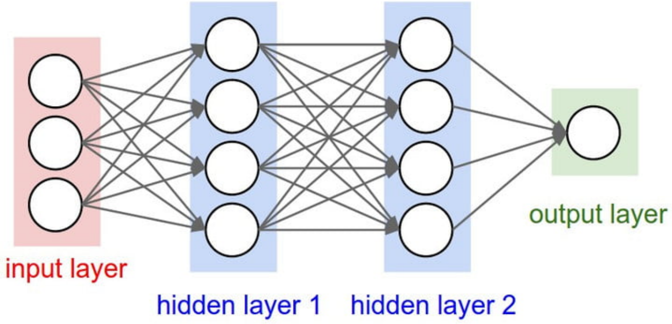


Fig. 3. Artificial Neural Network

on clinical data. The efficiency and accuracy of the classifier depend on the training parameters. The system needs to be trained in a manner that it should not be over or under trained. As there is no cookbook and one solution fit for all kinds of things, the training should be carried out wisely. There is every likelihood that for one clinical problem a classifier works well but for another, the same may fail to converge in a desirable manner. Once the training phase is over new dataset is given to the classifier for prediction and classification purposes. In the next subsection, we will elaborate on the most common type of machine learning methodologies.

A. Artificial Neural Networks (ANNs)

Artificial neural networks are mathematical models that mimic the human brain in the learning process [22]. To find the hidden pattern in the data and classification ANN uses a complex network of artificial neurons that are interconnected. ANN has an input layer and an output layer and most important of all, the hidden layer (one or more). The input layer takes some input and the output layer generates one or more than one outputs while the hidden layer trains itself to learn the hidden patterns from the data. The whole scenario is depicted in Figure 3.

The latest research on the validation and discovery of Alzheimer's disease was carried out in [68], where an artificial neural network pipeline was used for the said task. The research is a major contribution to the domain of dementia, which has no established cause and cure. Pipeline analyzes a public dataset with a continuous and categorical algorithm and further infers through network inference to generate novel markers for disease prediction. However, the research lacked performance and interpretability indices due to overfitting and noisy data. Another research [3] predicts the Dengue virus by implementing an artificial neural network on patient data for disease classification. Five performance objectives were employed including mean absolute error and accuracy. The research outnumbers the CART algorithm, however, improvement can be acquired if iterative non-parametric imputations are applied. ANN-based research was produced in [7] that used an artificial neural network (ANN) and Fuzzy analytic hierarchy process (Fuzzy_AHP) for an integrated decision support system. This hybrid approach got limitation for the automatic calculation of weights for the attributes which are used to train the ANN classifier.

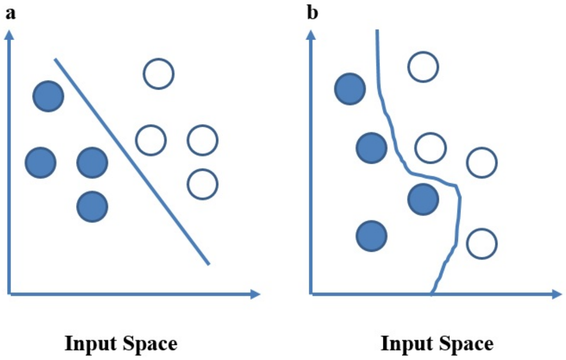


Fig. 4. Support Vector Machine

Artificial neural network like other computer methodologies has pros and cons. The ANN algorithm does not require rigorous training and is simple to implement with an ability to detect a non-linear relationship between variables. However, their hidden layer is not transparent to the user and has the tendency to over-fit when there are some noise and error in the data.

B. Support Vector Machine (SVM)

The basic theme behind the support vector machine resembles artificial neural networks, but SVM is much more powerful and effective when the classification of a complex task is required. The values from the dataset under classification are presented in the form of a point in a solution space [32]. The whole idea behind the support vector machine is that the data is projected in the solution space in the form of point and the point belongs to a designated class. For example, a dataset of disease classification will depict the presence or absence of a particular disease when the data is mapped on the same solution space. The data value will become part of one of the classes based on which side of the hyperplane it falls in. Figure 4 (a) depicts a linear classifier in which ANN can perform well, however, 4 (b) depicts a nonlinear classifier in which ANN cannot perform well and the best choice to solve such problem is support vector machine.

The novel research was presented in [31] where Parkinson's disorder was classified by diffusing proton spectroscopy, tensor imaging, and photometric data to obtain quantitative markers for the consumption of SVM. To achieve high accuracy a new graph-based technique was commissioned but the research lacked its full potential due to the fact that small data set with classifiable Parkinsonism was considered. Another research [56] classified healthy and Asthmatic individuals with the help of electronic nose through gas emission in exhaled breath by applying SVM in feature extraction and classification.

The system showed 78.8% accuracy when non-linear binary SVM was used instead of linear with a high rate of sensitivity. Adaptive SVM was commissioned for the first time to diagnose chest disease with high precision value by computing the appropriate bias term value to SVM [66]. In [58] a support vector machine and radial base function network structure is presented to predict the heart disease in the patients but a uni classifier is used in the research. [40] Proposed an automatic logistic regression and support vector machine-based prediction and classification for Parkinson's disease. The support vector machine along with Radial Basis Function (RBF) kernel achieves more accurate classification however results can further be improved if the ensemble framework is commissioned instead of uni-classifier. [62] Used various classification techniques for breast cancer prediction but the classification of support vector machine outnumbered all of them in performance. But this research has the same limitation of using a single classifier.

SVM's perform well when less training data is available and also suitable for multi-dimensional data which is unbalanced in nature. The algorithm hides the internal details of the working methodology but performs well by using mathematical modeling on the unstable dataset.

IV. Knowledge Representation

Knowledge Representation is principally based on vocabulary and in the clinical domain it generates a clinical knowledge descriptive language. This vocabulary is comprehensible and exploitable by the computer system and an amalgamation of an automated reasoning and inference system is formed [8]. In the medical domain patient's data can constitute vocabulary that can be automatically reasoned using clinical practicing guidelines to infer about patient's health.

A. Fuzzy Logic

Fuzzy logic is based on a probabilistic model that enumerates the human reasoning in approximate values [63]. The results generated are not just true and false but also enlightens the end-user with the degree of truth and false. Thus the results of fuzzy rules are more accurate because of their non-discrete nature. The fuzzy rule set is simple if-then-else rules that can be apprehended from natural language. Fuzzy logic has a Knowledge base that is built on the combination of Rule-base and database and a Fuzzy Inference Engine (FIE) as shown in Figure 5. Following lines express the rules for the fuzzy system:

$$\begin{aligned} R_i : \textbf{IF } x_1 \ is \ F_1 \ \textbf{\textit{AND}} \ x_2 \ is \ F_2 \\ \textbf{AND } x_m \ is \ F_m \ \textbf{\textit{THEN}} \ y = C_i \\ \textbf{\textit{WHERE i=1,2,3,4, \ldots,M}} \end{aligned} \tag{1}$$

Very important research carried out in [35] integrates the fuzzy standard additive model (SAM) with a genetic algorithm (GA), called GSAM was adopted and wavelet transformation was employed to extract discriminative features for high-dimensional datasets. GA was used to optimize the number of fuzzy rules before supervised learning. GSAM dominates PNN, SVM & ANFIS on classification accuracy but has disadvantage regarding computational cost compared to these competitive methods. Fuzzy logic was amalgamated with the modular neural network to diagnose the risks of hypertension [30]. The age, risk, and blood pressure were a major deriving force of the research and modular neural network utilized three modules, one looking after heart rate and the remaining two looking after systolic and diastolic readings. Two fuzzy inferences were incorporated for heart rate and night profiling of the subject. High accuracy and interpretability can be achieved if meta-heuristic models are applied along with type 2 fuzzy inferences. The befitting model of fuzzy logic was presented with a name PreFurGe that has the capability to predict the chances of invitro fertilization to help gynecologists and embryologist [14]. The model can be further improved by generating better rules and by ingesting GA. A low cost and accurate framework is presented in [46] that used a matrix-oriented fuzzy rule-based predictive model for heart disease. In [60] a learning membership function that uses neural networks in addition to fuzzy logic systems is proposed. However, this framework lacked desirable accuracy that would have been achieved if the multi-layer approach can be used instead of a single layer approach.

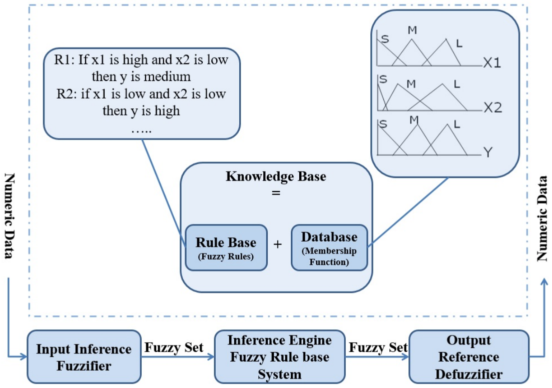


Fig. 5. Fuzzy Inference Engine

Fuzzy logic is an excellent representation of providing linguistic variables into computing with the liberty of dealing non-discrete/non-linear and imprecise problems. This power of fuzzy logic makes it a perfect choice for the predictive systems that require high accuracy. However, there is a handicap of tuning the membership function and other parameters manually that might cause inconvenience in terms of time and effort. Fuzzy logic cannot scale well for large problem set but still are a lucrative choice for medical domain predictive models.

V. Text Mining

Text mining is an emerging field of decision mining that adopts statistics, machine learning and linguistic techniques to extract high-quality information from unstructured data repositories [18]. Text mining has shown some remarkable results in medical data classification as the data generated by the medical domain is diversified and carries huge volumes. The most important text mining methodologies include information retrieval and natural language processing. The next subsection explains natural language processing in brief.

A. Natural Language Processing (NLP)

The large corpus of medical data has emerged as a major problem area for the domain users. The free context of medical data, medical notes and reports have given a new dimension to the text mining paradigm. The information retrieval represents document under scrutiny as a collection of predefined words however natural language processing takes this a step forward

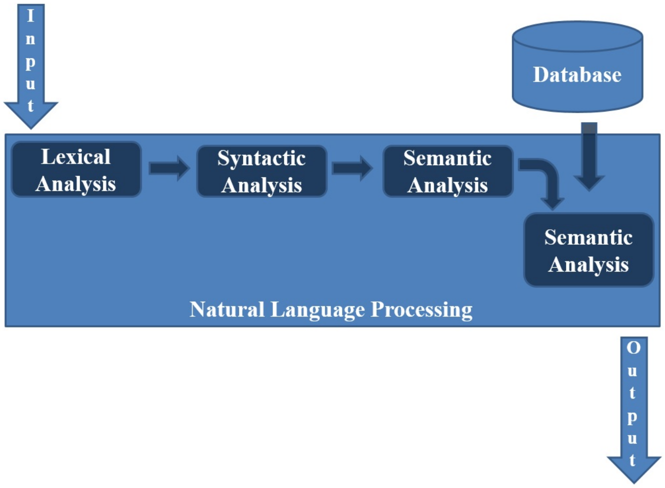


Fig. 6. Natural Language Processing

and generates meaning from natural language and human beings and is depicted in Figure 6 [21].

Latest research presented MetaMap that aimed to reduce the error rate by identifying eligibility for Intravenous Thrombolytic Therapy (IVT) in stroke patients using natural language processing [57]. MetaMap handicapped itself in the generalization of outcomes due to a small sample size and tend to acquire long processing time which makes it a hard choice for real-time large datasets. NLP was used in another research for Healthcare-Associated Infections (HAI) monitoring [61]. The major objectives were sensitivity and specificity. The major areas of medical sciences for this study were digestive, neuro and orthopedic surgery including adult intensive care. The performance factor can further be improved if semantics is applied with expert rules. Similarly, a similar kind of study was conducted on the Mayo Clinic health record for predetermined asthma criteria using NLP [65]. Natural language processing is one of the most widely used methodologies of artificial systems that involves natural language for the representation of knowledge. The main advantage of using NLP is that it is highly expressive and virtually can articulate any real-world situation, emotion, ideas, and pictures. The human intuition by default understands the semantics and vocabulary. NLP perform well when the domain in which it is applied has clarity and is narrowed to the deep understanding. However, NLP tends to have difficulties in elaborating syntax and semantics both when the domain is divergent. If not properly implied, NLP has a tendency of having little uniformity in the sentences that make the grammar ambiguous. NLP is widely used in medical decision-making and performs well where its precise and limited scope in establishing decision making or predictive analysis. Conventional disease classification algorithms perform well when the problem set is simple. With the advancement in the information technology and business intelligence methodologies, their era is converging to more complex and improved variants of disease classifications. A tabular representation of some of the conventional classification algorithms and techniques is presented in Table 1 below with their major characteristics and pros and cons elaborated in detail.

VI. Multi-Objective Optimization Techniques

In CDSS exclusive, potentially useful and original information is extracted from data set to enhance interpretability, accuracy, decrease time and cost [17]. The extracted knowledge is represented in legitimate, valuable and reasonable structures, trend and patterns. Various techniques of classification, clustering, association rule mining, and forecast deliver pronounced help to experts in the earlier detection and diagnosis of the disease [5]. The prompt evolution of data analysis techniques has empowered the production of CDSS to be much more tolerable than ever before. To solve mono and multi-objective problems, evolutionary algorithms have been evolved and their basic design procedure is depicted in Figure 7 [47]. Evolutionary algorithms are designed to allow survival of fittest theory in which the algorithm initiates itself by the selection of the population randomly. Then the method uses a sequence of generations, in which the best design point in a selected population is taken as most fit and is allowed to reproduce. Mathematical modeling is employed by simulating selection, breeding and mutation process. Multi-objective optimization evolutionary algorithms (MOEAs) are efficient with problems of two or three objectives. Evolutionary multi-objective (EMO) algorithms such as Non-Dominant Sorting Genetic Algorithm-II (NSGA-II), MOEA/D, GDE3, SPEA2, and others, have displayed remarkable results in addressing many scientific application problems of real-time nature, engineering and economics problems that pivotally cater two to five objectives. Nevertheless, if a solution to solve a greater number of objectives problem (termed as the multi-objective optimization that usually includes more than three objectives), most of the EMO algorithms fall short to unearth well spread & well converged non-dominated solutions due to the decrease in fitness evaluation function's selection pressure that results in compromised accuracy and interpretability.

Multi-objective optimization was used by employing a variant of classical SVM that is sequential minimal optimization (SMO). The base classifier was used in collaboration with an evolutionary algorithm Elephant Herding Optimization, Decomposition based MOEA and NSGA-II to constitute a framework named CEHO [44]. The system can be modified by invoking ELM and deep neural networks. [17] Presented a multi-objective optimization approach that is based on genetic fuzzy logic. The major objectives used in the research were accuracy and interpretability. However, the model lacked in determining the final non-dominated solutions with a high spread and well-balanced distribution in the objective space. It is evinced from the research resources that NSGA algorithms display better performance when the operators are selected optimally, such as random polynomial mutation to produce the offspring, differential evolution and simulated binary crossover.

Multi-objective optimization techniques have the capability and power of simultaneously dealing with a set of candidate solutions. Unlike to their counterpart, exploration of quite a few members of Pareto optimal set in a single run. The only point of concern is that they require a large number of iterative

TABLE I. COMPARISON OF CONVENTIONAL CLASSIFICATION TECHNIQUES

Sr.	Algorithm	Advantage	Disadvantage	Characteristics
		Classification		
1.	Decision Tree [41]	Easily understandable and resilient to replication and data with noise	Not good with inconsistent data & accuracy is also compromised	Recursive algorithm, uses greedy approach.
2.	Artificial Neural Network [50]	Resilient to data replication with ability to handle complex relationships	Susceptible to the data anomalies & hides internal details	Involves more than one layer with a precondition of holding one layer bare minimum.
3.	Rule-based Algorithms [17]	Easily understandable with resilience to inconsistent data	Compromised accuracy	Based on if-then-else rules
4.	Support Vector Machine [69]	Ideal when less training data is available and also suitable for multi-dimensional data	Hides internal details & require parameters	Unorthodox approach and involves mathematical calculations
5.	Naïve Bayes Algorithm [38]	Resilient to all sorts of data abnormalities	Compromise in accuracy & require earlier probability	Involves statistical calculations
6.	K- Nearest Neighbor [53]	Straightforward algorithm with utmost elasticity	Susceptible to replication and noise.	Lethargic approach with prediction results based on local data
		Clustering		
7.	K- Means Algorithms [27]	Straightforward algorithm with utmost processing speed	Not feasible for heterogeneous data and susceptible to data noise	Algorithm uses prototype based results
8.	Hierarchical Method [10]	No parameters required and less prone to initial value	Susceptible to data noise and involves complexity both in terms of space and time	Bottom up approach and uses graph theory
9.	Density-Based Spatial Clustering of Applications with Noise [1]	Noise resilient & ability to handle random density and size.	Involves complexity both in terms of space and time	Approach based on density
10.	Fuzzy C-Means Algorithms [49]	Straightforward algorithm with utmost processing speed	Not feasible for heterogeneous data and susceptible to data noise	Algorithm uses prototype based results
		Association Rule Mining		
11.	Apriori Algorithms [45]	Straightforward algorithm with wide acceptance	Involves complexity both in terms of input/output and time	It is a recursive approach that uses prior knowledge
12.	Dynamic Itemset Counting [55]	Ability to handle input / output complexity	Susceptible to non-heterogeneous data	Dynamic algorithm discovering lost patterns
13.	Direct Hashing and Pruning [42]	Ability to handle candidate patterns count	Cannot handle hash table anomalies	Based on the concept of hashing
14.	E-CLAT Algorithm [26]	Less complex with input /output	Cannot handle large data sets and complex with respect to space	Uses lattice theory and it's a bottom-up approach
15.	D – CLUB Algorithm [26]	Automatically adjusts to the situation and efficiently handle time and space complexity	Compromised accuracy	Best suited for no centralized databases and can support parallel processing

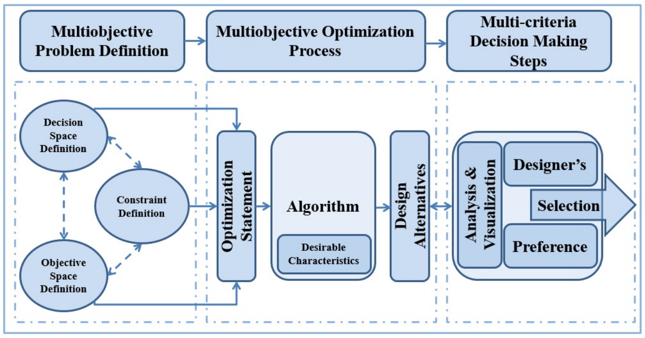

Multiobjective Problem Definition
Multiobjective Optimization Process
Multi-criteria Decision Making Steps
Decision Space Definition
Constraint Definition
Objective Space Definition
Optimization Statement
Algorithm
Desirable Characteristics
Design Alternatives
Analysis & Visualization
Designer's
Selection
Preference

Fig. 7. Multi-objective Optimizing Technique

runs that require high computational effort to generate Pareto front.

In [9] an Evolutionary Algorithms - EAs are used to propose an optimal searching feature subset. This is achieved by introducing a penalty term, to minimize feature count in the selected subset without affecting classification accuracy. In order to achieve the proposed objective, various Evolutionary Computational Algorithms (ECAs) are applied using penalty-based fitness function, which evaluates the next optimal feature subset. ECAs end up with high accuracy for higher-dimensional datasets. The proposed work has been tested using dimensions up to 10,000. However, for feature subset selection, ECAs lack in reducing residual features from final selection and they are costly too. In [28] development of a robust optimized machine learning - ML system is presented. The aim is to improve risk stratification accuracy by replacing outliers with median configuration, which is based on assumption. Concrete classes were used to accurately classify diabetes in patients. The proposed machine learning system is designed, developed and evaluated using a feature selection strategy. It is then combined with several kinds of classifiers. The proposed approach has shown stable and reliable results. It also improved the performance of existing systems by replacing outliers with median computations. With this approach, the results have become more accurate. However, the system works for only Indian diabetic medical data and its classification. The final solution is costly and time consuming for medical specialists as well as for patients. In [43] the fundamental purpose of the study is to enhance accuracy, sensitivity and specificity rates on Z-Alizadeh Sani dataset. This work proposes a hybrid method, which is highly accurate to diagnose coronary artery disease. This study achieves high performance as far as neural networks are concerned. This performance enhancement is attained by applying Practical Swarm Optimization (PSO). The proposed study helps to reduce the cost considerably along with no major side effects. Thus, coronary artery disease is detected without the need for invasive diagnostic methods, using clinical data. With the help of this approach, multi-objective of accuracy, specificity and accuracy rates on Z-Alizadeh Sani dataset are evaluated. Keeping in mind the dynamics of the problem, designing of a proper network can be a very tough task because of its dependency on problem dynamics. In

[12] Bio-inspired Multi-objective algorithm is presented. The process of gene selection is carried out using microarray data classification. Refined formulations of BA have been used along with MO search techniques and specialized operators. Variable selection is also done using binary domain called MOBBA-LS. The proposed algorithm called BA produces best subsets with lesser number of genes with the highest accuracy. These genes have excellent relevancy. Proposed work showed low performance, which needs to be enhanced. The other drawback is the increased time-complexity. Another study is based on the notion to build a simple classifier using multiclass classification strategy, which has integrated various multi-objective namely feature selection as well as its construction. It further models the intelligibility objective into a distance-based classifier [23]. The proposed approach optimizes data models by using genetic programming. This model named (M4GP) is based on an innovative stack-based program method, which makes the multi-dimensional solutions simpler to construct. This methodology gives M4GP an edge when a comparison is made with M2GP and M3GP (both of these models applied tree-based structure). The results of this model show that the final solution is interpretable and more accurate. M4GP also offers an efficient and flexible solution for providing accurate classifiers. Moreover, it also yields the best classification for small dimensional operations. Since M4GP works on the population-related domain, therefore it incurs a higher computational price. In [2] the main theme of this paper is to deal with such MO problems having high uncertainty. This uncertainty is represented by triangular fuzzy numbers. It involves solving the problem using fuzziness propagation to fitness functions. The proposed approach consists of a fuzzy Pareto dominant solution and then to apply EAs to reach to a solution. One of the advantages is that it uses transformations of other shapes using operations like projections, linguistic classifiers, and compositions. TFN can be deduced using the above operations. Regardless of a lot of suggested methodologies, there are still many open issues to be addressed for this domain. For instance, there is no real /close remedy to deal with uncertainty, which prevails as the core aspect in the area of multi-objective problem domains. In [4] the main purpose of this research is to present a model that is able to extract health indicators HI's. The aim is to keep track of various signals during operation. Thus, to study the component degradation during this process /operation. The proposed idea is derived from the usage of the feature extraction method. The selection of Health Indicators has three steps; firstly, feature extraction is done. Secondly, the selection is done and at the third stage, fusion is done by applying the BDE algorithm (multi-objective Binary Differential Evolution). This method has produced much satisfactory HI's. It has shown more satisfactory health indicators than found in other research studies. A set back of the proposed study is that unsatisfactory prognostic performances can result due to RUL. In [59] the main aim of this research is to study various entropy-based design optimization schemes and attempt to decrease the gap between them. This proposed research study is based on the notion of join entropy schemes that are independent. It is further applied to a real-life problem related to the water distribution network. Various stages include maximizing the joint entropy along with applying a penalty-free genetic algorithm with three objectives. The benefits of using such a methodology is that it presents a relatively simpler and easier way to be assimilated using multi-objective optimization algorithms. Another aspect is its efficiency in generating results. In short, this study shows a balance between computational budgeting and flexibility. One of the major drawbacks is that a significant increase in the available feasible solutions is gained when compared to previous research studies. This increase in entropy values made infeasible solutions to be vague and distorted. The main theme of another research is to present an innovative model that is able to enlighten occurrence of asthma, and also detect two markers of allergy namely IgE antibody against common allergens, and skin prick test positivity for common allergens (SPT) [64]. The technique was based on MGGP (multi-objective grammar-based genetic programming). The medical dataset contains details of nutritional, psychosocial, socioeconomics, atmospheric and infectious factors gathered from children who were part of the study/process. MGGP model achieved higher accuracy and results were also easy to interpret. The performance of MGGP model for each iteration takes 28.1 h along with its limitation/absence to offer parallelism capability for proposed work for now. In [51] presents the optimum aspects of SPIF parameters for titanium denture plate. The present paper attempts to measure the likelihood of generating customized Commercially Pure Titanium Grade 1 (CP-Ti Gr.1) denture plate with reasonable accuracy. The proposed working aims to control some process parameters that affect the quality of the final product with the prime objective of geometrical accuracy. The proposed strategy of optimization of multi-objective is based on numerical simulation using a Multi-Objective Genetic Algorithm and the Global Optimum Determination. This is done by Linking and Interchanging Kindred Evaluators algorithm to find the optimum solutions. Minimizing sheet thickness, its ultimate along with increasing forming force were main objectives. Achieving robustness for the selection of optimum factors in SPIF is the core advantage. It also results in improving geometric errors, especially in the base area. However, a large number of errors in the part wall section needs to be addressed in further development specifically to validate the quality of the surface after forming sections. Another research applied age prediction using neuroimaging with the help of ML schemes. The core theme of this study is to improve accuracy for age prediction by Bayesian parameter optimization pertaining to age prediction [24]. Bayesian optimization is done in an iterative manner to check the sample space for various parameters to achieve accuracy in the resultant space. This approach improves the notion to distinguish young and old brain. Neuroimaging data is the basis for the whole idea. The Bayesian optimizations achieve optimum voxel size thus improves performance. Keeping in view the complexity in neuroscience because of its multi-disciplinary nature, the research analysts may not hold expertise in every area. Thus, further unbiased optimization parameters may give more benefit.

Due to their precise predictive nature, these algorithms are utilized in the medical domain to evolve CDSS by amalgamating them with fuzzy logic.

VII. Challenges and Practices

As the rule of thumb, all domains of sciences have their own challenges and rigorous efforts are always underway by the researchers of respective domains to rectify those challenges. Like other domains, disease classification and CDSS has a number of challenges, in particular, some of them are enlisted below:

- ***Critical heterogeneous data [25]:*** The data available in the medical domain is possibly the most difficult in nature including numeric, alphanumeric, pictorial, and continuous, etc forms. To handle such data there is a requirement of a comprehensive data cleansing and normalization mechanism to eradicate the heterogeneity, missing values, and outliers.
- ***Clinical Data Privacy [54]:*** Privacy and secrecy of the clinical data are one of the major concern in health care. Most of the time this data is reused by various stakeholders for the betterment of health care but the concern of this data being misused can not be ruled out. The general attitude of the stakeholders toward its reuse must be analyzed carefully before the sharing of this critical information to mitigate this threat.
- ***Compromise on Identification of risk factors [16]:*** Most of the clinicians and technical hands involved in the development of CDSS neglect the identification of risk factors in the earlier design. The risk factors are generic and some of them are specific to the domain (a type of disease, type of CDSS, etc). The earlier detection, identification, and eradication of all sorts of risk factor makes the CDSS more robust and resilient to any change and advancement in technology.
- ***Incremental sensitivity and specificity [5]:*** Sensitivity is the true positive rate and specificity is the true negative rate. Both attributes have an important role in the construction of CDSS. It is desired that the CDSS should increase the sensitivity and specificity incrementally by training and fitting its model.
- ***Effectiveness [5]:*** The CDSS can only be effective if it is easy to use, generate accurate predictions, the decisions are interpretable with high sensitivity and specificity and has a low computational cost. The CDSS should be designed by keeping all these objectives to an acceptable level.
- ***Scalability and Adaptability [5]:*** CDSS should be developed with the ideology of being scalable and adaptable; as the medical domain is an evolving field of science a tunnel approach in the design of CDSS leads the clinicians and policymakers toward an ineffective CDSS design. It is always recommended that the CDSS should be constructed by keeping in view the scalability and adaptability perspective in mind.
- ***Accuracy & Interpretability [17]:*** Accuracy is the ability to provide a clinical decision process that generates an outcome with high precision value. However, Interpretability is the feature that provides the user with condensed and comprehensible enlightenment and explanation with reason, of the proposed decision. A major concern of the medical domain is to evolve CDSS that can improve the accuracy and interpretability simultaneously and considerably reduce the computational cost.
- ***Different Knowledge Areas [37]:*** CDSS's are systems that facilitate the clinicians but are evolved from the amalgamations of various knowledge areas that include computer sciences, statistics, mathematics, bioinformatics, medical sciences, etc. To work on CDSS require multifaceted knowledge of different domains of sciences. That is the reason that clinicians and policymakers of the medical domain working on the development of CDSS find it a difficult choice.
- ***Extraction of the relationship between variables [13]:*** It is considered essential in the construction of CDSS that all the variables along with their relationship should contribute. The relationship of the variables makes the system much more comprehensible and gives the utmost interpretability of the clinical decisions.
- ***The requirement of multiple objective decisions [13]:*** Clinical decisions are critical in nature and require multiple objectives to accomplish accurate predictive models. Most of the CDSS have their focus on achieving accuracy, however, over the period of time, multiple objectives like interpretability, specificity, sensitivity, and computational complexity have evolved as a major deriving force in accessing a CDSS. So the inclusions of multi-objective decision support systems have wider acceptability and are recent research areas of clinical decision making and disease prediction.
- ***Transfer Learning [52]:*** CDSS are primarily based on various decision making computer methodologies that require time to learn different pattern over time, it is, therefore, desired that the CDSS should be capable of transferring this inferred knowledge to its inherited system.

VIII. SWOT Analysis

SWOT analysis refers to the Strengths, Weaknesses, Opportunities, and Threats allied to a solution. Our survey opines that invoking a CDSS has certain pros and cons allied to it that can be sorted out for the better results in healthcare. In the upcoming section we will explain all the above mentioned important analysis factors in more detail and the same is depicted in Figure 8 below for consideration while opting CDSS invocation:

A. Strengths

CDSS has shown remarkable improvement in health care by diagnosing complex imaging results which remained unidentifiable by the human eye. The same is true for numeric data values. The major strength lies in their ever-evolving/training nature that makes them less prone to errors. As automated systems are free from sentiments/behavioral aspect that makes their performance to remain at an optimum threshold level. The results generated are not only accurate but they are comprehensible by the stakeholders. Automation brings scalability and transfers learning opportunities as part of a package that helps in an ongoing evolution of even better systems.

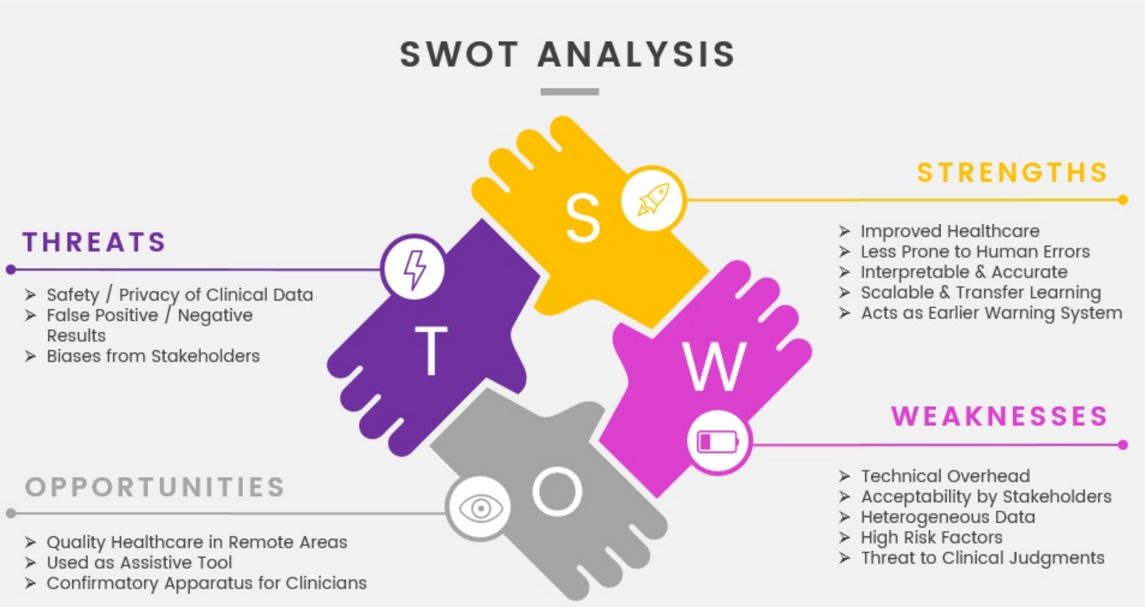

SWOT ANALYSIS
S
W
O
T
STRENGTHS
➢ Improved Healthcare
➢ Less Prone to Human Errors
➢ Interpretable & Accurate
➢ Scalable & Transfer Learning
➢ Acts as Earlier Warning System
THREATS
➢ Safety / Privacy of Clinical Data
➢ False Positive / Negative Results
➢ Biases from Stakeholders
WEAKNESSES
➢ Technical Overhead
➢ Acceptability by Stakeholders
➢ Heterogeneous Data
➢ High Risk Factors
➢ Threat to Clinical Judgments
OPPORTUNITIES
➢ Quality Healthcare in Remote Areas
➢ Used as Assistive Tool
➢ Confirmatory Apparatus for Clinicians

Fig. 8. SWOT Analysis

B. Weaknesses

Most of the studies discussing the commissioning of CDSS highlight the issue of acceptability from the stakeholders as they consider it to be an overhead instead of helping hand in assisting their assertions/decisions. Clinicians consider the decision support systems to be an evaluation apparatus for their decisions and prognosis. Clinicians have to perform diagnostic tasks that are time-consuming. Another very important factor that makes these systems complicated is the heterogeneous data. There is every likelihood that the outliers, missing values, and typo errors may cause daunting results that can lead to confusion and further investigation that is an overhead in terms of time and cost.

C. Opportunities

Revolution in technology brings opportunities along-with and CDSS has given great ease of use to hospital administration, health care policymakers and clinicians. The major opportunity comes in the remote areas where specialized healthcare resources in terms of man and material both are not available. These systems act as an assistive tool in justifying and conforming to the decisions made by the doctors. Hospital administration and policymakers can also monitor the true positive/negative and false positive/negative diagnosis very easily and may take measures to improve healthcare.

D. Threats

One of the major considerations that need to be catered to is the safety, privacy, and secrecy of clinical data. Clinical data contains health information of masses that can not be revealed to someone irrelevant. As the adaptability of these systems and data allied to it is growing exponentially, data safety has become a vital concern. Similarly, the behavioral aspect should also be taken as a threat as most of the mindsets don't allow change to the conventional methodologies in practice. CDSS being a relatively new methodology in the health domain may face these biases from the stakeholders.

IX. Conclusion

CDSS is expected to improve medical healthcare quality by assisting the doctors in making clinical decisions. The healthcare data classification mechanism assists medical experts in the early identification and management of medical malfunctioning and symptoms arisen in the patient. This substantial contribution has a pivotal role in the quality enhancement of healthcare by assisting doctors in decision-making. The contribution of decision support in general fields has shown great acceptance but specific to the medical domain and disease classification their acceptance is still scarce. A large amount of investigation and research in evolving an effective CDSS for the medical domain and disease classification are studied, the mechanisms used to service this research area include classification, clustering, ensemble, artificial neural networks, evolutionary algorithms like genetic algorithms (GA) and deep learning. The major challenge of CDSS is to attain the utmost accuracy, which has the ability to provide a clinical decision process that generates an outcome with high precision value. The aim of this review is to appraise the clinician and policymakers of the medical domain to gain the technical knowledge of renowned computer-based methodologies that are employed to construct the decision support models of medical domains. By doing so, the clinician and policymakers can give sensible and informed input during the analysis, design, development and deployment stages of CDSS. In the implementation stage, the clinicians could provide guidance on which of the methodology described above yields better results based on the clinical problem, the type of CDSS required and the dataset. With the deep insight on the methodologies surveyed in this research, the clinicians and policymakers will have the confidence to advocate the importance and significance of the system that will result in improved medical care and quality in the validation phase. Another depiction of relatively very recent and relevant multi-objective methodologies in evolving CDSS is highlighted in this research. It is further opined that multi-objective optimization techniques have shown remarkable results especially in the field of medical decision-making and is gaining a fast reputation for their accurate and interpretable results.

REFERENCES

[1] Amineh Amini, Teh Ying Wah, and Hadi Saboohi. On density-based data streams clustering algorithms: A survey. *Journal of Computer Science and Technology*, 29(1):116–141, 2014.

[2] Oumayma Bahri, El-Ghazali Talbi, and Nahla Ben Amor. A generic fuzzy approach for multi-objective optimization under uncertainty. *Swarm and Evolutionary Computation*, 40:166–183, 2018.

[3] K Balasaravanan and M Prakash. Detection of dengue disease using artificial neural network based classification technique. *International Journal of Engineering & Technology*, 7(1.3):13–15, 2018.

[4] P Baraldi, G Bonfanti, and E Zio. Differential evolution-based multi-objective optimization for the definition of a health indicator for fault diagnostics and prognostics. *Mechanical Systems and Signal Processing*, 102:382–400, 2018.

[5] Saba Bashir, Usman Qamar, and Farhan Hassan Khan. Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble. *Quality & Quantity*, 49(5):2061–2076, 2015.

[6] Riccardo Bellazzi and Blaz Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2):81–97, 2008.

[7] Thomas Bernd, Markus Kleutges, and Andreas Kroll. Nonlinear black box modelling–fuzzy networks versus neural networks. *Neural computing & applications*, 8(2):151–162, 1999.

[8] RJ Brachman, HJ Levesque, and M Pagnucco. Knowledge representation and reasoning: Knowledge representation and reasoning, 2004.

[9] Basabi Chakraborty and Atsushi Kawamura. A new penalty-based wrapper fitness function for feature subset selection with evolutionary algorithms. *Journal of Information and Telecommunication*, 2(2):163–180, 2018.

[10] Kerina Blessmore Chimwayi, Noorie Haris, Ronnie D Caytiles, and N Ch SN Iyengar. Risk level prediction of chronic kidney disease using neuro-fuzzy and hierarchical clustering algorithm (s). 2017.

[11] Krzysztof J Cios and G William Moore. Uniqueness of medical data mining. *Artificial intelligence in medicine*, 26(1-2):1–24, 2002.

[12] M Dashtban, Mohammadali Balafar, and Prashanth Suravajhala. Gene selection for tumor classification using a novel bio-inspired multi-objective approach. *Genomics*, 110(1):10–17, 2018.

[13] Kalyanmoy Deb. Multi-objective optimisation using evolutionary algorithms: an introduction. In *Multi-objective evolutionary optimisation for product design and manufacturing*, pages 3–34. Springer, 2011.

[14] José Gonzalez Enríquez, V Cid, N Muntaner, J Aroba, José Navarro, FJ Domínguez-Mayo, María José Escalona, and I Ramos. Behavior patterns in hormonal treatments using fuzzy logic models. *Soft Computing*, 22(1):79–90, 2018.

[15] Nura Esfandiari, Mohammad Reza Babavalian, Amir-Masoud Eftekhari Moghadam, and Vahid Kashani Tabar. Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, 41(9):4434–4463, 2014.

[16] Paolo Fraccaro, Panagiotis Plastiras, Chiara Dentone, Antonio Di Biagio, Peter Weller, et al. Behind the screens: Clinical decision support methodologies–a review. *Health Policy and Technology*, 4(1):29–38, 2015.

[17] Marian B Gorzałczany and Filip Rudziński. Interpretable and accurate medical data classification–a multi-objective genetic-fuzzy optimization approach. *Expert Systems with Applications*, 71:26–39, 2017.

[18] S Inzalkar and Jai Sharma. A survey on text mining-techniques and application. *International Journal of Research In Science & Engineering*, 24:1–14, 2015.

[19] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: a survey. *IEEE Transactions on knowledge and data engineering*, 16(11):1370–1386, 2004.

[20] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[21] Wahab Khan, Ali Daud, Jamal A Nasir, and Tehmina Amjad. A survey on the state-of-the-art machine learning models in the context of nlp. *Kuwait journal of Science*, 43(4), 2016.

[22] Miroslav Kubat. Artificial neural networks. In *An Introduction to Machine Learning*, pages 91–111. Springer, 2015.

[23] William La Cava, Sara Silva, Kourosh Danai, Lee Spector, Leonardo Vanneschi, and Jason H Moore. Multidimensional genetic programming for multiclass classification. *Swarm and evolutionary computation*, 44:260–272, 2019.

[24] Jenessa Lancaster, Romy Lorenz, Rob Leech, and James H Cole. Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction. *Frontiers in aging neuroscience*, 10:28, 2018.

[25] Nada Lavrač. Selected techniques for data mining in medicine. *Artificial intelligence in medicine*, 16(1):3–23, 1999.

[26] Jing-Song Li, Yi-Fan Zhang, and Yu Tian. Medical big data analysis in hospital information system. *Big Data on Real-World Applications*, page 65, 2016.

[27] A Malav, K Kadam, and P Kamat. Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy. *International Journal of Engineering and Technology*, 9(4):3081–3082, 2017.

[28] Md Maniruzzaman, Md Jahanur Rahman, Md Al-MehediHasan, Harman S Suri, Md Menhazul Abedin, Ayman El-Baz, and Jasjit S Suri. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. *Journal of medical systems*, 42(5):92, 2018.

[29] Gunjan Mansingh, Kweku-Muata Osei-Bryson, and Han Reichgelt. Using ontologies to facilitate post-processing of association rules by domain experts. *Information Sciences*, 181(3):419–434, 2011.

[30] Patricia Melin, Ivette Miramontes, and German Prado-Arechiga. A hybrid model based on modular neural networks and fuzzy systems for classification of blood pressure and hypertension risk diagnosis. *Expert Systems with Applications*, 107:146–164, 2018.

[31] Rita Morisi, David Neil Manners, Giorgio Gnecco, Nico Lanconelli, Claudia Testa, Stefania Evangelisti, Lia Talozzi, Laura Ludovica Gramegna, Claudio Bianchini, Giovanna Calandra-Buonaura, et al. Multi-class parkinsonian disorders classification with quantitative mr markers and graph-based features using support vector machines. *Parkinsonism & related disorders*, 47:64–70, 2018.

[32] Janmenjoy Nayak, Bighnaraj Naik, and H Behera. A comprehensive survey on support vector machine in data mining tasks: applications & challenges. *International Journal of Database Theory and Application*, 8(1):169–186, 2015.

[33] Eric WT Ngai, Yong Hu, YH Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569, 2011.

[34] Eric WT Ngai, Li Xiu, and Dorothy CK Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602, 2009.

[35] Thanh Nguyen, Abbas Khosravi, Douglas Creighton, and Saeid Nahavandi. Classification of healthcare data using genetic fuzzy logic system and wavelets. *Expert Systems with Applications*, 42(4):2184–2197, 2015.

[36] Federico Paoli, Ingrid Schmidt, Olivia Wigzell, and Andrzej Ryś. An eu approach to health system performance assessment: Building trust and learning from each other. *Health Policy*, 123(4):403–407, 2019.

[37] Vimla L Patel, Edward H Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. The coming of age of artificial intelligence in medicine. *Artificial intelligence in medicine*, 46(1):5–17, 2009.

[38] Shadab Adam Pattekari and Asma Parveen. Prediction system for heart disease using naïve bayes. *International Journal of Advanced Computer and Mathematical Sciences*, 3(3):290–294, 2012.

[39] Tadeusz Pietraszek and Axel Tanner. Data mining and machine learning-towards reducing false positives in intrusion detection. *Information security technical report*, 10(3):169–183, 2005.

[40] R Prashanth, Sumantra Dutta Roy, Pravat K Mandal, and Shantanu Ghosh. Automatic classification and prediction models for early parkinson's disease diagnosis from spect imaging. *Expert Systems with Applications*, 41(7):3333–3342, 2014.

[41] Yinsheng Qu, Bao-Ling Adam, Yutaka Yasui, Michael D Ward, Lisa H Cazares, Paul F Schellhammer, Ziding Feng, O John Semmes, and George L Wright. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical chemistry*, 48(10):1835–1843, 2002.

[42] E Ramaraj et al. An efficient pattern mining analysis in health care database. 2009.

[43] P Ranjitha and Vanishri Arun. Decision making for heart disease detection using hybrid neural network-particle swarm optimization algorithm. 2018.

[44] Nalluri MadhuSudana Rao, Krithivasan Kannan, Xiao-zhi Gao, and Diptendu Sinha Roy. Novel classifiers for intelligent disease diagnosis with multi-objective parameter evolution. *Computers & Electrical Engineering*, 67:483–496, 2018.

[45] P Sambasiva Rao and T Uma Devi. Applicability of apriori based association rules on medical data. *International Journal of Applied Engineering Research*, 12(20):9451–9458, 2017.

[46] P Sambasiva Rao and T Uma Devi. Improving accuracy of fuzzy rule based mining for heart disease detection using cost matrix. *International Journal of Advanced Research in Computer Science*, 9(1), 2018.

[47] Gilberto Reynoso-Meza, Javier Sanchis, Xavier Blasco, and Sergio García-Nieto. Physical programming for preference driven evolutionary multi-objective optimization. *Applied Soft Computing*, 24:341–362, 2014.

[48] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[49] Indira Rustempasic and Mehmet Can. Diagnosis of parkinson's disease using fuzzy c-means clustering and pattern recognition. *Southeast Europe Journal of Soft Computing*, 2(1), 2013.

[50] Oluwarotimi Williams Samuel, Grace Mojisola Asogbon, Arun Kumar Sangaiah, Peng Fang, and Guanglin Li. An integrated decision support system based on ann and fuzzy_ahp for heart failure risk prediction. *Expert Systems with Applications*, 68:163–172, 2017.

[51] M Sbayti, R Bahloul, H BelHadjSalah, and F Zemzemi. Optimization techniques applied to single point incremental forming process for biomedical application. *The International Journal of Advanced Manufacturing Technology*, 95(5-8):1789–1804, 2018.

[52] Manjeevan Seera and Chee Peng Lim. A hybrid intelligent system for medical data classification. *Expert Systems with Applications*, 41(5):2239–2249, 2014.

[53] Parul Sinha and Poonam Sinha. Comparative study of chronic kidney disease prediction using knn and svm. *International Journal of Engineering Research and Technology*, 4(12):608–12, 2015.

[54] Lea L Skovgaard, Sarah Wadmann, and Klaus Hoeyer. A review of attitudes towards the reuse of health data among people in the european union: The primacy of purpose and the common good. *Health Policy*, 2019.

[55] T Smitha and V Sundaram. Association models for prediction with apriori concept. *International Journal of Advances in Engineering & Technology*, 5(1):354, 2012.

[56] Hari Agus Sujono, Muhammad Rivai, and Muhammad Amin. Asthma identification using gas sensors and support vector machine. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 16(4), 2018.

[57] Sheng-Feng Sung, Kuanchin Chen, Darren Philbert Wu, Ling-Chien Hung, Yu-Hsiang Su, and Ya-Han Hu. Applying natural language processing techniques to develop a task-specific emr interface for timely stroke thrombolysis: a feasibility study. *International journal of medical informatics*, 112:149–157, 2018.

[58] Tuan Zea Tan, Chai Quek, and Geok See Ng. Ovarian cancer diagnosis by hippocampus and neocortex-inspired learning memory structures. *Neural Networks*, 18(5-6):818–825, 2005.

[59] Tiku T Tanyimboh and Anna M Czajkowska. Joint entropy based multi-objective evolutionary optimization of water distribution networks. *Water resources management*, 32(8):2569–2584, 2018.

[60] Feyzullah Temurtas. A comparative study on thyroid disease diagnosis using neural networks. *Expert Systems with Applications*, 36(1):944–949, 2009.

[61] Nastassia Tvardik, Ivan Kergourlay, André Bittar, Frédérique Segond, Stefan Darmoni, and Marie-Hélène Metzger. Accuracy of using natural language processing methods for identifying healthcare-associated infections. *International journal of medical informatics*, 117:96–102, 2018.

[62] Elif Derya Übeyli. Implementing automated diagnostic systems for breast cancer detection. *Expert systems with Applications*, 33(4):1054–1062, 2007.

[63] Fevrier Valdez, Patricia Melin, and Oscar Castillo. A survey on nature-inspired optimization algorithms with fuzzy logic for dynamic parameter adaptation. *Expert systems with applications*, 41(14):6459–6466, 2014.

[64] Rafael V Veiga, Helio JC Barbosa, Heder S Bernardino, João M Freitas, Caroline A Feitosa, Sheila MA Matos, Neuza M Alcântara-Neves, and Maurício L Barreto. Multiobjective grammar-based genetic programming applied to the study of asthma and allergy epidemiology. *BMC bioinformatics*, 19(1):245, 2018.

[65] Chung-Il Wi, Sunghwan Sohn, Mir Ali, Elizabeth Krusemark, Euijung Ryu, Hongfang Liu, and Young J Juhn. Natural language processing for asthma ascertainment in different practice settings. *The Journal of Allergy and Clinical Immunology: In Practice*, 6(1):126–131, 2018.

[66] Amani Yahyaoui and Nejat Yumusak. Decision support system based on the support vector machines and the adaptive support vector machines algorithm for solving chest disease diagnosis problems. *Biomedical Research*, 29(7), 2018.

[67] Duen-Yian Yeh, Ching-Hsue Cheng, and Yen-Wen Chen. A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications*, 38(7):8970–8977, 2011.

[68] Dimitrios Zafeiris, Sergio Rutella, and Graham Roy Ball. An artificial neural network integrated pipeline for biomarker discovery using alzheimer's disease as a case study. *Computational and Structural Biotechnology Journal*, 16:77–87, 2018.

[69] Rahmat Zolfaghari. Diagnosis of diabetes in female population of pima indian heritage with ensemble of bp neural network and svm. *Int. J. Comput. Eng. Manag*, 15:2230–7893, 2012.

Missing Data Prediction using Correlation Genetic Algorithm and SVM Approach

Aysh Alhroob[1], Wael Alzyadat[2], Ikhlas Almukahel[3], Hassan Altarawneh[4]
Department of Software Engineering, Faculty of Information Technology Isra University, Jordan[1,3]
Department of Software Engineering Faculty of Science and Information Technology Al-Zaytoonah University, Jordan[2]
Department of Computer Science Faculty of Information Technology Middle East University, Jordan[4]

***Abstract*—Data exists in large volume in the modern world, it becomes very useful when decoded correctly to inform decision making towards tackling real word issues. However, when the data is conflicting, it becomes a daunting task to get obtain information. Working on missing data has become a very important task in big data analysis. This paper considers the handling of the missing data using the Support Vector Machine (SVM) based on a technique called Correlation-Genetic Algorithm-SVM. This data is to be subjected to the SVM classification technique after identifying the attribute's correlation and application of the genetic algorithm. The application of the correlation enables a clear view of the attributes which are highly correlated within a particular dataset. The results indicate that apart from the SVM, the application of the proposed hybrid algorithm produces better outcomes identification rate and accuracy is considered. The proposed approach is also compared with depicts the Mean Identification rate of applying the neural network, the result indicate a consistent accuracy hence making it better.**

Keywords*—Missing data; Support Vector Machine (SVM); genetic algorithm; hybrid algorithm; correlation

I. Introduction

Data missing is the most common issue in various real worlds because it affects taking a timely decision using the acquired data. This research addresses missing data issues of data preprocessing that can have a significant impact on generalization performance of classification accuracy towards meaningful data. Various dataset suffer from an unavoidable problem of missing values for many reasons such as not enough data in report results; missing in industrial experiment, or failures automatic machine while collecting data [1] medicinal dataset contains missing data because some patient's record needs some critical value, not all possible tests to investigate it [2].

Data missing may happen at two stages; during training time as training data or at the prediction time while testing the data. The machine learning algorithms are mainly concerned with the identification of the missing values at the training time with less focus on missing values during prediction time. There are various techniques for treating missing data, examples include imputation techniques, ignoring techniques, and model-based techniques. The ignoring technique includes complete case analysis, which involves analyzing the case to have any missing data in any of the variables. The particular case is omitted from the analysis part. Another technique in ignoring is pair-wise deletion in which each of the features is considered and the value missing in any field is not much minded. Treating missing data requires thorough analysis process involving estimation of missing value without losing the statistical perspective of the dataset. These two criteria are contradictory and use the information from the partially completed data and at the same time maintaining the statistical perspective of the dataset while imputing the missing values [3]. Some techniques have been discussed to handle of missing data[3], such as remove cell containing missing data other using imputation with appropriate values. The main difference between the two approaches is that, removing missing data is more suitable with small number of instances to avoid information decrease while Imputation methods can be practical with big data and large missing value. Consequently, imputation methods are a accepted approach dealing with missing value.

Correlation is a technique which identifies the relationship between variables. The correlation factor helps in identifying the suitable relation between the variables of a particular dataset. The support vector machine (SVM) is a supervised learning technique which initially helped in two-class classification problem. The kernel functions may also be applied to optimize the parameters. Given a set of training data, SVM produces optimum hyper plane by using the concept of supervised learning. Basically a hyper plane is one which acts as a line that plays an important role in dividing the plane into two parts which belongs to each of the class. The SVM plays a drastic role when there is a clear-cut division of the two classes along X and Y plane. When there is no clear discrimination of the two classes through a particular line, then there is a need to use the third axis Z. There arises the use of kernals. Therefore by using some tuning factors in support vector machine and by changing them according to the problem enables to achieve non-linear classification. This type of classification helps in achieving higher accuracy in limited amount of time. Kernal plays a very important role in learning the hyper plane in SVM which helps in changing the problem using linear algebra. The SVM plays a major role in text categorization removing the need for labeled training data. Image classification is also possible through SVM which provides higher accuracy rate than existing systems. Image segmentation also has the usage of SVM in it.The SVM helps in classification of proteins in biological science and also enables recognition of the handwritten text. SVM also has few disadvantages. The SVM algorithm avoids probability estimation on data which are stable. The input data needs to be fully labeled. The applicability of SVM is more towards two-class problem and further multi-class structure needs to be looked upon.

The genetic algorithm has the basic steps of selection of population, crossover and mutation. The fitness function determines the quality of the individual. Fitness passed individuals are inherited to another generation. The genetic algorithm initially originates with a set of solutions and later variants them for different generations. For increasing the performance of the algorithm, random search is performed on the old data for new search items. Therefore genetic algorithm allows global search thereby trying to improve the global optimum through various available solutions.

The genetic algorithm has the necessary steps for selection of population, crossover, and mutation. The fitness function determines the quality of the individual, and individuals who pass fitness test are inherited to another generation. The genetic algorithm initially originates with a set of solutions and later variants them for different generations. For increasing the performance of the algorithm, a random search is performed on the old data for new search items. The genetic algorithm creates an opportunity for global search hence improves global optimum through the available solutions. The genetic algorithm tries to identify the attributes with the missing values [4]. Once the attribute has been identified, it engages in finding domain values of missing data values. Values of the missing attributes are then replaced with identified domain values such that possible set of domain values are identified for the missing attribute values. A similar concept applies to all attributes with missing values. With an overall bunch of arrived values, the set of values are chosen. Crossover on the set of selected instances is made. The fitness function is determined and validation is done against it. This helps in the determination of classification accuracy on the decision tree [5] [6] . If the selected instance is classified, then the substituted values are classified or else they are deleted. The process is repeated until a bunch of values is obtained. The proposed paper tries to address the missing data using the concept of correlation, which identifies the relation. Then the genetic algorithm and SVM are applied to handle the missing data and efficiently classify the data.

The paper is organized as follows: Section II deals with the related works in handling missing data, Section III is the proposed work and Section IV deals with the implementation and Section V deals with results and discussion and section VI deals with conclusion.

II. Related Work

Handling missing data is very important in term of use these data. Many Techniques used to optimize the data findings and use. Optimization and Machine Learning algorithms are used to enhance the data processing. The Genetic Algorithm (GA) [7] was used to optimize the initial weight and threshold values of support vector machine. The proposed GA-SVM was used to forecast the CO2 emissions of Beijing [8]. The factors contributing to this was identified to be residential growth, economic factors and the CO2 emissions were found to be more than 0.5. The cancer data is classified using support vector machine and genetic algorithm[9] to find the better accuracy in classification. Radial basis and polynomial kernel [10] function are used in this proposed technique. The proposed technique is compared to the existing techniques based on the runtime also.

In model selection using support vector machine [11], genetic algorithm is being used. The fitness function is calculated and various kernel parameters [12] [13] are determined. The proposed model selection technique is applied on four datasets to observe if it satisfies the criteria. The proposed estimator outperforms giving best fitness criterion that yielded more models. Authors in paper [14] proposed a genetic algorithm for optimizing the parameters of support vector machine. This involves image classification based on object-oriented classification. The proposed system is compared with the grid algorithm and found to be superior in terms of time and accuracy factors.

For the purpose of identifying the damage on the bridge, support vector machine along with genetic algorithm which is customized to get best kernel parameters. The proposed GA-SVM [15] is compared with other back propagation techniques to arrive at the best technique. With the error rates of other technique, it is being concluded that the proposed technique has higher accuracy rate in finding the damage. The least square SVM [16] technique is being proposed which helps in making the complex problem to linear regression one. Then by applying genetic algorithm over this LS-SVM [17], optimal parameters [18] [19] are obtained. The proposed system is compared with other existing systems like artificial neural network and it is found that the LS_SVM based system perform far better than that.

The classifier works in [23], presents how a classifier works if there are missing values in the data. Initially non-parametric technique is used for the data processing. But it narrows to simpler SVM if no missing data is present in the data. Further an analysis of Least square SVM [4] [24] is done to understand the classifier better.

The work in paper [1], is based on the objective to identify the missing rate in a selective manner. The proposed technique helps in achieving a good Mean Identification Rate (MIR) through less imputation method. By understanding the technique, the proposed method is evaluated for the parameter to check if the system is working properly. The paper [2]is based on the functional dependency related technique which is targeted with machine learning. The algorithm namely K-nearest neighbour algorithm is used to find the functional dependency in the given data. The concept of using data dictionary also yields effective results. The parameter namely missing rate [4] is taken into account for evaluation.

Additive least square technique with application of support vector machine which helps in performing classification of the data which are missing is presented in [3]. Cross validation strategy with ten folds is performed to correctly classify the data. The strategy is verified by measuring the accuracy factor through mean and standard deviation values [6] for the given data. The research in [20] provides embedding based calculation of the missing data through non-linear technique that bind the vector label. The proposed system is evaluated for its performance and also by the time taken for training the dataset.

The method of finding the missing data and grouping them is done through sampling in [21]. This method helps in omitting the missing data by calculating the error. Based on the accuracy and error calculation, the proposed system is evaluated. SVM based model which does not require selection of planes is investigated in[16]. The system is evaluated

TABLE I. VARIOUS EXISTING MISSING DATA HANDLING TECHNIQUES

Reference	Year	Proposed Approach	Merits	Evaluation Parameters
[1]	2016	Missing Rate Oriented selective (MROS) algorithm	Achieve effective mean Identification rate (MIR) with minimal imputation effort.	Mean Identification Rate
[2]	2018	Functional dependency based techniques, Machine learning based KNN	Functional dependency and data dictionary provide efficient results	Missing Rate
[3]	2018	Novel transfer-based additive least square support vector machine (LS-SVM)	Perform direct classification on the missing data	Ten fold cross validation strategy, Calculation of mean and SD of accuracy
[20]	2019	Embedding based method that non-linearly embeds label vector	Accurate prediction of tail labels	Prediction performance and Training Time
[21]	2014	ImputationTechnique based on sampling method	Overcomes the missing data using copulas using small error	Accuracy error
[16]	2008	SVM Kriging Technique	This model does not requires selecting variogram models	Mean error, Mean absolute error, Root mean square prediction error.
[22]	2016	BayesianNetwork, Multilayer perceptron, C.4	Higher accuracy rate is provided by KNN and optimal endurance by MLP	Accuracy and endurance

using the parameters like root mean square and absolute error [6] which helps in effective determination of the proposed technique.

Table I summarizes the various techniques for handling missing data and the merits and evaluating parameters.

III. PROPOSED METHODOLOGY

A. Panel Data

Panel data is a multidimensional-format data involving measurements over varied time. The multi-dimensional format represents the various attributes of a dataset constituting a complete dataset. Time series data also comes under the panel data. The dataset with the primary data element occurring n number of times in a particular time series makes it worth investigating. A balanced panel is one in which the panel data is continuously observed in every time interval as represented in Equation 1.

$$NO = P * T \tag{1}$$

Where, NO is the Number of Observations in the data, P denotes the panel members and T denotes the time period.Where, I (I=1....n) is the individual factor and T (T=1....t) is the time factor. At the same time, Equation 2 represent the panel data

$$P_{(IT)} \tag{2}$$

Using the panel data, the model can be constructed as shown in Equation 3 and Equation 4:

$$Q_{IT} = \mu + \sigma P_{IT} + F_{IT} \tag{3}$$

$$F_{IT} = \gamma_I + G_{IT} \tag{4}$$

Where, G_{IT} is the component which varies with time and γ_I is the specific thing to a particular member and fixed for a time interval.

B. Pre-Processing

Data in reality has a noisy and incomplete [25]. To address the preprocessing, various techniques of data cleaning integration, and reduction are incorporated to make it more consistent[26] . For the proposed technique CGA-SVM, the data cleaning process involves identification and addressing of missing values. Further data values are sorted and arranged into their respective buckets; a process called binning. Values that do not reflect any cluster are identified, differentiated through outlier detection techniques. Redundant values in data aggregation process are then removed after which the process of normalizing values is done. During data reduction, the attribute dimension which in this case is the size is reduced and data compressed making it easier to handle. The pre-processing technique for CGA-SVM is initially based on the data cleaning procedures as shown in Fig 1. Data cleaning has the ability to fill missing values based on mean or median values of all the values. When a missing value occurs, that individual tuple may be omitted from the general dataset. Such omission may not be efficient as key details could be missed. However, when

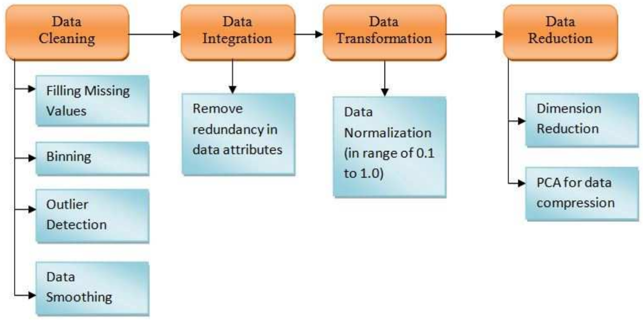


Fig. 1. Data Pre-processing Techniques for CGA-SVM Technique

many columns of that tuple are missing, the omission technique would work.

Missing values can be keyed in manually but only small datasets with fewer tuples. Replacement of missing data with global constant fixed based on the relevant dataset can also apply in such a situation. Data becomes noisy by using measuring instruments that make faulty calibrations. Binning is the next technique as it helps in classifying data into several buckets. Smoothing is also a data cleaning strategy, which involves replacing bin values by either mean or with the close boundary value. The values, which do not fit any of the group, are termed as outliers and are being handled with the dataset. Therefore, the pre-processing takes place efficiently starting from cleaning and further proceeds until reduction with intermediate steps being executed.

C. Correlation

Correlation is a measure of association between two attributes and also the nature of the relationship[27]. The correlation coefficient value lies between -1 to +1. Correlation a mathematical value which describe a relationship between one or more independent variables with dependent variable. For example, a correlation can be a connection between two variables (numeric) values. If increasing happened to one variable value, then the other one also will increase (or decrease). However, potential predictive power in Correlations make it valuable: use or act on the value of one variable to predict or modify the value of the other. Furthermore, the correlation does not imply causation and a correlation does not tell us about the core cause of a relationship. The correlation method is a systematic praxis with roots going far back in human history. It is also used to analyses extremely large datasets correctly and efficiently that plays a critical role in the science of the future.

There are various correlations namely pearson, kendall and spearman which are used to measure the relationship between two attributes or variables. Pearson Correlation is a measure of the degree of relationship between linearly associated variables. The variables are required to be normally distributed. It is given as the ratio of covariance of the variables to the product of their standard deviation as shown in Equation 5.

$$\rho = \frac{cov(a,b)}{\sigma_a \sigma_b} \tag{5}$$

Where, ρ is the Pearson coefficient and a,b are the two variables.

Kendall correlation is non-parametric where there is a dependency between two variables and is represented as tau. It analyses the relationship between two variables, and provide solution between discordant and cordant pairs. Spearman correlation is ranked, depending on the variable's rank value for the operation. It is denoted as ρs.

When the three correlation techniques are adopted in both panel datasets, the outcome shows how Pearson coefficient provides better correlation values as indicated. Once the correlations between the attributes are obtained, it is subjected to the SVM classification with genetic algorithm applied.

D. CGA-SVM Technique

Data generated after analysis of the correlation factor is examined with a genetic algorithm, which is further analysed by the SVM technique. This provides room for treatment of missing values hence minimising their effects.

CGA-SVM Algorithm:

- Input: Dataset with missing values and correlation details.
- Output: Classified data with addressed missing data.

1) *Initialize each individual and then produce which is in accordance with X,i=1,2. . . .ln.*
2) *Arrive based on the signs of and*
3) *Calculate the fitness values using the fitness function for the individuals as follows: F(z)=j-f'(where f'(is the derivative of the objective function for optimization based on the correlation result and j is the constant for scaling the fitness function*
4) *If condition achieved then stop. Else move to step 5.*
5) *If the best fitness value is less than the threshold, Go to next iteration.*
6) *Do selection based on F(z). Then perform crossover between the chromosomes with same attribute values.*
7) *The Leave one out cross validation is applied as a condition for SVM with GA.*
8) *Perform new iteration of variables generation using the steps 3-7 and exit.*

In the algorithm CGA-SVM, the genetic algorithm determines the fitness value of the used variables. The fitness function is defined by determining the fitness value of variables using the objective function. it is an iteration process until the best fitness value is achieved, compared to the threshold. Mutation and crossover are also performed with the matching chromosomes. Validation is done using Leave one out technique which ensures that the proposed fitness value meets the threshold and CGA-SVM outperforms the existing techniques. The flowchart in Fig. 2 represents the correlation based Genetic algorithm and SVM combination of classification which involves parameter setting in SVM and then finding the fitness value. Once the fitness function is determined, and the required condition is satisfied, SVM model is evaluated and missing values are addressed. If condition is not met, then further genetic algorithm is being applied with the activities of selection, crossover and mutation to arrive at desired objective function.

IV. Implementation

Four benchmark panel datasets [28] are used to support the findings of this study have been selected from the UCI machine learning repository as follow:

1) Ionospher
2) Iris Plant
3) Parkinson
4) SPECTF

First correlation is applied to identify the association among the attributes in a dataset. Correlation value GA will be applied then followed by classification by SVM for meaningful data and testing of results. Table II shows the main characteristics

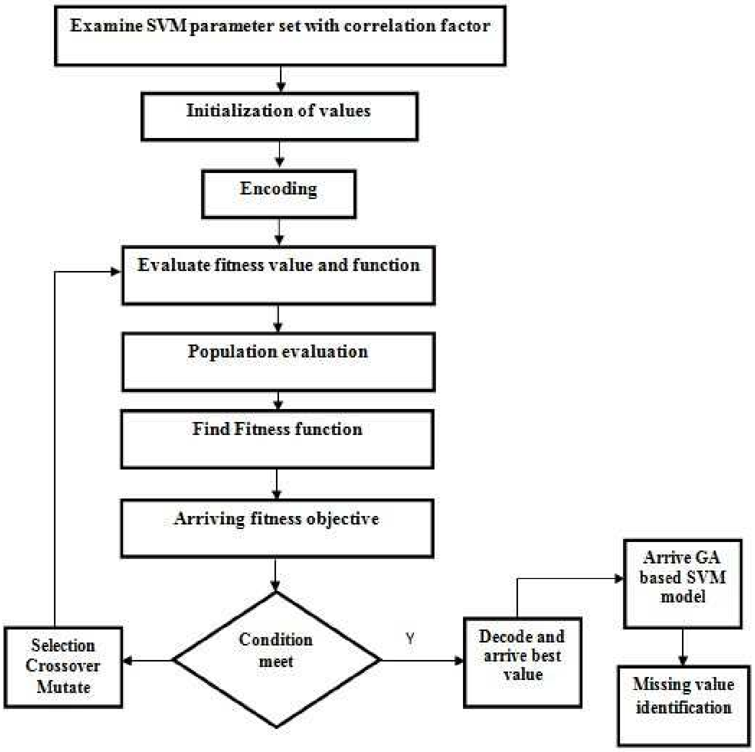

Examine SVM parameter set with correlation factor
Initialization of values
Encoding
Evaluate fitness value and function
Population evaluation
Find Fitness function
Arriving fitness objective
Condition meet
Y
Selection Crossover Mutate
Decode and arrive best value
Arrive GA based SVM model
Missing value identification

Fig. 2. Flow of CGA-SVM Approach

of each dataset with the number of instances and the number of attributes. For each dataset, the missing data with random values were used to present missing values in terms of correlation values.

TABLE II. The Main Characteristics of Eeach Dataset

Name	Class	Instances	Features	Missing Data (Random)
Ionosphere	2	351	34	17%
Iris Plant	3	150	4	25%
Parkinson's	2	267	44	15%
SPECTF Heart	2	267	44	36%

A. Correlation Methods

Using Tidy verse package in R software, the dataset is being read and Tidy up the dataset by making every row and column clear for observation. A variable is one way to visualize the rearranged data, making the relationships between measure, class, and part a little clear. Correlation values between the various datasets parts and corresponding measures are based on what class the attributes of each data set happens to be. Reshape2 package helps to reshape benchmark dataset and establishment of variables for correlation of every measurement against the other. Pass the benchmark dataset, grouped by class, to (cor_list) function, which calculates correlations by applying the Reshape2 package. This will generate N rows as shown the Equation 6.

$$RowsNumber = (mclasses * (Nmeasurements)^2) \quad (6)$$

relation coefficients for every measurement pair. The last step of correlation is the visualization of correlations between measurements, grouped by class. Fig. 3 summarizes correlation plot box measurement according to Pearson correlation coefficients values which greater than or less than 0. We then omit 0 values indicating no relation among attribute. Fig. 4 indicates the strengths of correlation coefficients of correlation measure within four different features of used datasets.

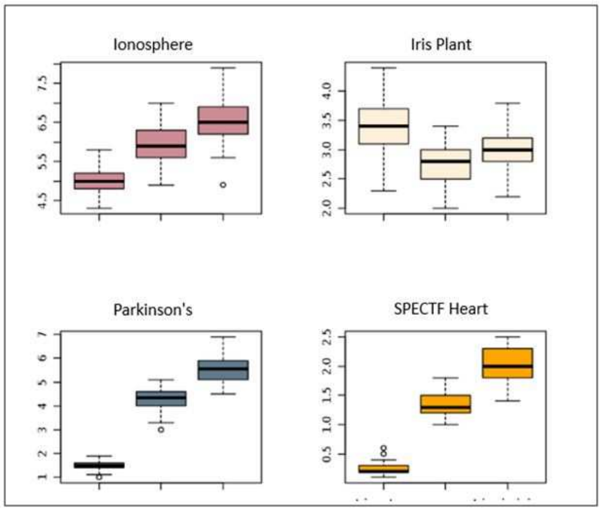


Fig. 3. Box Plot Determine the Correlation Among Different Datasets and Missing Data

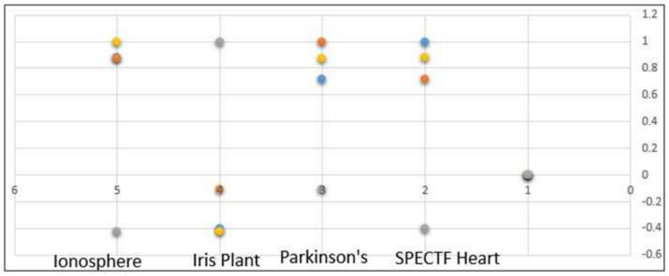


Fig. 4. Comparison of Strengths Correlation Coefficients in Correlation Measure for Four Futures within Datasets

B. Genetic Algorithm

The benefit of the proposed approach is its imputation approach based on Genetic Algorithm. Input to the algorithm is a dataset with missing values and correlation values not equating 0 and that needs imputation. Primarily, the dataset is randomly imputed. This step is relevant as it produces a temporary complete dataset for enhancements in later steps through crossover and mutation. The algorithm repeats itself continuously regressing each attribute with missing values on other attributes. Implementation of Genetic programming in the experiment adopted the Genetics package. For 100 iterations, calculate the accuracy independently running for each benchmark dataset, and the result summary is as shown in Table III and Fig. 5.

TABLE III. Genetic Algorithm with Accuracy values for 21 Rounds

Run	Ionosphere	IrisPlant	Parkinsons
1	94.9	94.5	85.5
2	96.32	96.1	90.35
3	90	89	88.23
4	91.62	97.4	90.12
5	95.54	95.4	89.12
6	92.65	91.4	88.2
7	93.5	96.19	78.15
8	89.8	93.33	90
9	97.2	91.2	96
10	97.23	94.6	85.5
11	95.32	92.3	90.35
12	88.9	87.8	88.23
13	92.9	95.28	90.12
14	96.3	92.64	89.12
15	92.8	87.14	88.2
16	96.8	84.28	78.15
17	97.1	91.42	93.5
18	95.2	94.76	92.1
19	93.2	96.56	89
20	94.7	96.19	91.4
21	96.3	96.4	88.01
Max	97.23	97.4	96
Min	88.9	84.28	78.15
Median	94.9	94.5	89.12
Mean	94.20380	93.04238	88.54048

V. Results and Discussion

The proposed system is SVM with correlation and GA applied. It is compared with the simple SVM providing accuracy and error comparison. The prediction accuracy (%) (with errors less than 10%) using SVM with correlation and GA is Accuracy1 and without correlation and GA is Accuracy2.

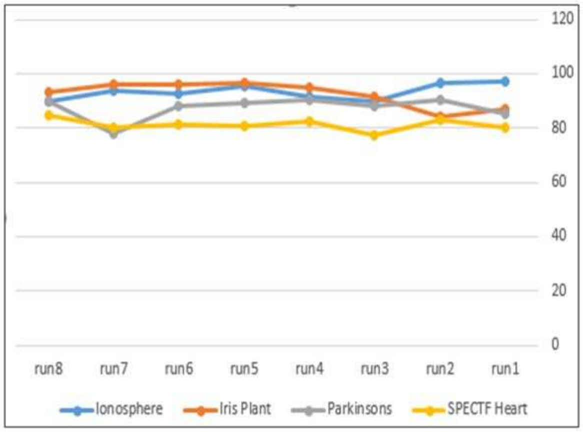


Fig. 5. GA Plot for Different Datasets

TABLE IV. Accuracy Comparison of SVM and CGA-SVM Techniques

Dataset	Accuracy1	Accuracy2
Ionosphere	97.23	94.55
Iris Plant	96.56	95.64
Parkinson’s	90.35	89.5
SPECTF Heart	84.44	80.17

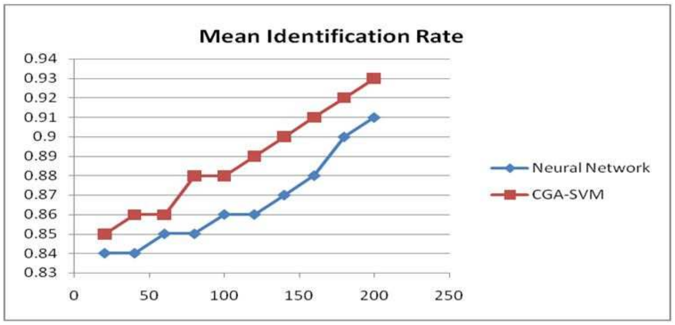


Fig. 6. Mean Identification Rate Comparison of Neural Network and CGA-SVM.

Table IV shows accuracy comparisons of various datasets for the SVM technique and CGA-SVM technique. Fig. 6 depicts the Mean Identification rate of applying the neural network approach and the proposed CGA-SVM approach. Regarding to Table IV and Fig. 6 show the best accuracy achieved in the experiments, after training with CGA-SVM. The proposed system is also compared with depicts the Mean Identification rate of applying the neural network approach by (91%) and the proposed CGA-SVM approach(93%) which mean the existing systems to handle missing values, where the results indicate a consistent accuracy hence making it better.

VI. Conclusions

Addressing missing data in the big dataset is very important. The proposed system handles the missing data through correlation technique followed by genetic algorithm imposed on support vector machine. This variant of SVM performs well as it effectively handles the missing data. The proposed system is first subjected to correlation technique comparing the various techniques and then evaluating it using the fitness function of the genetic algorithm. The proposed CGA-SVM provides better accuracy than the existing techniques based on the standard SVM. Further, mean identification rate is used for the comparison of the proposed technique with the existing neural network approach, and the results show that the proposed technique has a higher percentage accuracy of 2% accuracy compared to existing methods.

In Future work, Grey Wolf optimization algorithm will be used to avoid the irrelevant and redundant attributes significantly, after the features are forwarded to the SVM.

References

[1] X. Li, G. Li, and R. Fishbune, "A novel missing-rate-oriented selective algorithm for handling missing data by minimizing imputation," in *2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE, 2016, pp. 234–237.

[2] I. Ezzine and L. Benhlima, "A study of handling missing data methods for big data," in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*. IEEE, 2018, pp. 498–501.

[3] G. Wang, J. Lu, K.-S. Choi, and G. Zhang, "A transfer-based additive ls-svm classifier for handling missing data," *IEEE transactions on cybernetics*, vol. 50, no. 2, pp. 739–752, 2018.

[4] T. Yeoh, S. Zapotecas-Martínez, Y. Akimoto, H. Aguirre, and K. Tanaka, "Genetic algorithm assisted by a svm for feature selection in gait classification," in *2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*. IEEE, 2014, pp. 191–195.

[5] Y. Ding and J. S. Simonoff, "An investigation of missing data methods for classification trees applied to binary response data," *Journal of Machine Learning Research*, vol. 11, no. Jan, pp. 131–170, 2010.

[6] K. Sijtsma and L. A. Van der Ark, "Investigation and treatment of missing item scores in test and questionnaire data," *Multivariate Behavioral Research*, vol. 38, no. 4, pp. 505–528, 2003.

[7] D. Ugryumova, R. Pintelon, and G. Vandersteen, "Frequency response function estimation in the presence of missing output data," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 2, pp. 541–553, 2014.

[8] J. Li, B. Zhang, and J. Shi, "Combining a genetic algorithm and support vector machine to study the factors influencing co2 emissions in beijing with scenario analysis," *Energies*, vol. 10, no. 10, p. 1520, 2017.

[9] D. Nithya, V. Suganya, and R. S. I. Mary, "Feature selection using integer and binary coded genetic algorithm to improve the performance of svm classifier," *Journal of Computer Applications (JCA)*, vol. 6, no. 3, p. 2013, 2013.

[10] P. Stoica, J. Li, J. Ling, and Y. Cheng, "Missing data recovery via a nonparametric iterative adaptive approach," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 3369–3372.

[11] S. Lessmann, R. Stahlbock, and S. F. Crone, "Genetic algorithms for support vector machine model selection," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. IEEE, 2006, pp. 3063–3069.

[12] S. Di Martino, F. Ferrucci, C. Gravino, and F. Sarro, "A genetic algorithm to configure support vector machines for predicting fault-prone components," in *International conference on product focused software process improvement*. Springer, 2011, pp. 247–261.

[13] L. Garg, J. Dauwels, A. Earnest, and K. P. Leong, "Tensor-based methods for handling missing data in quality-of-life questionnaires," *IEEE journal of biomedical and health informatics*, vol. 18, no. 5, pp. 1571–1580, 2013.

[14] M. Li, X. Zhou, X. Wang, and B. Wu, "Genetic algorithm optimized svm in object-based classification of quickbird imagery," in *Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*. IEEE, 2011, pp. 348–352.

[15] H.-B. Liu and Y.-B. Jiao, "Application of genetic algorithm-support vector machine (ga-svm) for damage identification of bridge," *International Journal of Computational Intelligence and Applications*, vol. 10, no. 04, pp. 383–397, 2011.

[16] W. Huizan, Z. Ren, L. Kefeng, L. Wei, W. Guihua, and L. Ning, "Improved kriging interpolation based on support vector machine and its application in oceanic missing data recovery," in *2008 International Conference on Computer Science and Software Engineering*, vol. 4. IEEE, 2008, pp. 726–729.

[17] S. Moridpour, T. Anwar, M. T. Sadat, and E. Mazloumi, "A genetic algorithm-based support vector machine for bus travel time prediction," in *2015 International Conference on Transportation Information and Safety (ICTIS)*. IEEE, 2015, pp. 264–270.

[18] G. Wang, Z. Deng, and K.-S. Choi, "Tackling missing data in community health studies using additive ls-svm classifier," *IEEE journal of biomedical and health informatics*, vol. 22, no. 2, pp. 579–587, 2016.

[19] W. Shi, Y. Zhu, S. Y. Philip, T. Huang, C. Wang, Y. Mao, and Y. Chen, "Temporal dynamic matrix factorization for missing data prediction in large scale coevolving time series," *IEEE Access*, vol. 4, pp. 6719–6732, 2016.

[20] A. H. Akbarnejad and M. S. Baghshah, "An efficient semi-supervised multi-label classifier capable of handling missing labels," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 229–242, 2018.

[21] R. Houari, A. Bounceur, A. K. Tari, and M. T. Kecha, "Handling missing data problems with sampling methods," in *2014 International Conference on Advanced Networking Distributed Systems and Applications*. IEEE, 2014, pp. 99–104.

[22] O. M. Prabowo, K. Mutijarsa, and S. H. Supangkat, "Missing data handling using machine learning for human activity recognition on mobile device," in *2016 International Conference on ICT For Smart Society (ICISS)*. IEEE, 2016, pp. 59–62.

[23] K. Pelckmans, J. De Brabanter, J. A. Suykens, and B. De Moor, "Handling missing values in support vector machine classifiers," *Neural Networks*, vol. 18, no. 5-6, pp. 684–692, 2005.

[24] D. P. Mesquita, J. P. Gomes, F. Corona, A. H. S. Junior, and J. S. Nobre, "Gaussian kernels for incomplete data," *Applied Soft Computing*, vol. 77, pp. 356–365, 2019.

[25] W. J. Alzyadat, A. AlHroob, I. H. Almukahel, and R. Atan, "Fuzzy map approach for accruing velocity of big data," *Compusoft*, vol. 8, no. 4, pp. 3112–3116, 2019.

[26] A. Alhroob, W. J. Alzyadat, I. H. Almukahel, and G. M. Jaradat, "Adaptive fuzzy map approach for accruing velocity of big data relies on fireflies algorithm for decentralized decision making," *IEEE Access*, vol. 8, no. 1, pp. 2169–3536, 2020.

[27] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[28] C. Dua, Dheeru & Graff, "Uci machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

Ranking System for Ordinal Longevity Risk Factors using Proportional-Odds Logistic Regression

Nur Haidar Hanafi[1], Puteri Nor Ellyza Nohuddin[2]
Institute of Visual Informatics[1,2]
Universiti Kebangsaan Malaysia
43650 UKM, Bangi
Selangor, Malaysia
Faculty of Computer and Mathematical Sciences[1]
Universiti Teknologi MARA
70300 Seremban
Negeri Sembilan, Malaysia

***Abstract*—Longevity improvements have traditionally been analysed and extrapolated for future actuarial projections of longevity risk by using a range of statistical methods with different combinations of statistical data types. These methods have shown great performances in explaining the trend movements of the longevity rate. However, actuaries believe that knowing the trend movements is not enough, especially in controlling the impact of the longevity risk. Accessing the effects of each level of the risk factors, especially ordinal risk factors, towards the improvements of the longevity rate would provide significant additional knowledge to the trend movements. Therefore, this study was conducted to determine the potentiality of Proportional-Odds Logistics Regression in ranking the levels of the ordinal risk factors based on their effects on longevity improvements. Based on the results, this method has successfully reordered the levels of each risk factor to be according to their effects in improving longevity rate. Hence, a more meaningful ranking system has been developed based on these new ordered risk factors. This new ranking system will help in improving the ability of any statistical methods in projecting the longevity risk when handling ordinal variables.**

***Keywords*—Longevity risk; ordinal risk factors; risk ranking; proportional-odds ratios; effect analysis**

I. Introduction

Longevity improvements indicate a good sign that people are enjoying better fitness and better health conditions than the previous generations. A high life expectancy is considered as a success story to the enforcement of public health policies and socio-economic development systems. The increasing number of older people due to the improvement in life expectancy has undeniably changed the landscapes of every society and economic activity across various industries. These older people have become an increasing consumer group with specific needs and significant spending patterns, especially on health care and mortality protection policies.

Providing readily accessible and available health care and insurance policies can significantly contribute to equal social, economic, political, and cultural participation of these older people. There are increasing demands for insurance companies and private pension managers to bring to the market more products that are suited to the needs and wants of this consumer group. However, according to Hanafi and Nohuddin [1], the policymakers are more concerned about the impact that the life expectancy improvement may have on their financial position and reserve status due to the exposure of longevity risk.

This concern arises due to various negative impacts of longevity risk which include higher financial responsibilities for governments and annuity policy providers [2], risk of outliving resources during old ages for individuals [3][4]; and reduction in the ability of younger members of a family to take care of the older ones due to the extended mobility of the workforce [5]. To address this concern, it is essential for the policymakers to accurately assess the size of longevity risk in ensuring that product pricing, risk management, and asset allocation can be done smoothly.

Several statistical models have been developed to accurately predict longevity risk including cause-of-death specific models, disease-based models, and population-based models. These models were generated based on a diverse spectrum of mortality rate projections tools. Some of the models involved application of life tables [6][7], stochastic mortality models [8], generalized dynamic factor models with vine-copulae simulations [9]; and various data mining techniques including logistic regression technique and decision tree technique [10][11].

Policymakers find these statistical models to be more appealing to them because they help in improving the underwriting process by providing analytics-based approaches using readily accessible customers information to yield a more accurate, consistent, and efficient decision. This helps in simplifying policy applications for smaller face amounts, reducing data acquisition and storing cost, refining underwriting requirements, and processing data via automated software packages. However, actuaries believe that predicting the lonvegity risk movements is not enough, especially in controlling the impact of the longevity risk.

Accessing the effects of each level of the risk factors, especially ordinal risk factors, towards the improvements of the longevity rate would provide significant additional knowledge. Therefore, this study was conducted to determine the potentiality of Proportional-Odds Logistics Regression in ranking the levels of the ordinal risk factors based on their effects on longevity improvements. This new ranking system will help in improving the ability of any statistical methods in projecting the longevity risk when dealing with ordinal risk factors.

This paper is divided into different important sections to ease the discussion process. The first section discussed the motivation behind this research which include the different risk factors influencing the longevity improvements and the gap analysis on the existing risk ranking systems, especially when using ordinal risk factors. The second section discussed the nature of Proportional-odds Logistic Regression and its potentiality as an alternative to the current risk ranking system. The third section showed the empirical illustration using data on death records with a combination of different risk factors data types. The fourth section combined all the information from the third section into a meaningful risk ranking system. Conclusion and recommendations are then included in the last section.

II. Motivation

A. Risk Factors

Identifying significant risk factors is the most important phase in developing statistical models to predict longevity risk. Various studies have been done to identify significant risk factors in influencing life expectancy improvements. These studies were done using historical data; either non-disease data or genomic data or a combination of both. Non-disease data is readily accessible demographic data from the customers' database while genomic data is medically obtained database such as complete sets of DNA data including all of the customer's gene maps. Different combination of the risk factors would change how the models perform in predicting the longevity risk.

In the underwriting process, selecting a method that is cheaper and producing faster results would benefit the policymakers. Genomic data would require a large space to store and complex purpose-built software to analyse as compared to non-disease data. Therefore, most policymakers would prefer to use non-disease data when developing the models for future projections of longevity risk. Based on past studies using non-disease data, there are five most significant risk factors which include gender, race, residential status, marital status and education level [12][13][14][15][16][4][17][18][2].

B. Risk Ranking System

A major drawback of using statistical models is that they produced very complex methods which are difficult to be explained to the stakeholders and non-statistical practitioners. Therefore, multiple attempts have been done to transform them into a meaningful risk ranking system. A risk ranking system is a phase in the risk management process whereby the identified risks are assessed either using quantitative or qualitative analysis to determine which risks have the highest consequence of occurrence in order of importance. This method is considered to be a simpler mechanism as compared to the more complex statistical methods. It simplifies the estimation of risks, increases the visibility of the risks and assists the policymakers in decision making.

One notable method among a limited number of risk ranking systems that are available is a risk matrix model. A risk matrix model is an m by n risk matrix which provides the assessment of the size of individual risks versus the assessment of group risks along with the amount of overall exposure of the insurance company, particularly with regard to general and life insurance products [19]. One major advantage of this model comes from its relatively simple and transparent characteristics as well as its ability to trace risk trends over time. However, the simplicity of this method leads to inconsistent possible results.

Using qualitative risk parameters in the risk matrix model is a subjective process of numerical interpretation of the risk parameters by means of crisp intervals. This type of interpretation violates the real-life gradual transition between intervals. This problem has long been pointed out and a fuzzy risk ranking approach has been introduced in place of the crisp framework in order to overcome it [20].

A fuzzy model risk ranking system is a system developed by defining risk factors as fuzzy sets [21]. By doing so, an insurance company can utilize multiple prognostic factors that are imprecise and vague. This model provides a more realistic way of modelling longevity risks since it allows for interactions between multiple risk factors at once. Furthermore, it captures expert knowledge and allows for expertise descriptions to be done in a more intuitive and human-like manner.

When dealing with ordinal risk factors, most of the existing risk ranking models have failed to provide a thorough knowledge of the effects of each level of the risk factors towards the improvements of life expectancy. Analysing, extrapolating and scoring the degree of longevity risk while ignoring the substantive features of the risk factors will produce biased results. Requirements for models with the potential to accurately represent the actual characteristics of each of the risk factors should be met.

This study is conducted to assess the potentiality of Proportional-Odds Logistics Regression (POLR) in ranking the ordinal risk factors' levels based on their individual effects on longevity improvements. Hence, a more meaningful ranking system can be developed based on these new ordered risk factors. This new ranking system will help in improving the ability of any statistical methods in projecting the longevity risk when handling ordinal variables.

III. Proportional-Odds Logistic Regression

An ordinal regression model is a regression model specifically developed for ordinal dependent variables based on discrete or continuous covariates. It is similar to binary and multinomial logistic regression whereby it uses the same iterative procedure called maximum likelihood estimation. There are several types of ordinal logistic regression models. The most frequently used in practice is the proportional odds model [22][23][24].

$$logit[P(Y \leq j)] = \alpha_j - \Sigma \beta_i X_i \tag{1}$$

The above equation represents the proportional-odds models where $j = 1, \cdots, J-1$ and $i = 1, \cdots, M$. Let Y denotes the response category in the range $1, 2, \cdots, j$ such that $j \geq 2$. On the right side of the equation is a simple linear model with one slope, β, and an intercept α_j that changes depending on the category j in which the intercepts depend on j, but the slopes are all equal. According to this equation, the model is basically generating the probability of being in one category lower level versus being in categories above it.

Some advantages of the logit link are worth to be mentioned here. The model yields constant odds ratios across each split with interpretations very similar to logistic regression. It represents both orderings as well as categorical nature without any substantial increase in the difficulty of interpretation, such that it decreases variability and increases interpretability of the subject matter. Thus, the model has fewer terms than a multinomial regression model.

IV. Empirical Illustration

The analysis procedures illustrated in this section play a major purpose in highlighting the potentiality of POLR in ranking the ordinal risk factors' levels based on their effects on longevity improvements. The POLR has the ability to capture the strength of the effects that the independent variables have on a given dependent variable, thus it can be used to understand how much the dependent variable changes when the independent variables are changed.

A. Data Descriptions

The death records of Americans who had died at the age of 70 and above from 2013 until 2015 were used in this study with a total of 3,765,210 recorded deaths. Only deaths that occurred at the age of 70 and above were selected for this study because the size of longevity risk is more significant amongst those within this age group. Such datasets were used because they represent real-life risk exposures of people being able to live and die beyond their life expectancy; including both typical and extreme ones.

Any unnatural deaths would disturb the accuracy of this study, thus only deaths caused by natural events were included; while those caused by suicide, homicide or other unnatural events were removed. Only six non-disease risk factors were included in this study which is the age at death, gender, race, residential status, marital status and education level. These variables are a combination of nominal and ordinal variables without any range variables. Age at death was selected as the dependent variable where each age group represents the level of longevity risk exposure.

The descriptions of these non-disease risk factors along with their variable types and codes as coded in the RStudio software are presented in Table I. The R software version 1.1.383 installed in Windows 10 with processor Intel(R) Core (TM) i7-6500U CPU at 2.5GHz was used in the analysis process. Different R packages have been used for various purposes which included readr package for reading rectangular data [25], plyr package for splitting, applying and combining data [26], ggplot2 package for data visualisation [27], ordinal package for ordinal regression modelling [28], stats package for running a Kruskal-Wallis Test [29]; and effects package for effect displays [30][31][32].

TABLE I. Data descriptions.

Variable	Description	Variable Type
Age_Death	The age at death; recorded as a single age, was then discretized into 5-year age groups so that it became categorical data. {Age_Death = ”70-74”, ”75-79”, . . . , ”100+”}	Ordinal
Residential	{Residential = “Residents”, “Intrastate NR”, “Interstate NR”, “Foreign R”}	Nominal
Education	The education status of the deceased; recoded using the revised 2003 education codes. {Education = “Primary”, “Secondary”, “Diploma/GED”, “Degree”, “Master”, “PHD/Professional”}	Ordinal
Gender	{Gender = “Female”, “Male”}	Nominal
Marital_Status	{Marital_Status = “Single”, “Married”, “Widowed”, “Divorced”}	Nominal
Race	{Race = “White”, “Black”, “American Indian”, “Asian/Pacific Islander”}	Nominal

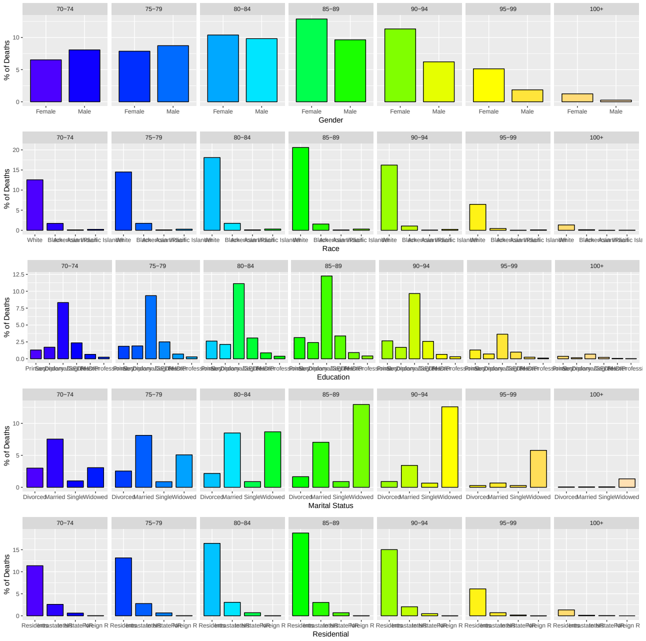


Fig. 1. Data composition of each risk factor with respect to age at death.

B. Descriptive Analysis

The basic features of each risk factor with respect to each category of the dependent variable based on the percentage of deaths was visualised using bar chart plots as shown in Figure 1. Separating each risk factor based on their item composition is an important process in getting to know the structure of this dataset, thus highlighting potential relationships between variables and giving extra information for some early presumptions of the longevity risk exposure.

According to Figure 1, the majority of the deaths among citizens aged 70 up to age 79 were male before being dominated by female starting from age 80 onwards. White coloured people, having diploma/GED as their highest education level and residents of the United States of America have recorded the highest percentage of deaths for all age groups. From ages 70 to 79, the majority of the deaths occurred among married people before being dominated by widowers starting from age 80 onwards.

The correlation between the dependent variable and each risk factor is represented by mosaic plots as shown in Figure 2. The size of the boxes corresponds to the size of death records within each level of each risk factor as in Figure 1. The shading density of each box is based on the Pearson residual values. The blue and red coloured boxes represent the level of the residual for that category. More specifically, a blue coloured box means that more observations in that box that would have been expected under the null model; whereas a red coloured box means that there are fewer observations in that box than the one would have been expected. This is known as positive and negative relationships between each category of the dependent variable with each category of the risk factors.

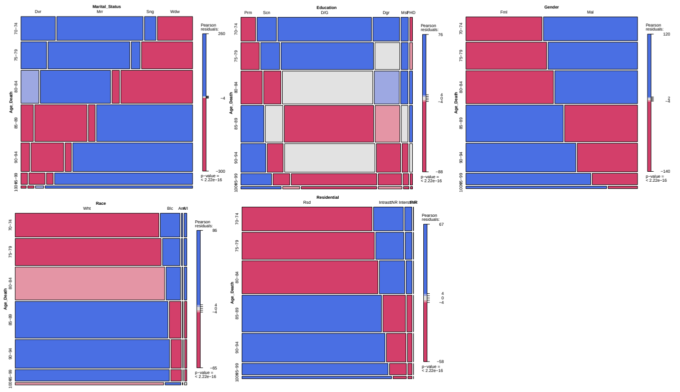


Fig. 2. Classic mosaic plot for correlation between age at death and the risk factors.

The p-value of the Pearson Residuals indicates the significant departure from the independence of the association between any given two risk factors. Therefore, when a p-value is greater than 0.05 then there is no association between the two risk factors. From Figure 2, the values of Pearson residuals for all the risk factors used in this study are less than 0.05. This result shows that all five risk factors are significantly affecting the deceased's life expectancy.

There are a few interesting findings that could be discussed further with respect to Figure 2. Even though all risk factors are significantly associated with age at death as proven by the p-values, there exist some levels within the Education risk factor that are not associated with age at death. This condition can be seen from grey coloured boxes. For example, owning a diploma/GED certificate did not affect those who died between ages 80 to 84 and between ages 90 to 94.

C. Proportional-Odds Assumption

One of the assumptions underlying POLR is that the relationship between each pair of the dependent variable's categories is the same, thus it is called a parallel regression assumption. A POLR study can only be done if and only if this assumption is checked to ensure that the coefficients of the relationship between the lowest category versus all higher categories of the dependent variable are the same as those that describe the relationship between the next lowest category and all higher categories.

Figure 3 shows the graphical tool used in assessing the parallel slopes assumption using all observations. Since the relationship between all pairs of the dependent variable's category is the same, there is only one set of coefficients. The plots as displayed in this graph are predictions from the logit model used to model the probability that ages at death is greater than or equal to a given value using one risk factor at a time. The normalisation of all the first set of coefficients by setting them to be zero was done so that there is a common reference point for all risk factors.

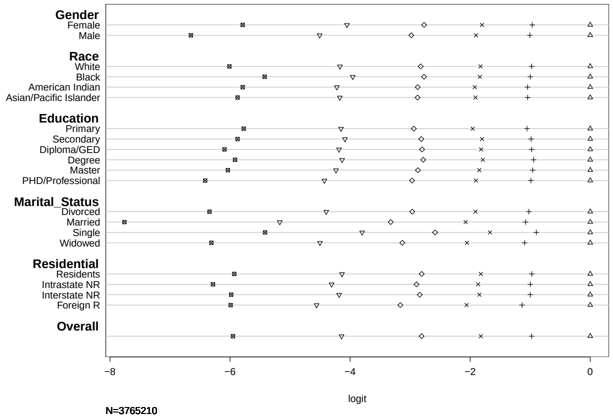


Fig. 3. Parallel slopes assumption.

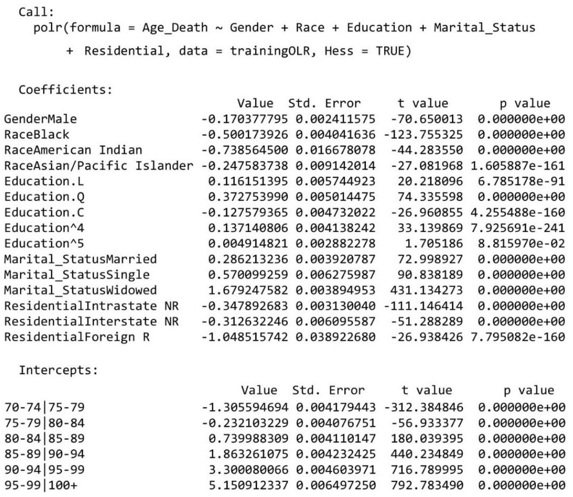
If the proportional-odds assumption holds, for each risk factor, the distance between the symbols for each set of categories of the dependent variable should remain similar. Looking at the coefficients for the variables paired in Figure 3, it can be seen that the distance between all sets of coefficients are almost similar which indicate that this dataset met the requirement for the proportional-odds assumption. In contrast, the markers are much further apart on the line for the married category under the marital status risk factor which suggests that this assumption may not hold only for this particular item.

D. POLR Model Fitting

The standard interpretation of the ordered logit coefficient is that for a one-unit increase in a particular risk factor, the dependent variable level is expected to change by its respective regression coefficient in the ordered log-odds scale given that the other risk factors in the model are held constant. The coefficients from the model can be somewhat difficult to interpret because they are scaled in terms of logs. Therefore, the coefficients are converted into proportional-odds ratios. The proportional-odds ratios are interpreted pretty much as would be for the odds ratios from a binary logistic regression.

The results of the POLR model fitting for this study as produced by RStudio can be found in Figure 4. The p-value for all the risk factors is less than 0.05, hence they are statistically significant in influencing age at death at a 95% confidence interval. These findings are similar to the findings for the Pearson correlation between the dependent variable and each risk factor in Figure 2.

The noticeable difference between the results produced by the multinomial regression model as compared to the results produced by the POLR model can be seen from the generation of intercept values between two categories of the dependent variable. These values represent the relationship between the lowest category versus all higher categories of the dependent variable. Mathematically, the intercept $70-74|75-79$ corresponds to $logit[P(Y \leq 1)]$ and can be interpreted as the log of odds of occurrence that a person is going to die between age 70 to 74 versus the log of odds of occurrence that a person is going to die within other age groups.



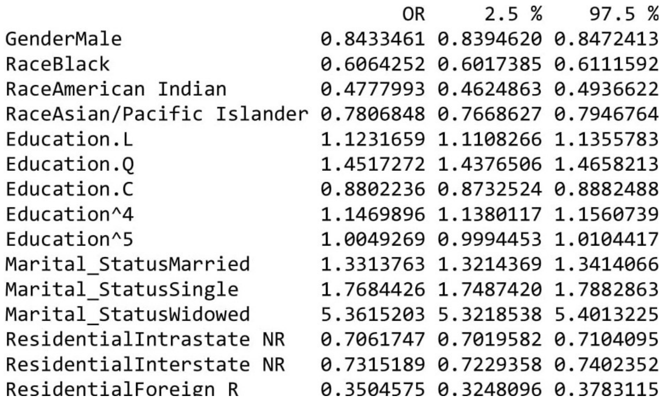
```
   Call:
     polr(formula = Age_Death ~ Gender + Race + Education + Marital_Status
         + Residential, data = trainingOLR, Hess = TRUE)

   Coefficients:
                                  Value  Std. Error    t value      p value
 GenderMale                 -0.170377795 0.002411575  -70.650013  0.000000e+00
 RaceBlack                  -0.500173926 0.004041636 -123.755325  0.000000e+00
 RaceAmerican Indian        -0.738564500 0.016678078  -44.283550  0.000000e+00
 RaceAsian/Pacific Islander -0.247583738 0.009142014  -27.081968 1.605887e-161
 Education.L                 0.116151395 0.005744923   20.218096  6.785178e-91
 Education.Q                 0.372753990 0.005014475   74.335598  0.000000e+00
 Education.C                -0.127579365 0.004732022  -26.960855 4.255488e-160
 Education^4                 0.137140806 0.004138242   33.139869 7.925691e-241
 Education^5                 0.004914821 0.002882278    1.705186  8.815970e-02
 Marital_StatusMarried       0.286213236 0.003920787   72.998927  0.000000e+00
 Marital_StatusSingle        0.570099259 0.006275987   90.838189  0.000000e+00
 Marital_StatusWidowed       1.679247582 0.003894953  431.134273  0.000000e+00
 ResidentialIntrastate NR   -0.347892683 0.003130040 -111.146414  0.000000e+00
 ResidentialInterstate NR   -0.312632246 0.006095587  -51.288289  0.000000e+00
 ResidentialForeign R       -1.048515742 0.038922680  -26.938426 7.795082e-160

   Intercepts:
                                  Value  Std. Error    t value      p value
 70-74|75-79                -1.305594694 0.004179443 -312.384846  0.000000e+00
 75-79|80-84                -0.232103229 0.004076751  -56.933377  0.000000e+00
 80-84|85-89                 0.739988309 0.004110147  180.039395  0.000000e+00
 85-89|90-94                 1.863261075 0.004232425  440.234849  0.000000e+00
 90-94|95-99                 3.300080066 0.004603971  716.789995  0.000000e+00
 95-99|100+                  5.150912337 0.006497250  792.783490  0.000000e+00
```

Fig. 4. The POLR model fitting.

The coefficients were then converted into proportional-odds ratios for ease of interpretation of the results since they are interpreted pretty much the same as the odds ratios from a binary logistic regression. The values of the proportional-odds ratios for each category of each risk factor can be found in Figure 5 below. These proportional-odds ratios were calculated based on a pre-selected baseline category from each risk factor, thus all interpretations must be done based on this pre-selected baseline. For example, for male, the odds of being more likely to live longer is 15.7% lower than female, given that all other risk factors are constant. Another example of interpretation, for widowers, the odds of being more likely to live longer is 5.36 times that of single people, given that all other risk factors are constant.



```
                                  OR     2.5 %    97.5 %
GenderMale                 0.8433461 0.8394620 0.8472413
RaceBlack                  0.6064252 0.6017385 0.6111592
RaceAmerican Indian        0.4777993 0.4624863 0.4936622
RaceAsian/Pacific Islander 0.7806848 0.7668627 0.7946764
Education.L                1.1231659 1.1108266 1.1355783
Education.Q                1.4517272 1.4376506 1.4658213
Education.C                0.8802236 0.8732524 0.8882488
Education^4                1.1469896 1.1380117 1.1560739
Education^5                1.0049269 0.9994453 1.0104417
Marital_StatusMarried      1.3313763 1.3214369 1.3414066
Marital_StatusSingle       1.7684426 1.7487420 1.7882863
Marital_StatusWidowed      5.3615203 5.3218538 5.4013225
ResidentialIntrastate NR   0.7061747 0.7019582 0.7104095
ResidentialInterstate NR   0.7315189 0.7229358 0.7402352
ResidentialForeign R       0.3504575 0.3248096 0.3783115
```

Fig. 5. The proportional-odds ratios and their 95% confidence intervals.

Some interesting findings can be discussed further from the results in Figure 5. For gender risk factor, a female American is more likely to live longer than a male American. For race risk factor, Black, American Indian and Asian/Pacific Islander people are less likely to live longer than White people; with American Indian to have more than 50% fewer odds compared to White. For education risk factor, a person with degree qualification is less likely to live longer than those with primary school qualification. For the other education categories, they are more likely to live longer than those who have primary school qualification. For marital status risk factor, being married, single or widowed are more likely to live longer compared to being divorced. For residential risk factor, intrastate non-resident, interstate non-resident or foreign resident is less likely to live longer than those who are a permanent resident of the United States.

The values of these proportional-odds ratios were then used to generate the ranking of longevity risk based on the likeliness of each category of each risk factor in influencing the longevity risk with respect to the pre-selected baseline category. However, depending only on these values would only generate a one-way ranking because the arrangements of the ranking were done based on only one category. A good risk ranking must also be able to explain the interaction between each category of each risk factor with each other. Thus, an effect analysis was carried out to overcome this situation.

E. Effect Analysis with Visualisation

The main purpose of doing an effect analysis is to determine the interactions between each category of each risk factor with each other. The first step in doing such analysis is to check if there are statistically significant differences between the independent variables. The most common statistical tests for analysing such relationships is the ANOVA test. However, using an ordinal data type as the dependent variable means that the assumption that the data follows a normal distribution will be violated. Given that the assumption of normality is violated, a typical ANOVA test in this situation would at best lack sensitivity, and at worst provide spurious estimates.

Fortunately, there are non-parametric versions of the ANOVA test which do not depend on the assumption of normality, and so are quite suitable for the ordinal data type. One of them is the Kruskal-Wallis test which is most appropriate for statistical testing between an ordinal level dependent variable and nominal level independent variables. This test assumes that the population has the same distribution, except for a possible difference in the population medians. The cross-sectional test for the dataset used in this study produced p-values of $2.2e^{-}16$ for each combination of risk factors. This shows that all of the independent variables are significantly different from one another with the p-value of each interaction is less than 0.05. Thus, the outcome of the dependent variable is influenced by each independent variable without any disturbance from the relationships among them.

Visualizing how the dependent variable responses with the changes across different level of the independent variables through effect display would help in determining the interactions between each category of each risk factor with each other. The effect display is carried out by allowing the independent variables to range over their combinations of values while holding other independent variables at fixed values. Figure 6 until Figure 15 show the effect of changing the value of the independent variables on the probability of being included into each level of the dependent variable based on a different combination of the independent variables while holding the other independent variables at fixed values.

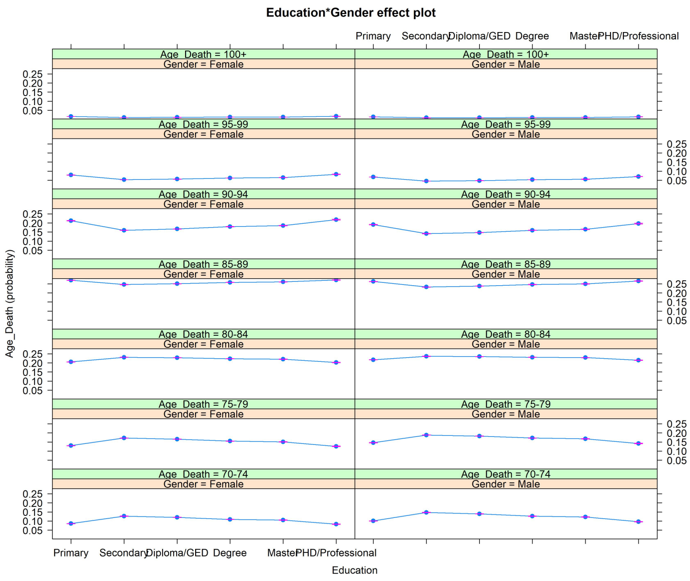


Fig. 6. Gender versus Race effect plot.

Based on Figure 6, it was found that, regardless of gender and race, the probability of dying at the age of 100 and above stays low and constant. Being a White person would increase the probability of dying within the age group of 85 to 89, regardless of gender. Changing the race value from White to Black would reduce the probability of dying between the age of 85 to 94, regardless of gender, with only a slight reduction in probability for the age group of 95 to 99. Being a male American would have more significant in the probability of dying between the age of 70 to 79, with a slight increment in probability for age 80 to 84. No dramatic differences between female and male categories.

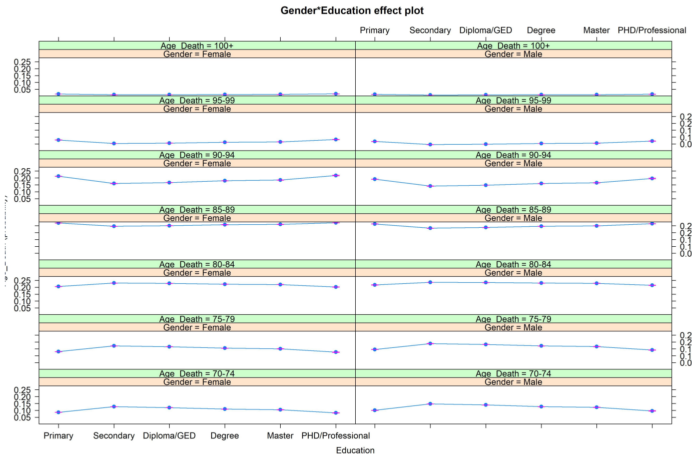


Fig. 7. Gender versus Education effect plot.

The effect of gender and education on the age of death also reveals some interesting information, as in Figure 7. The effect of changing the values of education level shows only small differences in the probability of dying across all ages for both genders. The probability of dying within the age group of 85 to 89 is dramatically high, regardless of gender and education level. The probability of dying at the age of 100 and above stays low and constant, regardless of gender and education level. There is a major drop in the probability of dying between ages 90 to 94 given that the level of education is changed from primary to secondary for both genders.

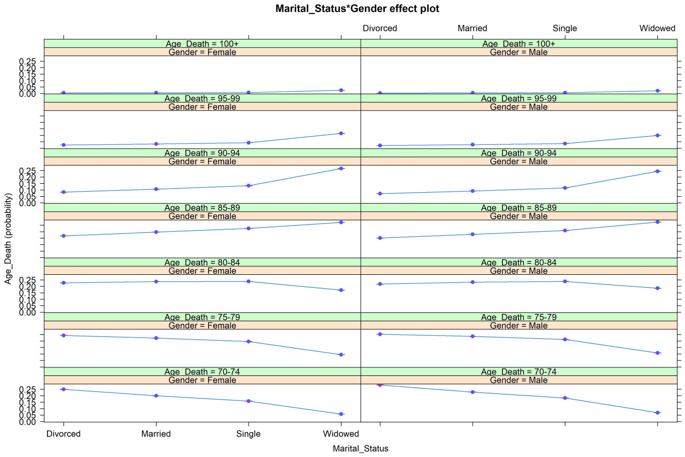


Fig. 8. Gender versus Marital Status effect plot.

Figure 8 shows the effect of gender and marital status on the age of death. The probability of dying at the age of 100 and above stays low and constant, regardless of gender and marital status. No significant differences in the effect of changing marital status on the age of death for both genders. There are some dramatic changes in the probability of dying for all age groups when the marital status is changed from single to widowed for both genders, especially for the age group of 70 to 74 and 90 to 94. There exist some reversed effects between these two age groups across all levels of marital status, for both genders.

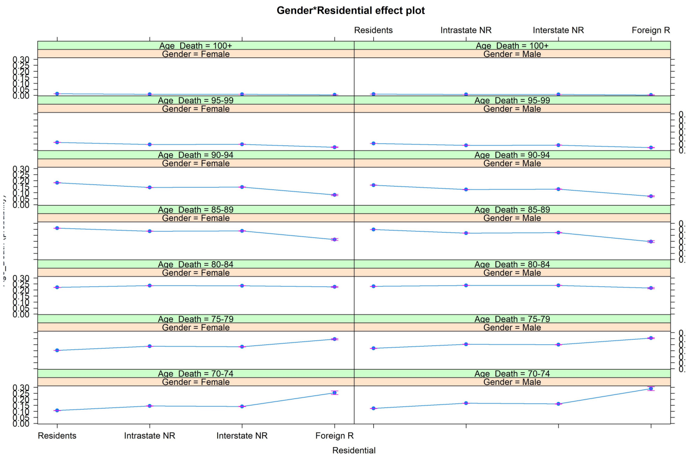


Fig. 9. Gender versus Residential effect plot.

The probability of dying at the age of 100 and above stays low and constant, regardless of gender and residential status, as shown in Figure 9. The effect of changing the values for residential status, for both genders, on the probability of dying within the age group of 95 to 99 and 80 to 84, is almost not visible. The effect of intrastate non-resident and interstate non-resident is almost the same for all age groups. However, dramatic changes can be seen between residents and foreign residents for all age groups regardless of gender. The most obvious effect could be found for the age group of 70 to 74.

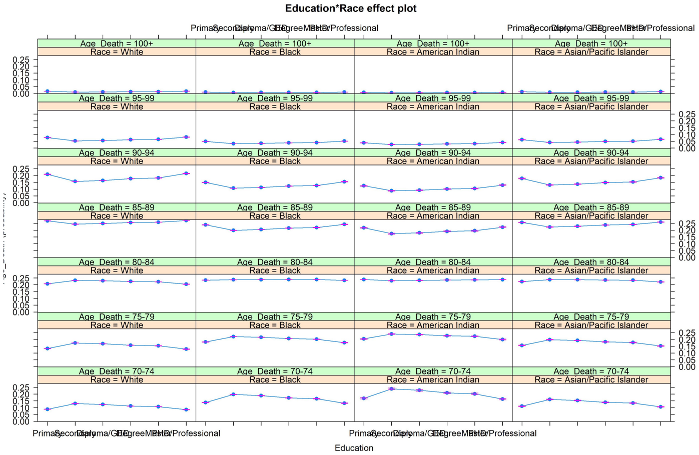


Fig. 10. Race versus Education effect plot.

Figure 10 shows the effect analysis for race and education level on the probability of dying for all age groups. The probability of dying at the age of 100 and above stays low and constant regardless of race and education level. The effect of all levels of education is the same across race on the probability of dying within the age group of 80 to 84 and 95 to 99. The probabilities of dying within the age group of 70 to 74 and 75 to 79 are highly affected by the changes in the education level of Black and American Indian people. The same thing can be said for the age group of 90 to 94, but with a reversed effect. The probability of a White person to die within the age group of 85 to 89 is constantly high, regardless of the level of education.

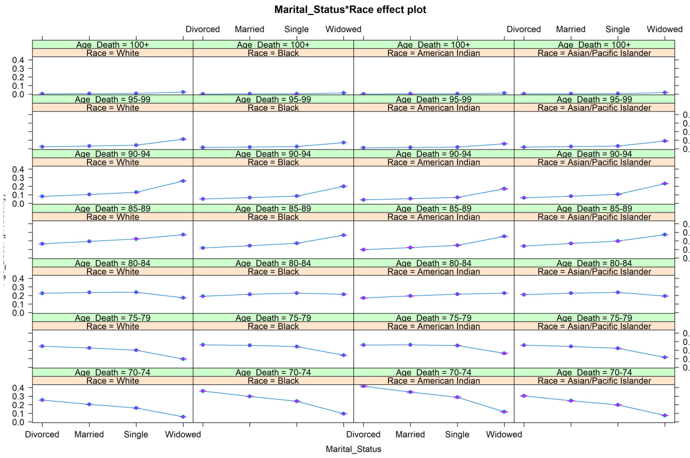


Fig. 11. Race versus Marital Status effect plot.

Figure 11 shows the effect of changing the values in the race and marital status on the probability of dying within each age group. The probability of dying at the age of 100 and above stays low and constant, regardless of race and marital status. Dramatic changes in the probability of dying for every age group are visible if there are changes in the marital status for all races. Being a divorced American Indian has the highest probability of dying within the age group of 70 to 74, whereby being a widowed person would have a higher probability of dying within higher age groups, regardless of race.

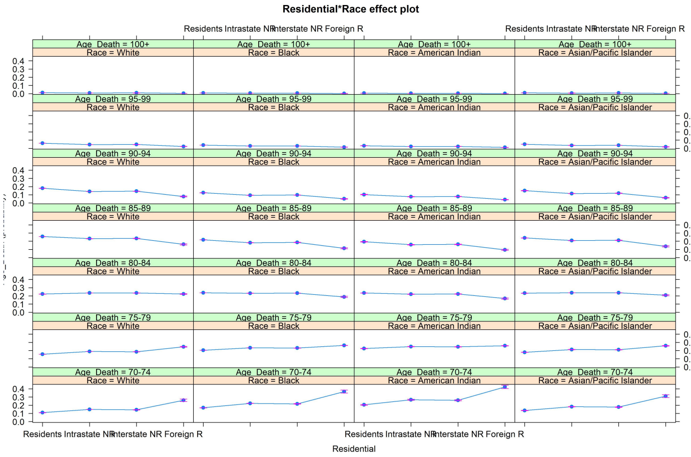


Fig. 12. Race versus Residential effect plot.

Figure 12 shows the effect of changing the values in the race and residential status on the probability of dying within each age group. The probability of dying at the age of 95 and above stays low and constant, regardless of race and residential status. The probability of dying within the age group of 80 to 84 stays moderate and constant, regardless of race and residential status. Dramatic changes in the probability of dying for the other age groups are visible, given that there are changes in the residential status for all races. The most dramatic changes can be seen in the probability of dying within the age group of 70 to 74 across all races. Based on this, foreign residents would have a higher probability of dying within this age group as compared to residents across all races, with the most dramatic increment for American Indian.

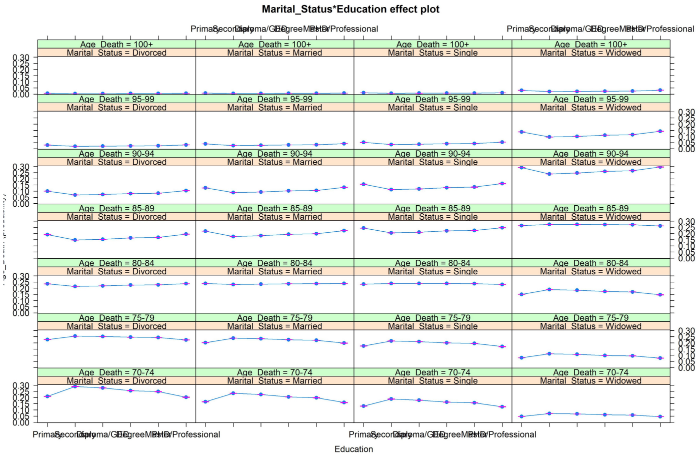


Fig. 13. Marital Status versus Education effect plot.

Figure 13 shows the effect of changing the values in marital status and education level on the probability of dying within each age group. The probability of dying at the age of 100 and above stays low and constant, regardless of marital status and education level. The effect of education levels stays low for age group 95 to 99 for divorced, married and single people. The most dramatic changes can be seen in the probability of dying within the age group of 70 to 74 for all education levels. Changing the marital status from divorced to widowed would dramatically reduce the probability, regardless of the education level. Being widowed will also increase the probability of dying within the age group of 90 to 94 which has recorded a very high increment.

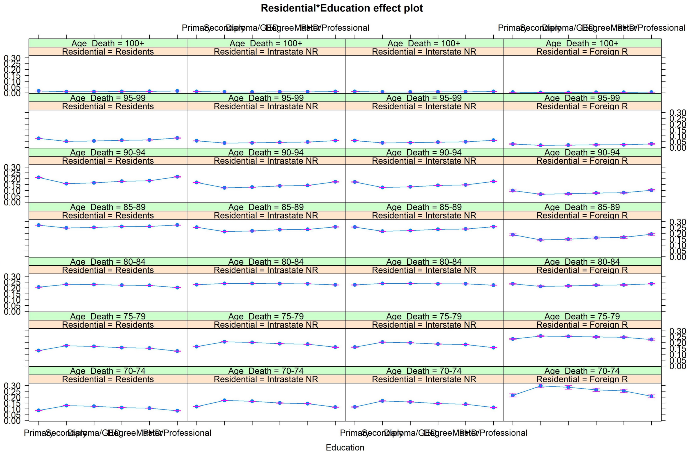


Fig. 14. Residential versus Education effect plot.

Figure 14 shows the effect of changing the values in residential status and education level on the probability of dying within each age group. The probability of dying at the age of 95 and above stays low and constant, regardless of residential status and education level. A reversed situation can be found in the age group of 80 to 84, whereby the probability of dying within this age group stays high and constant. The effect of changing the education level among foreign residents for the age group of 70 to 74 is very significant, whereby the probability of dying within this age group is higher than the other residential statuses.

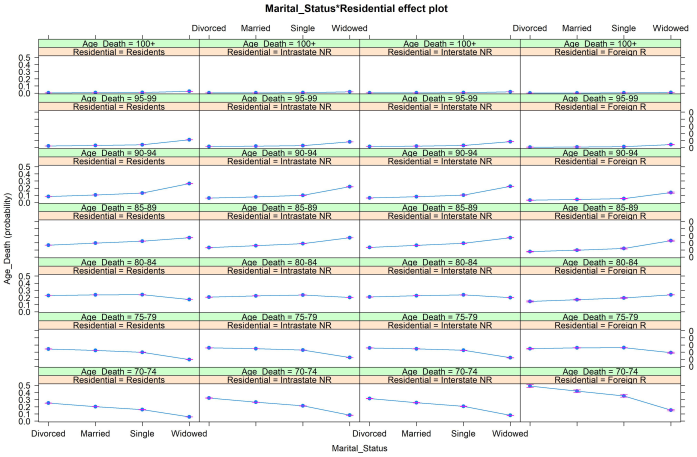


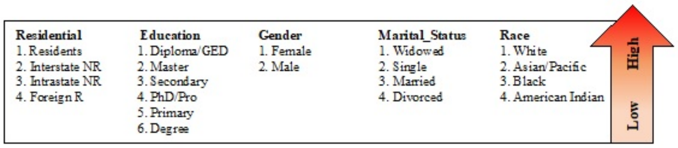
Fig. 15. Residential versus Marital Status effect plot.

Figure 15 shows the effect of changing the values in marital status and residential status on the probability of dying within each age group. The probability of dying at the age of 100 and above stays low and constant regardless of marital status and residential status. The probability of dying within the age group of 95 to 99 shows a slight increment between single and widowed people for all residential statuses. Dramatic changes in the probability of dying can be found for the age group of 70 to 74 across all marital statuses with the highest changes could be seen between residents and foreign residents. Changing the marital status of divorced to married would significantly reduce the probability of dying within this age group.

V. Longevity Risk Ranking

The findings from the odds-ratio analysis, along with the information abstracted from the effect display plots, were transformed into a meaningful risk ranking system, as shown in Figure 16. All levels of the risk factors were rearranged according to the effect level that they have on the probability of dying for each age group, whereby a higher age group represents a higher level of longevity risk. Based on the risk ranking diagram as shown in Figure 16, a high level of longevity risk is predicted to exist in those who are a White female widowed and the resident of the United States with the highest education level is diploma/GED. Those with such profile are predicted to live longer as compared to the other profiles.

One interesting characteristic of this risk ranking system is on the nature of each risk factor. At the beginning of this study, all the risk factors were treated as nominal variables. Rearranging them according to their level of effect on the probability of dying for all age groups has transformed them to be ordinal risk factors, thus improved their ability in providing more knowledge on the longevity risk. This ranking system is considered as a simpler mechanism compared to the more complex statistical methods. It simplifies the estimation of risks, increases the visibility of the risks and assists the policymakers in decision making.



Residential	Education	Gender	Marital_Status	Race
1. Residents	1. Diploma/GED	1. Female	1. Widowed	1. White
2. Interstate NR	2. Master	2. Male	2. Single	2. Asian/Pacific
3. Intrastate NR	3. Secondary		3. Married	3. Black
4. Foreign R	4. PhD/Pro		4. Divorced	4. American Indian
	5. Primary			
	6. Degree			

Fig. 16. Level of longevity risk according to the rank of each level for each variable.

VI. Conclusion and Recommendations

The POLR model has been proven to have good potential as an alternative model in ranking the ordinal risk factors' levels based on their individual effects on longevity improvements. A more meaningful ranking system has been developed based on a set of ordinal non-disease risk factors. This new ranking system has the potentiality of improving the ability of any statistical methods in projecting the longevity risk when using ordinal variables. Thus, it increases the visibility of the longevity risk and could be used to assist the policymakers in decision making.

However, some limitations need to be highlighted in this study. There is no means of classifying this ranking system and comparing it with the other countries, thus it's generalisation cannot be proved further and can only be applied for US longevity risk projections. As a recommendation for future researchers, the same methods used in this study could be applied to data from other countries and a comparative study should be conducted in comparing the findings from this study with other countries. It would be better if the procedures in this study could be replicated using data from various insurance companies and pension providers to see the impact that this method has in projecting the longevity risk within these industries.

References

[1] N. H. Hanafi and P. N. E. Nohuddin, "Data Mining Approach in Mortality Projection: A Review Study," Advanced Science Letters, vol. 23, no. 3, pp. 1612-1615, 2018.

[2] A. Bozikas and G. Pitselis, "An empirical study on stochastic mortality modelling under the age-period-cohort framework: The case of Greece with applications to insurance pricing," Risks, vol. 6, no. 2, p. 44, 2018.

[3] R. Zugic, G. Jones, C. Yiasoumi, K. McMullan, A. Tacke, M. Held and B. Moreau, "Longevity CRObriefing Emerging Risks Initiative Position Paper," CRO Forum, Amsterdam, 2010.

[4] S. Haberman, V. Kaishev, P. Millossovich, A. Villegas, S. Baxter, A. Gaches, S. Gunnlaugsson and M. Sison, "Longevity basis risk: A methodology for assessing basis risk," The Institute and Faculty of Actuaries & The Life and Longevity Markets Association, London, 2014.

[5] J. Bravo, P. Real and C. Silva, "Participating life annuities incorporating longevity risk-sharing arrangements," Portugal, 2009.

[6] A. D. Lopez, J. Salomon, O. Ahmad, C. J. Murray and D. Mafat, "Life tables for 191 countries: data, methods and results," World Health Organization, Geneva, 2001.

[7] R. I. Ibrahim, "Expanding an abridged life table using the Heligman-Pollard model," MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics, vol. 24, pp. 1-10, 2008.

[8] M. Denuit and J. Trufin, "From regulatory life tables to stochastic mortality projections: The exponential decline model," Insurance: Mathematics and Economics, vol. 71, pp. 295-303, 2016.

[9] H. Chulia, M. Guillen and J. M. Uribe, "Modelling longevity risk with generalized dynamic factor models and vine-copulae," ASTIN Bulletin: The Journal of the IAA, vol. 46, no. 1, pp. 165-190, 2016.

[10] L. Guo and M. C. Wang, "Data Mining Techniques for Mortality at Advanced Age," 2007.

[11] L. Guo, "Predictive Modeling for Advanced Age Mortality," in Living to 100 and Beyond Symposium, Florida, 2008.

[12] E. Daly, A. Mason and M. J. Goldcare, "Using mortality rates as a health outcome indicator: A literature review. Report to the Department of Health," National Centre for Health Outcomes Development, Oxford, 2000.

[13] S. Vinnakota and N. S. Lam, "Socioeconomic inequality of cancer mortality in the United States: a spatial data mining approach," International journal of health geographics, vol. 5, no. 1, p. 9, 2006.

[14] M. Heron, "National vital statistics reports," National Center for Health Statistics, 2007.

[15] World Health Organization, "Global health risks: mortality and burden of disease attributable to selected major risks," World Health Organization, Geneva, 2009.

[16] P. Berry, L. Tsui and G. Jones, "Our New 'Old'Problem–Pricing Longevity Risk in Australia," in 6th International Longevity Risk and Capital Markets Solutions Conference, Sydney, 2010.

[17] Office for National Statistics, "Mortality Statistics: Metadata," Office for National Statistics, South Wales, 2015.

[18] D. Allen and S. Lee, "Modelling Life Insurance Risk Prudential Insurance Data Set," in SAS Student Symposium Forum, 2018.

[19] D. Drljača, "Risk assessment through matrix model in insurance companies," Poslovna ekonomija, vol. 10, no. 2, pp. 43-65, 2016.

[20] O. Abul-Haggag and W. Barakat, "Application of fuzzy logic for risk assessment using risk matrix," International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 1, pp. 49-54, 2013.

[21] P. Horgby, "Risk classification by fuzzy inference," The Geneva Papers on Risk and Insurance Theory, vol. 23, no. 1, pp. 63-82, 1998.

[22] R. Brant, "Assessing proportionality in the proportional odds model for ordinal logistic regression," Biometrics, pp. 1171-1178, 1990.

[23] R. Bender and U. Grouven, "Using binary logistic regression models for ordinal data with non-proportional odds," Journal of clinical epidemiology, vol. 51, no. 10, pp. 809-816, 1998.

[24] R. Williams, "Generalized ordered logit/partial proportional odds models for ordinal dependent variables," The Stata Journal, vol. 6, no. 1, pp. 58-82, 2006.

[25] H. Wickham, J. Hester and R. Francois, "readr: Read Rectangular Text Data," 2018.

[26] H. Wickham, "The Split-Apply-Combine Strategy for Data Analysis," Journal of Statistical Software, vol. 40, no. 1, pp. 1-29, 2011.

[27] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, New York: Springer-Verlag New York, 2016.

[28] R. H. B. Christensen, ordinal: Regression Models for Ordinal Data, 2019.

[29] R Core Team, R: A Language and Environment for Statistical Computing, Vienna: R Foundation for Statistical Computing, 2019.

[30] J. Fox and J. Hong, "Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the effects Package," Journal of Statistical Software, vol. 32, no. 1, pp. 1-24, 2009.

[31] J. Fox and S. Weisberg, "Visualizing Fit and Lack of Fit in Complex Regression Models with Predictor Effect Plots and Partial Residuals," Journal of Statistical Software, vol. 87, no. 9, pp. 1-27, 2018.

[32] J. Fox and S. Weisberg, An R Companion to Applied Regression, 3rd ed., CA: Thousand Oaks, 2019.

Lexical Variation and Sentiment Analysis of Roman Urdu Sentences with Deep Neural Networks

Muhammad Arslan Manzoor[1], Saqib Mamoon[2], Song Kei Tao[3], Ali Zakir[4], Muhammad Adil[5], Jianfeng Lu*[6]
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing, China

***Abstract*—Sentiment analysis is the computational study of reviews, emotions, and sentiments expressed in the text. In the past several years, sentimental analysis has attracted many concerns from industry and academia. Deep neural networks have achieved significant results in sentiment analysis. Current methods mainly focus on the English language, but for minority languages, such as Roman Urdu that has more complex syntax and numerous lexical variations, few research is carried out on it. In this paper, for sentiment analysis of Roman Urdu, the novel "Self-attention Bidirectional LSTM (SA-BiLSTM)" network is proposed to deal with the sentence structure and inconsistent manner of text representation. This network addresses the limitation of the unidirectional nature of the conventional architecture. In SA-BiLSTM, Self-Attention takes charge of the complex formation by correlating the whole sentence, and BiLSTM extracts context representations to tackle the lexical variation of attended embedding in preceding and succeeding directions. Besides, to measure and compare the performance of SA-BiLSTM model, we preprocessed and normalized the Roman Urdu sentences. Due to the efficient design of SA-BiLSTM, it can use fewer computation resources and yield a high accuracy of 68.4% and 69.3% on preprocessed and normalized datasets, respectively, which indicate that SA-BiLSTM can achieve better efficiency as compared with other state-of-the-art deep architectures.**

Keywords*—Sentiment analysis; Self-Attention Bidirectional LSTM (SA-BiLSTM); Roman Urdu language; review classification

I. Introduction

Sentiment analysis is a fundamental task that classifies the feedback, feelings, emotions, and gestures in natural language processing domain [1]. Recent theoretical developments have revealed that every discussion on social media, forums, blogs, chats has a great influence on society regardless of the region or the language. This situation is considerable for vast number of societies and business communities in terms of feedback, to conquer deficiencies and enhance productivity. The growing demand for the computational learning of text, further results in sentence classification, aspect categorization, and opinion detection.

Convolutional Neural Networks (CNNs) have achieved impressive results on the important task of sentence categorization [2]. Further, Recurrent Neural Networks (RNNs) and their variants such as LSTM, BiLSTM, and GRU have produced better results for sequence and language modelling [3], [4]. Previous studies show that most of the neural networks require more time and memory resources to train and run the model and difficult to optimize. A solution to this problem is proposed by Bahdanau et al. [5] which emphasized that Attention keeps track of the source input sequence by building a shortcut between encoder hidden states and context vector. This study gave a great break through at the Language Modeling Planet by introducing Transformer [6] that is merely based on Attention mechanism, drops off recurrence and convolutions thoroughly in terms of training and striking results. In terms of training performance, Attention mechanism is more stable as there are no large number of hidden states to update and maintain. After it, Attention networks have been applied to multiple tasks [7]–[9] i.e., image classification, text summarizer , and sentiment analysis.

Another key limitation is that most of these models are unidirectional. In order to address this issue, a novel framework "The Self-Attention Bidirectional LSTM (SA-BiLSTM)" is proposed in this study. In SA-BiLSTM, Self-Attention mechanism focuses only the relevant word embedding to correlate in the whole sentence which influences polarity and BiLSTM supervise context representations of these attended embedding in forward and backward direction. Studies mentioned above are evidenced that Self-Attention can produce better result and consume less resources because of its selective nature and Bidirectional LSTM is integrated to conquer the limitation of unidirectional model.

A lot of research work has been done on English language Analysis [10]. According to our best knowledge, no previous research is carried out to classify the sentences of subcontinent language (Urdu/Hindi) with Neural Networks. Urdu is the native language of Pakistan and currently being spoken and understood in several parts of India, Bangladesh, and Nepal [11]. Roman Urdu is one of those languages which is usually used on social media for communication and comments [12]. There is no dataset of Roman Urdu available that is ready to apply deep learning models. The most challenging task was to preprocess and normalize the Roman Urdu dataset then made it usable for Sentiment Analysis.

In view of the existing gap, our contributions is as follows:

- Preprocessed the 10,000 Roman Urdu Sentences (negative and positive reviews), and normalized more than 3000 sentences.

*Corresponding author

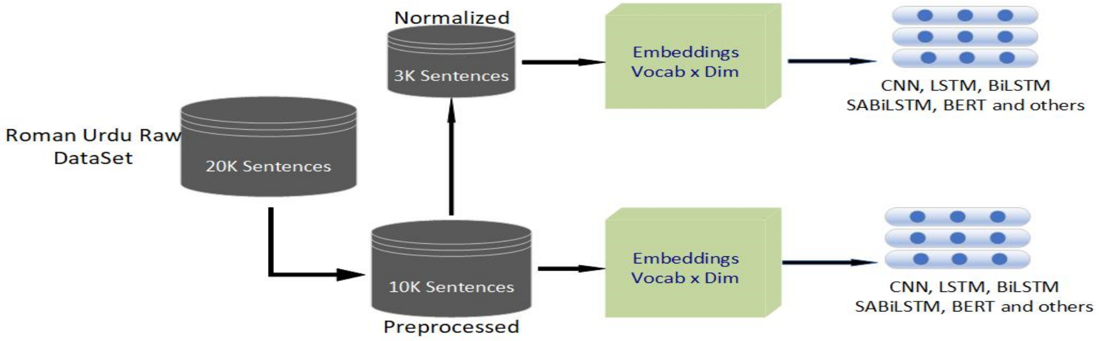


Fig. 1. Dataflow diagram from preprocessing of dataset to implementing the models

- Bidirectional LSTM is integrated with Self-Attention to deal with the complexity of sentences and variations of words in Roman Urdu.
- Self-Attention with Bidirectional LSTM (SA-BiLSTM) is trained and evaluated on Roman Urdu sentences, then comparison and analysis are made with other models.

This study has made a significant contribution by addressing the problems of Roman Urdu scripts with deep neural networks. The layout of this paper is as follows. Initially, section II describes the existing networks for NLP tasks. Section III proposes the language architecture in methodology. Then, the Section IV reveals the dataset preprocessing and experiments. Moreover, Section V illustrates the results and analysis. Finally, Section VI concludes the research work and opens the future.

II. Related Work

In recent years, deep learning has become the key technique for many researchers to deal with sentiment analysis. It consists of effective models that are used to solve a variety of problems efficiently [13]. Convolution Neural Network (CNN) has achieved impressive results on the task of sentence categorization [2], [14]–[17]. The models need not be complex to realize strong results [2], regarding visual sentiment analysis, CNN enhanced its efficiency by growing its size and depth [18]. Results show that the proposed system achieves high performance without fine-tuning. Detailed research [19] using Convolutional Neural Network (CNN) has presented a summary of sentiment analysis related to micro-blog. However, in terms of sentence processing, CNN extracts the feature without correlating all the sentence that leads to producing low results and consume high resources.

Recurrent Neural Network (RNN) is well known for sequential information processing as they use internal memory states to process input sequences [20]. It produces output that is dependent on the computation of all previous input and hidden states. RNNs prefer terms that they get later in the sentence, despite the words they get earlier. RNN lacks in most applications because they demand high memory, time and hardware resources.

To deal with the shortcoming of standard RNN, researchers have developed sophisticated variants of RNN [21]. Bidirectional RNN is built on the idea that the outcome at each time may not only bases on the previous elements but also depends on the next elements in the sequence. LSTM cell uses forget gate, input gate and output gate for processing cell states to focus most concerning information which enhances the performance of this cell [22]. Gated Recurrent Unit (GRU) combined the forget gate and input gate to make it simpler but less efficient than LSTM for long sequences and large datasets [3]. BiLSTM (Bidirectional long short-term memory) is an extended version of LSTM with more information [23]. BiLSTM access both the preceding and succeeding contexts by considering the forward and the backward hidden layers employed by Chen et al. [24] for sentiment analysis task. All these variants of RNNs have achieved great success in numerous tasks. However, they are often called as black boxes, lacking interpretability and consume high resources [25]. Research efforts to solve this issue have steadily increased.

The Attention mechanism was presented to upgrade the RNN encoder decoder sequence-to-sequence network for NMT [5], [26]. Initially, Attention was defined as the process of determining a context vector for the next decoder step that consists of the most relevant information with the encoder hidden states. Seminal contributions have been made by Vaswani et al. [6] when Transformer architecture was proposed for machine translation. It depends only on Attention mechanisms, as the best replacement of either recurrent or convolution neural networks. For sequence processing and language modeling, Transformer has outperformed the recurrent neural network and their variants.

A closer look to the literature on neural networks for sentence classification [5], [6] reveal that Attention predicts based on only recent hidden states (unlike RNN, which predict based on entire history and reminds all the previous hidden states). The objective is to devise and implement a system that consists of Self-Attention to address the problem of complex structure of Roman Urdu Sentences. In this study, a more efficient and lightweight model Self-Attention Bidirectional LSTM is proposed for targeted problem, where Self-Attention takes charge of the complex formation by correlating the whole sentence and determining embedding that consists of the most relevant information. Bidirectional LSTM is integrated to strengthen the network as it extracts context representations to tackle the lexical variation of attended embedding in preceding and succeeding directions. Moreover, it promotes essential embedding by memorizing the contextual information for the long term. The results endorse that the integration of network leads to enhance the Self-Attention's performance. Besides, deficiencies of Bidirectional LSTM are conquered by Self-Attention module in the network.

III. Methodology

According to Bahdanau et al. [5], Attention's task is to compute the context vector for the succeeding decoder step that consists of maximum appropriate values of encoder hidden states after getting a weighted average of encoder hidden states. There was a factor of alignment score which represents the contribution to the weighted average between encoder states and previous decoder hidden states.

A. Self-Attention Mechanism

Following the above concept, Vaswani et al. [6] trained decoder hidden states as query vector which pay Attention to those hidden states of an encoder that have more influence in producing relevant output. Key, Value vectors are formed by hidden states of Encoder. Attention does not always take two different sentences and correlate them, it may take same sentence along column and row to extract the relation between different parts of it. Each sequence position is considered as Q and compared with the rest of sequence position K by correlating them and as a result V is produced that has most weighted relevance (Self-Attention). Initially, compatibility function determines the weights connecting the query and the keys in (1). Compatibility score is transformed by the softmax function into probability distribution as described in (2), this normalization helps Query (q) to consider the important tokens Key (k) for classification. Then, weighted average of Value (v) vectors corresponding (k) produced output. Feed forward layers and learned linear projections were applied to create (query, value, key) vectors. Taking a query q, values and keys, compatibility function is responsible to compute correlating outcome between k and q as follows

$$f(k,q) = \frac{(k)(q)^T}{\sqrt{d_k}} \tag{1}$$

d_k is served as a scaling operator, and maintains the numerical stability when the dimension of keys increases. The softmax function is applied to the compatibility score to compute Weighted sum α.

$$a = softmax\,\{f(k,q)\} \tag{2}$$

$$Z = \sum a(v) \tag{3}$$

Equation (3) represents the most relevant values with the query selected by the highest weights.

B. Bidirectional LSTM

Attention output is passed to Bidirectional LSTM to memorize only the most considerable Self-Attended preceding and succeeding embedding which would enhance the accuracy. LSTM support to reminisce those embeddings by means of its three gates architecture to impact the results efficiently. To strengthen the LSTM and deal the weakness (not accessing the forward hidden layers for the future token), Bidirectional LSTM is used to collect the contextual and relevance information from previous and future embedding values.

C. Positional Information

Input embeddings gather the positional Information of sequence ordering through the Position Embedding Layer. Absolute (or relative) positional information of each token in a sequence is passed to Attention layer. A method is proposed where positional encoding (PE) vectors are formed using sine and cosine functions of difference frequencies and then are appended to the input embeddings [6].

D. Network Architecture (SA-BiLSTM)

The Network architecture of SA-BiLSTM is represented in Fig. 2. The input sentences are passed through the embedding layer that uses pre-trained embedding generated from Word2Vec model. Embeddings are passed to Position Encoding (PE) to learn position representation as Attention receive the whole sequence and does not keep the positional information. The output embeddings of PE are sent to SA-BiLSTM module which is described below:

Self-Attention mechanism is applied to every position of source sentence. For each sentence position query, key, and value would behave as vectors. The absence of previous decoder state that behave as query made every position of input sequence as a set of Query vectors. In this step, by Keeping each query static, compatibility score with all the rest of keys of that sequence would be measured. It is applied to all values of vectors $O = (o_1, o_2, ..o_n)$ which creates Output vector which has the information of how much each sequence query is relative to the rest of queries and contributes in polarity of sentence.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

It led to multi-head Attention which implements parallel computing on the whole sequence by making groups of the query, keys, and values in Q, K, and V matrices respectively. After the above study, we propose a Self-Attention Bi-LSTM Sequential Model (SA-BiLSTM) for the normalized dataset which consists of three major sections. Input Sentences are converted into Embedding with Q, K and V vectors which are passed through the Position Embedding module that brings consideration to sequence ordering. These Embedding with positional Information is passed to Self-Attention Module which applies Attention mechanism as in (4) on each sequence position. It helps in a correlating the weights. Multi head Attention performed Attention h times on (Q, K, V) matrices of dimension d_{model}/h in (5). where each head performed Self-Attention to produce an output of dimension d_{model}/h in (6). Then the outputs are concatenated to produce matrices of

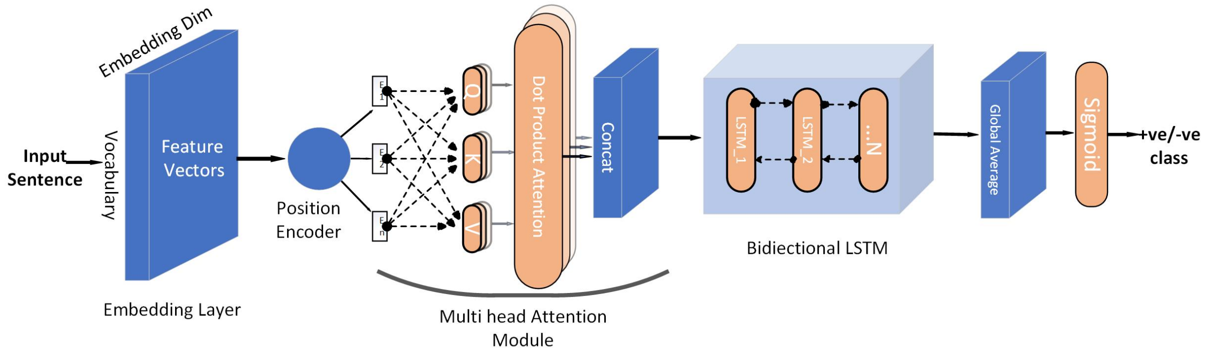


Fig. 2. Network Architecture of SA-BiLSTM

identical dimensionality to Self-Attention on the actual (Q, K, V) matrices. Feed forward layers pass the embedding to next module.

$$MultiHead(Q,K,V) = Concat(head_1, ..., head_h)W \quad (5)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (6)$$

These Self-Attended Embedding sent to stack BiLSTM for contextual semantic information on the backward and forward direction of embeddings. It selects those embedding which is going to influence polarity more by memorizing its previous effect in both directions.

$$\tilde{c} = tanh(W_c[h_{t-1}, emb] + b_c) \quad (7)$$

$$c_t = f_t.c_{t-1} + i_t.\tilde{c_t} \quad (8)$$

$$h_t = o_t.tanh(c_t) \quad (9)$$

Equation (7) denotes the input to the cell, emb is the value selected by Self-Attention passed as input and controlled by hyperbolic tangent function, W_c and b_c are learnable parameters, ht-1 is the hidden state value of previous time step. In (8) and (9), i_t, $f_t$and o_t are the input, forget and output gate values activate by sigmoid function at time t, respectively. Cell state at current time step is denoted as c_t and c_{t-1} is cell state for previous time steps. The final output of cell at time t is filtered by the output gate denoted as h_t. Global average pooling is applied to the final output of the BiLSTM. Finally, sigmoid classifier can output the class of sentence.

IV. Experimental Setup

This section describes the experimental setup to analyze the performance of SA-BiLSTM for Roman Urdu sentiment analysis. All the well-known deep learning language models along with the proposed model are implemented on two datasets (preprocessed and normalized) of Roman Urdu sentences. Each sentence of the dataset is labeled with positive or negative tag 0,1. Experiments were run using a single Titan Pascal XP 12G. All models are implemented in Keras 2.2.4 with Tensorflow-GPU 1.13.1 backend using cuDNN 7.3.1 and CUDA 10.1.

A. Parameter Setting

Extensive experiments were run using adadelta, sgd, rmsprop, and adam optimizer. After ablation study, it is observed that adam optimizer using a 0.0004 learning rate with batch size of 32 achieved more stable results than rest of the optimizers. Cross-entropy loss with L2-regularization is performed on the model parameters with a λ value 10-3. The dropout value was kept 0.3 to avoid underfitting.The word embedding of 200 dimension created by word2vec model. For the activation purpose in final dense layer, sigmoid function is used.

B. Dataset

The dataset that is used to evaluate the deep learning models for Roman Urdu analysis is comprised of sentences extracted from Urdu blogs, social and news websites, prepared by Sharf et al. [27], where reviews written by customers, such as social media users and fan followers of celebrities. The dataset available at resource* which contains more than 20,000 sentences (positive, negative and neutral) that belongs to 4 to 5 domains of online platforms.

TABLE I. Sample Sentences of Roman Urdu from Dataset

Roman Urdu	English
usay saalgira per **khoobsurat** tohfa mila	He got a beautiful birthday present
wo aik **kamyaab** shakhs hai	He is a successful person
us movie ka subject bohat **acha** hai	The theme of this movie is very good
isay **bura** samjha jata hai	This is considered bad

*1) **Preprocessed dataset:*** We mainly focused on binary (positive, negative) classification and selected 10,000 most appropriate sentences from resource for preprocessing and termed it as “Preprocessed” dataset in experiment. To make the dataset more reliable, non-textual symbols and characters were removed so as to implement language models for the Roman Urdu analysis and evaluation of proposed model. The sample sentences from the dataset are shown in Table I with English translation for better understanding the script style of Roman Urdu.

Lexical Variation in Preprocessed Dataset: The Roman script does not follow any standard which makes it more complicated than English language dataset. Different spelling

*https://archive.ics.uci.edu/ml/datasets/Roman+Urdu+Data+Set refer to same word and identical spelling refer to different contextual words. This phenomenon confuses the embedding of vocabulary and motivated us for normalization of Roman Urdu sentences. Previously, some approaches have been used for normalization purposes to reduce the variation of embedding for the same word: Urdu phone, Similarity function, Lex-C clustering algorithm, Stemming and Lemmatizing [27], [28]. These approaches depend upon some rules and there is 30% to 40% chance of failure attributed to these rules. Making a set of similar words and clipping suffix or prefix can negatively influence the embedding behavior towards sentence polarity.

Standards were followed to apply lexical normalization and standardization of words. As discussed in [29] each word of the Urdu language should follow the standard spelling as it comes to Roman transliteration of Urdu terms. Unification of vocabulary (where each word refers by unique characters not multiple combinations of characters) was done by same person to maintain the consistency for the whole dataset.

*2) **Normalized Dataset**:* From the preprocessed dataset, the sentences that have more polarized words are normalized manually for unification of vocabulary. Three different categories were mainly focused of preprocessed to normalize. For the sake of generalization and avoiding overfitting, sentences belong to different categories from different sources are normalized. These categories include news, reviews about celebrities and feedback about products received through online shopping. All of these categories have equal number of positive and negative sentences.

TABLE II. Lexical Variation of Words (from Roman Urdu Sentences in TableI)

English	Preprocessed Roman Urdu	Normalized Roman Urdu
Successful	kamiyaab, kamyab, kaamyaab, kamyaab	kamyaab
Beautiful	khubsurat, khubsoorat, khoobsurat, khoobsoorat	khoobsoorat
Good	acha, achi, ache	same as in preprocessed
Bad	bura, buri, bure	same as in preprocessed

The lexical variation of the words is represented in Table II that influences the polarity of the sentence. The Roman Urdu terms *kamyaab, kamiyaab, kamyab, kaamyaab* for *successful* and *khoobsoorat, khubsurat, khubsoorat, khoobsurat* for *beautiful* in first two sentences of Table I are normalized to *kamyaab* and *khoobsoorat* respectively as shown in Table II. Roman Urdu terms *acha, achi, ache* for *Good* and *bura, buri, bure* for *Bad* in the next two sentences of Table I depend upon the gender and number of subject word (singular or plural). Therefore, these terms are not normalized and will remain the same as in preprocessed dataset. Even though the normalization process increases the accuracy as mentioned in [30] but the existence of this limitation in the Urdu language leads to produce low results as compared to other languages. The resultant dataset called as normalized dataset. Considering the time limit and assessing the performance improvement of model, we normalized 3000 Sentences.

C. Effect of Normalization

The similarity of embedding-vectors measured by cosine distance sort the words in the vocabulary according to their ”similarity” in the embedding-space. Table III (a) indicates that similar embedding vectors of different words belonging to the same class (have same contextual meanings) appeared to give high results. Table III (b) indicates that words have different or opposite contextual meaning belonging to different classes must indicate less similarity and give low results. Tables are prepared of the same word for preprocessed (unnormalized) and normalized dataset. From results, it can be observed that similar words belonging to the vocabulary of normalized dataset show higher similarity than a preprocessed dataset. Those words that are contextually less similar show lower results for the vocabulary of a normalized dataset. Reason lies in the unification and Normalization of terms. A different variation of a word creates ambiguity that leads the network to become less efficient despite having a large number of sentences to train the model. On the other case, a small dataset with unique terms and no multiple variations for the same word make the dataset consistent on which network produces better results. For example, word successful *kamyaab* has multiple variations *kamiyaab*, *kamyab*, *kaamyaab* in Roman Urdu that has changed to one term *kamyaab* in the normalized dataset. This unique word has the highest similarity with itself as compare to different variation, this reason influences the results considerably.

TABLE III. Similarity %age in Normalized Dataset Denoted by Norm %Sim in the Table and Similarity %age in Preprocessed Denoted by Preproc %Sim in Table

(a) This table represents the similarity percentage of Similar words that belong to same class.

Similar word	Norm %Sim	Preproc %Sim
Pyar;sakoon; [love; calm]	**99**	92
Qeeemati;khoobsoorat; [Expensive; Beautiful]	**98**	95
Shohrat;Fatah; [Fame; Victory]	**99**	96

(b) This table represents the similarity percentage of dissimilar words that belong to different class.

Dissimilar words	Norm %Sim	Preproc %Sim
Zakhmi;sehat; [Injured; Health]	**70**	75
Janbahaq;zinda; [Died; live]	**58**	62
Shohrat;badnam; [Fame; disgrace]	**71**	76

V. Results and Analysis

This section exhibits the results of all language models evaluated on preprocessed and normalized Roman Urdu sentences. Comparison of experimental results, finding and contributions are discussed. Table IV and Table V contain multiple matrices including testing accuracy, recall (true positive rate) and precision (positive predicted value) to assess the efficiency of each language model as shown in equations (10), (11), and (12) respectively. Moreover, time complexity is represented as subscript of testing accuracy to show the time taken by each experiment for corresponding language model. Besides, accuracy on testing dataset and utilization of time resources, recall and precision show the exactness(quality) and completeness(quantity) of each language model.

$$Accuracy = \frac{tp + tn}{(tp + tn + fp + fn)} \tag{10}$$

$$Recall = \frac{tp}{(tp + fn)} \tag{11}$$

$$Precision = \frac{tp}{(tp + fp)} \tag{12}$$

The results demonstrate two aspects for the evaluation of the language models. First, the performance measure matrices including recall, precision, and accuracy. Second, the time in seconds represents the time complexity of the experiment. The slight difference in results for different evaluation metric indicates the consistent performance of the model.

TABLE IV. EXPERIMENTAL RESULTS ON PREPROCESSED DATASET.

Language models	Recall	Precision	$Accuracy_{seconds}$
Fasttext	63.8	62.3	$62.4_{(220)}$
CNN	60.6	60.6	$60.6_{(400)}$
LSTM	66.2	66.4	$66.2_{(362)}$
BiLSTM	66.9	67.0	$66.9_{(465)}$
Self-Attention	66.9	66.6	$66.8_{(230)}$
SA-BiLSTM	68.5	68.4	**68.4**$_{(260)}$

Results in the Table IV confirm these findings: CNN is least efficient in accuracy and time complexity as CNN does not extracts the important embedding from sentences as compared to Attention mechanism. Fasttext is an agile network and produce better results than CNN but still it is far low than other models as Fasttext is not as deep. The results produced by LSTM and BiLSTM (RNN variants) on Roman Urdu sentences is remarkably high than former networks. However, limitation of these methods are that they consume high time and memory cost. Results depicts that Self-attention is efficient in time memory complexity and achieves comparable accuracy with LSTM, BiLSTM. Therefore, it is generally accepted that Self-Attention addresses the issues arise in other RNN variants.

The proposed network, SA-BiLSTM outperformed all neural network by achieving highest accuracy of 68.4%. From the results, it is clear that proposed network utilized less time resources than CNN, LSTM and BiLSTM. These results support the effectiveness of model by attaining better outcome on all matrices when compared with other language models. The selective and bidirectional architecture of SA-BiLSTM results in the highest accuracy for complex sentence structure of Roman Urdu script possessing lexical variation of words.

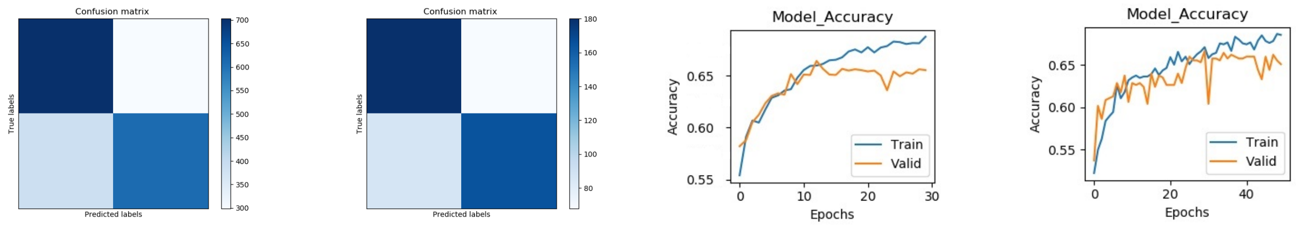
TABLE V. EXPERIMENTAL RESULTS ON NORMALIZED DATASET.

Language models	Recall	Precision	$Accuracy_{seconds}$
Fasttext	62.1	62.2	$62.1_{(100)}$
CNN	64.6	65.0	$64.6_{(214)}$
LSTM	67.8	68.1	$67.8_{(150)}$
BiLSTM	67.7	67.9	$67.6_{(240)}$
Self-Attention	67.2	67.1	$67.0_{(90)}$
SA-BiLSTM	69.4	69.3	**69.3**$_{(120)}$

Normalized dataset is 73% smaller in size than preprocessed dataset in terms of sentences and pre-trained word vectors. In spite of this fact, all language models have produced better results on this dataset even if the improvement is negligible as shown in Table V. Contrary to Previous experiments, CNN has yielded higher accuracy than Fasttext which shows that CNN performs well on Normalized dataset. The results produced by LSTM, BiLSTM and Self-Attention on Normalized dataset are in line with trend of results on preprocessed dataset. The proposed model delivered significantly better results for all matrices and highest accuracy i.e. 69%. It can be seen that SA-BiLSTM supersedes the existing models in all metrics. Even though deep learning models achieved adequate results, the limitation we faced thoroughly in the experiments was the unavailability of large pre-trained word embeddings due to the absence of a massive dataset like Google News or Wikipedia. As the previous study mentioned that less pre-train embedding did not produce good results, despite, the performance of SA-BiLSTM on Roman Urdu dataset is in line state-of-the-artwork.

Additionally, these results endorse our claim that consistent dataset with more polarized sentences, having normalized vocabulary can produce more efficient results, although it has trained on a smaller number of sentences and pre-trained word embedding. The confusion matrix shown in Fig. 3 upholds the normal behavior of proposed model. From Fig. 3a and 3b, it is obvious that SA-BiLSTM succeeded in detecting true positive and true negative by giving strong confusion matrix for normalized dataset.
Fig. 3 expresses the accuracy curves of model on preprocessed and normalized dataset respectively. The accuracy curve on training and validation set of data represent the learning and generalizing ability of SA-BiLSTM. For the case in Fig. 3c, the accuracy curve on validation data displays that the experiment stopped earlier (in 30 epochs). The curve hits maximum accuracy of 67% for training and and 65% for validation. The existence of noise in embedding space (in the preprocessed dataset), restricts the model to learn after reaching at certain limit (65%), even with more number of sentences and a large number of pre-trained embedding. It can be seen in Fig. 3d that the model accuracy curves enter in stabilized region after 40 epochs. It depicts that the model learns the features smoothly in more than 40 epochs. Even though the less number of sentences and short vocabulary to create pre-trained embeddings, training and validation curves of accuracy reached 70% and 68% respectively. Moreover, despite the complexities in processing Roman Urdu (as explained earlier) the difference between the training and validation curves in Fig. 3d is less than 2%, which is well in line for a models to be considered as a good fitted model. This upholds the validation of model on normalized dataset.

Fig. 4 illustrate the result summary of language models on both datasets. Accuracy on the normalized dataset is higher for each language model as compare to the preprocessed dataset. From figure 4, it must be pointed out that results are getting increasingly better from CNN to SA-BiLSTM on the normalized dataset. Experimental results prove that every model performs better on normalized dataset (which is 3x times smaller) than preprocessed dataset and utilizes less time cost. SA-BiLSTM results in the highest



(a) Preprocessed dataset

(b) Normalized dataset

(c) Preprocessed dataset

(d) Normalized dataset

Fig. 3. Confusion matrix and Accuracy curves of SA-BiLSTM

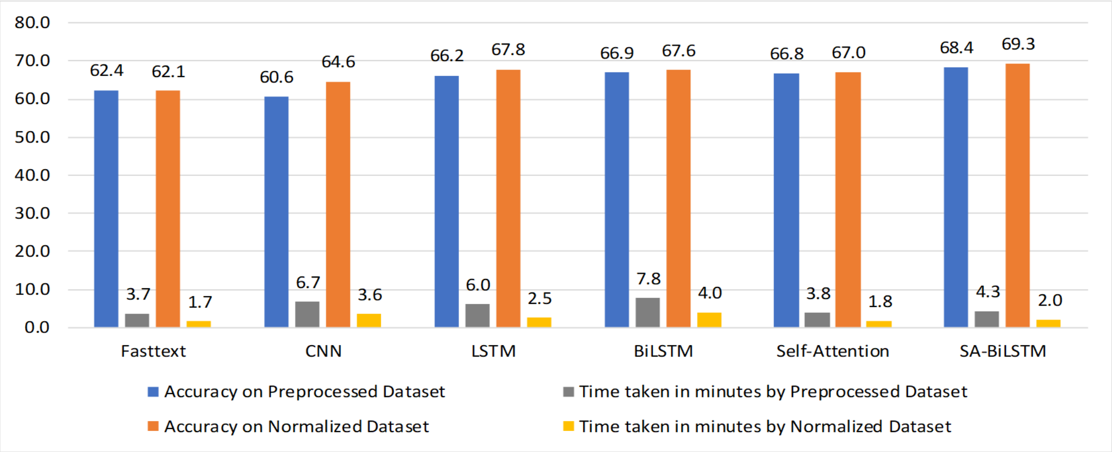


Fig. 4. Comparison of Accuracy achieved by all language models w.r.t time (minutes) on both dataset

accuracy because of its selective and bidirectional nature on both datasets which confirm that this integrated model is the best choice for sentiment analysis of Roman Urdu. Additionally, Normalization of the dataset is important for all non-English languages as it improves the performance of the model.

VI. Conclusion

In this paper, we present a novel deep learning model for sentiment analysis of Roman Urdu. This particular script hinders direct approaches owing to its complex sentence structure, and numerous lexical meaning. Proposed model utilizes the traits of Self-attention and Bidirectional LSTM (SA-BiLSTM) network to yields better results. Moreover, to make a fair comparison, we preprocessed and normalized the dataset. Experimental results indicate that SA-BiLSTM surpasses existing deep learning models in accuracy and requires fewer resources. SA-BiLSTM achieves a high accuracy of 68.4% and 69.3% for preprocessed and normalized datasets, respectively.
As for future research, we can try to enhance the efficiency of SA-BiLSTM and bring it to use for language inference and generation tasks, and these are very critical components to increase normalized vocabulary, vast pre-trained embedding, and massive datasets for better analysis.

References

[1] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.

[2] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[3] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[4] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[7] N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," *arXiv preprint arXiv:1802.05751*, 2018.

[8] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, and N. Shazeer, "Generating wikipedia by summarizing long sequences," *arXiv preprint arXiv:1801.10198*, 2018.

[9] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "Disan: Directional self-attention network for rnn/cnn-free language understanding," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—a review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16–32, 2018.

[11] K. Ravi and V. Ravi, "Sentiment classification of hinglish text," in *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*. IEEE, 2016, pp. 641–645.

[12] H. Kaur, V. Mangat, and N. Krail, “Dictionary based sentiment analysis of hinglish text,” *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, 2017.

[13] X. Ouyang, P. Zhou, C. H. Li, and L. Liu, “Sentiment analysis using convolutional neural network,” in *2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing*. IEEE, 2015, pp. 2359–2364.

[14] C. Zhou, C. Sun, Z. Liu, and F. Lau, “A c-lstm neural network for text classification,” *arXiv preprint arXiv:1511.08630*, 2015.

[15] P. Wang, J. Xu, B. Xu, C. Liu, H. Zhang, F. Wang, and H. Hao, “Semantic clustering and convolutional neural network for short text categorization,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2015, pp. 352–357.

[16] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, “Deep unordered composition rivals syntactic methods for text classification,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. 1, 2015, pp. 1681–1691.

[17] Y. Goldberg, “A primer on neural network models for natural language processing,” *Journal of Artificial Intelligence Research*, vol. 57, pp. 345–420, 2016.

[18] J. Islam and Y. Zhang, “Visual sentiment analysis for social images using transfer learning approach,” in *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom)*. IEEE, 2016, pp. 124–130.

[19] L. Yanmei and C. Yuda, “Research on chinese micro-blog sentiment analysis based on deep learning,” in *2015 8th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 1. IEEE, 2015, pp. 358–361.

[20] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[21] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[22] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[24] T. Chen, R. Xu, Y. He, and X. Wang, “Improving sentiment analysis via sentence type classification using bilstm-crf and cnn,” *Expert Systems with Applications*, vol. 72, pp. 221–230, 2017.

[25] R. C. Fong and A. Vedaldi, “Interpretable explanations of black boxes by meaningful perturbation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3429–3437.

[26] W. Zaremba, I. Sutskever, and O. Vinyals, “Recurrent neural network regularization,” *arXiv preprint arXiv:1409.2329*, 2014.

[27] Z. Sharf and S. U. Rahman, “Lexical normalization of roman urdu text,” *International Journal of Computer Science and Network Security*, vol. 17, no. 12, pp. 213–221, 2017.

[28] A. Rafae, A. Qayyum, M. Moeenuddin, A. Karim, H. Sajjad, and F. Kamiran, “An unsupervised method for discovering lexical variations in roman urdu informal text,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 823–828.

[29] T. Ahmed, “Roman to urdu transliteration using wordlist,” in *Proceedings of the Conference on Language and Technology*, vol. 305, 2009, p. 309.

[30] R. Satapathy, C. Guerreiro, I. Chaturvedi, and E. Cambria, “Phonetic-based microtext normalization for twitter sentiment analysis,” in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2017, pp. 407–413.

Knowledge based Authentication Techniques and Challenges

Hosam Alhakami[1], Shouq Alhrbi[2]
School of Computer Science and Information Systems
Umm Al-Qura University
Saudi Arabia, Makkah

***Abstract*—Knowledge-based Authentication (KBA) is an authentication approach, which verifying the user identity when accessing services such as finical websites. KBA requests specific information to prove personal identity of the owner. This paper discusses the challenges that are faced by KBA techniques. Memorability is the main obstacle in KBA since the users trying to utilize simple passwords or unify the passwords in various services, a step that cause problems and issues with compliance with security policies. Furthermore, the technique of mixing username/password is considered as another important challenge of KBA due to the recall-based authentication. This discussion includes a comparative analysis of KBA's techniques based on trade-off criteria to support making of decision. This study's results can support organizations in the recommendations process of a suitable KBA technique for organizations.**

***Keywords*—Knowledge-based authentication; artifact-based authentication; biometric-based authentication; usability; vulnerabilities; memorability; performance; cost**

I. Introduction

Authorization [1, 2] is the process of ensuring only authorized rights are exercised in the process of determining rights. Authentication is verifying the person's identity, such as (a user, or device) who intends to access data, resources, or applications. Confirming the identity of an entity proves a confidence relationship for interactions. Authentication [3] also allows accountability based on the possibility of mapping the access link and concurrent actions to identities. The techniques of authentication are classified into three essential categories which are token-based authentication, biometric based authentication [4] and knowledge-based authentication system [5]. Fig. 1 illustrates the types of user authentication types [6] but that differs in the focusing idea based on in each type.

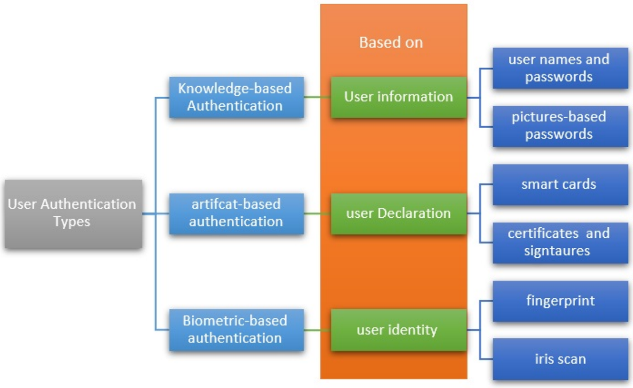

User Authentication Types
Knowledge-based Authentication
artifcat-based authentication
Biometric-based authentication
Based on
User information
user Declaration
user identity
user names and passwords
pictures-based passwords
smart cards
certificates and signtaures
fingerprint
iris scan

Fig. 1. Types of User Authentication.

Previous researches discuss the different identification and authentication techniques and their different key terms which include protect credentials, identity, password, biometrics, and others [6]. Any system requires to identify its users and authenticate them accordingly depending on the system's target and the target population. User authentication is of three types: knowledge-based, artifact-based, and biometric-based. Any system that relies on the secret user identity information such as text or image passwords that the user provided in the registrations process or when creating passwords is said to be dependent on knowledge-based authentication for its users authentication [7]. Any system that relies on authentication signature or smart issues is said to be dependent on artifact-based authentication for user authentication. Furthermore, any system that relies on the physical characteristics of the user such as fingerprints in the authentication process is considered to depend on biometric based authentication for authentication of its users.

This research targets studying KBA, and specifically emphasizing on security and usability challenges [8]. KBA is an authentication approach that searching the evidence to define of accessing a service. This study discusses different types of KBA and the requirements for each type of KBA. Authentication is necessary in this era of big data revolution on the internet that has affected the mode of human communication and the quality of services provided which all depends on sharing the information. KBA is a popular technique that is used by the largest population of IT systems users but it faces several challenges in this technique.

KBA is known for its simplicity, ease of revocation and legacy deployment that consists of textual and graphical password. Previous studies [9, 10] unearth several attacks that enabled attackers stealing user's identity and confidential information. KBA is defined by an authentication approach that looking for the evidence to define of accessing a service. Static KBA and Dynamic KBA are the main two types [11, 12] of KBA. Fig. 2 discusses these types [11] which figure includes the main feathurs and examples for each type. Static KBA refers to a pre-agreed set of shared secrets like passwords [13]. Dynamic KBA refers to questions generated from a wider based of personal information like registration or verification questions.

In Addition, the static KBA refers to the process that enable users to choose security questions and provide answers that are

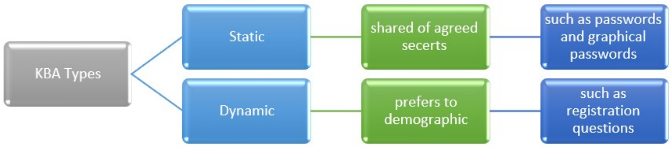

KBA Types
Static
Dynamic
shared of agreed secerts
prefers to demographic
such as passwords and graphical passwords
such as registration questions

Fig. 2. The Types of Knowledge-Based Authentication.

stored by an organization to be accessed later. Moreover, the dynamic KBA refers to go a step that generate questions that applies only to the intended end user and do not require a previous relationship with the customer. The most used technique for authentication is username and password which is classified into one of the knowledge-based techniques [14]. The essential cause of utilizing password as a popular technique is that it does not require any special target hardware to observe in and out operations on protected areas in the systems [15]. According to the literature, KBA was identified as the approach used to combine some challenges (i.e. questions) to verify claimed users where the answers of these challenges came from their knowledge [16].

This study discuses the definition, importance, types, techniques, and challenges of KBA. Also, it explores KBA techniques which are usability, memorability, performance and cost and any combination of the stated KBA techniques. This paper includes a comprehensive review in term of comparative analysis which will be taken into consideration to provide tradeoffs criteria to help decision makers in their organization so that they can be able to select the most suitable KBA technique. This study mentions recent research trends in this domain.

This paper is organized as the following: Section 2, examining the related works of knowledge-based authentication and its security issues, Section 3, Discussion, Section 4, presents open research challenges of knowledge-based authentication. Finally, Section 5 discusses the conclusion and future works.

II. Related Works

The main goal [17] of a user authentication mechanism is offer security to information systems. Attackers are using several strategies to attack authentication systems that are in use in different systems. Therefore, schemes must be measured with respect to vulnerabilities and susceptibility to various attacks which can indicate absence of enough security for any system that uses that specific scheme. Use of passwords and user-identity in processing of login is one of the most popular scheme. Knowledge-based authentication (KBA) mechanisms utilize the memorized authentication secret that can be a text password (as numbers and characters), a personal identification number (PIN) or a graphical/image password such as CAPTCHA. The benefits of using traditional passwords is that there is no specialized personnel, hardware or software required, simple to use, and easy to remember. But that causes of many problems of using the password, that it is more likely to suspicious attacks and speculation of passwords.

A. Knowledge-based Authentication (KBA)

The evaluation of Knowledge-based authentication (KBA) is best when based on the following criteria that includes static and dynamic type (which illustrated in Fig. 2). A suitable security question should be acceptable to the largest segment population, possess answers that are easy to remember, have no redundant answers for the correct answer, and the answer of security question should not be simple to speculation to find out in searching. For any technique, it is based on KBA that requires to depend on four dimensions [18], which are known as KBA techniques (which illustrated in Fig. 3), memorability, usability, performance, and cost. Previous researches focus on the memorability and usability [16, 17] that differs the Fig. 3 includes other collective techniques performance and cost. There are many research and applications that recommend combinations of these techniques to reach the good level of KBA. Memorability refers to the saving passwords is the browser. Usability is meaning that the uses of passwords in several applications that is vulnerable to attack easily. Performance refers to the strength of password. Cost targets reducing fraud from fraudulent claims.

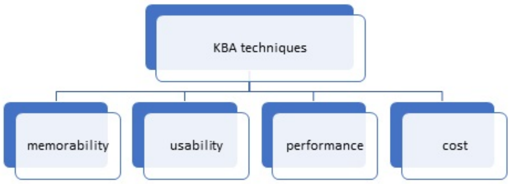


Fig. 3. The KBA Techniques.

Table I illustrates a comparative study between several motivation researches in authentication based on authentication type (which mentioned in Fig. 1), techniques, and the authentication mechanism. This comparison also mentions the advantages and disadvantages of these authentication mechanisms and techniques.

B. Knowledge-based Authentication Challenges

Knowledge-based authentication is the main target of study in this research. Fig. 4 mentions the challenges of KBA, it discusses main problems, security and usability as [8], but it includes the Characteristics of them that are divided into six Characteristics challenges. Table II discusses the comparison between the different types of KBA and their techniques. This research focus on Knowledge-based authentication with passwords credentials and properties. The Knowledge-based authentication has several techniques and challenges which are shown in the comparison Table II.

The two main challenges to KBA is Usability and Security. Each type of KBA has several challenges as the following: usability challenges includes usability in several applications, management problems, and the domino effect. The security challenge includes security issues, searchable personal data, and privacy. Mostly, attacks are the most feared challenge in all the mentioned challenges of KBA. The challenges are discussed in the following:

1) Security challenge

The main challenge of the KBA is how to be safe from attacks and hacks. The required challenge is how to

TABLE I. THE KNOWLEDGE-BASED AUTHENTICATION TECHNIQUES FOR SECURITY ASPECTS

No.	Authentication mechanism	Authentication Type	Technique	Pros	Cons
[19]	password	Knowledge	It includes sign in and sign up for users. It uses mathematical analysis and combination between years and visits number of passwords	prevent multiple brute-force attack	Has a long time
[20]	Finger print	Biometric	It is based on a comparison between four authentication schemes: facial recognition, finger print, pin code, and NFC ring	Higher accuracy And finger scan is easy to use and trus	Can not guarantee the trust in this technique for using in the public setting; Hardness to use facial recognition
[21]	password	Knowledge	efficient password protocols	Ensure the security of data that can be safer from attacks	Hardness security and usability
[22]	Graphical password	Knowledge	A proposed technique is entitled WYSWYE (where you see is what you enter) strategy.	Decrease guessing attacks	Improving accuracy with user's images
[23]	Voice Recognition	Biometric	Using text-to-speech technique and speech-to-text technique	Easy to use and minimize cost and memory	Improve efficiency
[24]	Multi-Factor based authentication(voice, text, iris, DNA, ...)	Biometric	It is based on Multi-Factor Authentication	Improve security level and prevent attacks	Hardness to apply it
[25]	Graphical passwords	Knowledge	The technqiue is based on Pictorial password systems	High secure and usability	Difficulty to upload personal images
[26]	graphical random authentication technique (gRAT)	Knowledge	The technique is based on the classification of the existing graphical password methods into recognition-based, cued-recall-based, pure-recall-based, and hybrid techniques.	More secure and powerful in usability	Complex implementation
[27]	Textual password	Knowledge	Studying twelve passwords schemes	Improving decision making and usability when combining the most used password the is text with fingerprint	That is very powerful for smartphone only
[28]	Dual-Factor Authentication Protocol	Knowledge	It does not employ a password verifies. s dynamically changed each time the user logs in	Increasing security with multi-factors from inside and outside attacks	Time- and energy-consuming

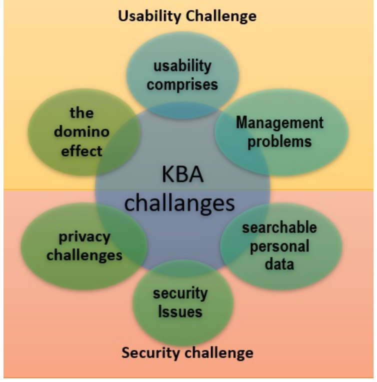


Fig. 4. KBA Challenges.

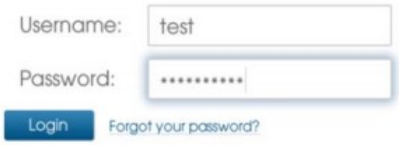

Username: test
Password: ••••••••••

Fig. 5. Example of username and password profile.

save personal information, as the username and password example are shown in Fig. 5, in various domains.

Previous researches discuss the state-of-the-art of knowledge-based user authentication mechanisms that are classified two dimensions: security and usability. Security authentication mechanisms discuss and compare the strength of each mechanism depends on various policies. The major discussion of this analysis and identify areas for further research and enhanced methodology with the target to drive this research towards the design of sustainable, secure and usable authentication approaches. This challenge divides into three parts: security issues of attack types, searching about the people's information or identity, and privacy challenges of the user accounts. Security challenge is divided into three types as the following:

a. Searchable personal

The use of passwords that are the same in different social networks simplifies things for users but that is considered a challenge because of the repetition and circulation of the password. That may be caused by easy attacks or guessing the passwords.

b. Security Issues

There are several types of attacks and hacks with fake account or stealing data. Hackers can steal personal data and accounts and sell these data to benefit from the information. The main challenge of security is guessing the account's passwords. There multiple online and offline password guessing techniques that are in use. The famous method to prevent guessing while online is inclusion of CAPTCHA in systems. Offline method does not need computational power, but it is based on several times of guessing passwords and writing them in a repetitive way.

c. Privacy challenges

Privacy is implemented using privacy laws that protects client privacy and aim at controlling access to client's data. So, there is always a need to make verification questions that are not private to users and not discriminate for specific users to avoid attacks.

2) Usability challenge

Usability is considered a critical challenge of managing user's accounts due to the ease of use of the same password in several domains and applications. But it is a threat that threatens the safety and confidentiality of data. It includes management problems of various systems, the problem effect on domains, and usability challenges which can be interpreted in the repeatedly used passwords.

a. Management problems

Management has several problems such as the organization authentication of several users who want to access the system due to the similarity of passwords and registration questions. There are several conditions for suitable questions as the following. They do not include default values, texts, and the organizations have quick recovery techniques for any sudden attacks.

b. Usability compromises

The ability of usability challenges provides to the user some capability. Graphical/audio challenges can be employed. Using the same password in several platforms becomes risk of user accounts. Users are threatened by attackers via guessing accounts users and passwords without the user's knowledge. These guessing of passwords have several policies to minimize the challenge of passwords memorability.

c. The domino effects

The accumulative impact introduces a group of similar events. The idiom is best known as a mechanical impact and is utilized as an analogy to a falling row of dominoes.

C. Knowledge-based Authentication Security Measurements

From previous researches, we found that it is very important to find a way to evaluate authentications for various platform's policies [9,10, 29]. The evaluation criteria are built based on a combination of three parts: password intensity, guessability percentage rate, and entropy measurement.

a. Password intensity: It refers to the strength of using characters, numbers, and the length of the password. The password should not be related to the name or email. A password's intensity can be in one of these types (weak, medium, strong, and very strong).

b. Guessability percentage rate refers to the numbers of speculations from attackers or hackers to guess the user's password. This rate depends on the password's parameters of guassability that is used for improving the password intensity and saving data.

c. Entropy measurement: It is defined by one of the security measurements for each policy. Entropy refers to the random number of ways that users can choose the passwords from given keys that are related to the hardness rate of guessing the textual passwords.

Table II discusses a comparative study of knowledge-based authentication challenges. It reviews the strengths and weaknesses of each technique and suitability in different applications.

TABLE II. A COMPARATIVE STUDY OF KBA CHALLENGES

Application scheme	**Password Length (minimum # of characters)**	**Guessability**	**Authentication advantages**
Google	8	Yes, that is easy to guess google passwords	Weak and short. Powerful and password Strong.
Yahoo	9	Yes, that is simple to suggest yahoo passwords	Can't enough authentication security.
Facebook	6	No, that is not simple enough that has complex rules for prevent many attacks types	It includes three categories: Weak, average or password Strong
Twitter	6	No, that is not easy to attack due to the complexity of passwords authentications It requires very short and clear.	That is classified into four classes Weak, Good, Strong, Very Strong
Instagram	6	No, that does not put complex rules for authentication security and save user's accounts	It has not enough security roles for passwords and length
Amazon	6	No, that is not suitable secure for this system and its users	It has not enough security roles for passwords and length
Booking	8	No, that is not enough secure the used passwords	It has not enough security roles for passwords and length
Linked in	6	No, that does not take care of the importance of authentication passwords	It is a medium password challenge and small length
Ebay	6	No, that has not authentication rules enough for secure system	It has not enough security roles for passwords and length
Dropbox	6	Yes	That is classified into four classes Weak, medium, Good, Great

III. DISCUSSION

The evaluation of Knowledge-based authentication (KBA) technique is satisfied when criteria for static and dynamic KBA are achieved. This criteria consists of:

A. Static: Any system requires strong password (fixed length such as from 6 to 8 characters). It has suitable number of characters and the password must include special characters and alphabetic letters. It also needs to minimize the complexity to make the passwords and authentication profiles and questions are easy to remember.

B. Dynamic: Any system requires to be dynamic to create suitable security question that are related to the large segments of the population. The answers to the question should be such that, they make it easy for users to remember them easily. But each question requires to have unique answer. This means that there should be no redundant answers for correct answer. The answer of security question should not be simple to speculation to find out in searching.

Since the main goal of user authentication mechanism is to improve the security of the information systems, several strategies are applied by the attackers to compromise the authentication to the system. Passwords have many challenges which include their high susceptibility to exposure to attacks, password guessing, and key-loggers. KBA includes techniques: memorability, usability, performance, and cost, and combinations of any of these techniques. Most of the challenges of implementing KBA techniques are in online services. Also, analysing and testing the strength are essential in comparing different KBA techniques. The comparison will focus on usability, memorability, security, and performance. The research will study cases of combining different KBA techniques, and the resulting framework, its strengths, weaknesses, and applications. Previous researches conclude that the importance of security challenge is bigger than usability. Several applications require to improve their security systems and authentication rules to protect users and to prevent attacks. This improvement might be necessary depending on the KBA security measures.

IV. Open Research Trends

This research can support researchers and students to make several motivations in this area to improve the performance of their security systems. First, they can work on solving the knowledge-based authentication challenges. In the Memorability challenge, the research can improve the memorability to make easy and simple to use passwords but while still adhering to the restrict rules. Use of the same password in several platforms should never be allowed. In the usability challenge, open research provides important information on how to make passwords and authentication for users based on KBA security measurements [30]. For the security issues challenges, open research goes forward to give information on how to prevent attacks and hacks. Second, dynamic KBA is very difficult to implement and is considered harder than Static KBA. Finally, there is no standard reusable model available for dynamic KBA that fits the need of all the organizations.

V. Conclusion and Future Works

This paper introduces the authentication survey and makes comparison of the different types of authentication mechanisms. It discusses the importance of knowledge-based authentication (KBA) from a security perspective. It also examines the challenges of knowledge-based authentication challenges and open more research areas. This survey concludes that there is a good criterion for knowledge-based authentication based on a textual methodology based on the types of KBA whether static or dynamic. Textual KBA is the most usable method although several platforms and studies suggest using an image or graphical authentication mechanisms. Textual KBA faces many challenges to be secure and safe from attackers and hackers. KBA includes four techniques as the following: memorability, usability, performance, and cost, and combinations of any of those techniques. The major challenges when it comes to implementing KBA techniques lies in online services. Also, the strength and analysis will be essential in comparing the different KBA techniques.

Acknowledgment

The research for this paper was financially supported by Umm Al-Qura University. We gratefully acknowledge the support and generosity of the University, without which the this study could not have been completed.

References

[1] Manjunath D, Nagesh A S,Sathyajeeth M P, Naveen Kumar J R, and Syed Akram, A Survey on Knowledge-Based Authentication, Volume 2, Issue 4, 2015.

[2] Mohammad A Alia, Adnan Hnaif, Ayman M. Abdalla, and Mohammad Abu Maria, An improved authentication scheme based on graphical passwords, ICIC Express Letters, Volume 12 (8), pp.775-783, 2018

[3] Alican Beydemir; İbrahim Soğukpinar, Lightweight zero knowledge authentication for Internet of things,International Conference on Computer Science and Engineering (UBMK), 2017.

[4] Hasini Gunasinghe and Elisa Bertino, PrivBioMTAuth: Privacy Preserving Biometrics-Based and User Centric Protocol for User Authentication From Mobile Phones, IEEE Transactions on Information Forensics and Security (Volume: 13) , Issue: 4, 2018.

[5] M. Yildririm, and I.Mackie, Encouraging users to improve password security and memorability,International Journal of Information Security,Volume 18, Issue 6, pp 741–759, 2019.

[6] Nurul Afnan Mahadi, Mohamad Afendee Mohamed, Amirul Ihsan Mohamad, Mokhairi Makhtar, Mohd Fadzil Abdul Kadir and Mustafa Mamat, A Survey of Machine Learning Techniques for Behavioral-Based Biometric User Authentication,Recent Advances in Cryptography and Network Security, 2018.

[7] Bhanushali, A., Mange, B., Vyas, H., Bhanushali, H. and Bhogle, P. (2015). " Comparison of Graphical Password Authentication Techniques". International Journal of Computer Applications (0975 – 8887) April 2015. Vol. 116, No. 1.

[8] Katsini, C., Belk, M., Fidas, C., Avouris, N. and Samaras, G., "Security and Usability in Knowledge-based User Authentication: A Review"., 2016.

[9] Jyoti Deogirikar and Amarsinh Vidhate, Security Attacks inIoT: A Survey, International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), 2017.

[10] Ramsha Fatima, Nadia Siddiqui, M. Sarosh Umar, and M. H. Khan, A Novel Text-Based User Authentication Scheme Using Pseudo-dynamic Password, Information and Communication Technology for Competitive Strategies pp 177-186, 2018.

[11] Yusuf Albayram, Mohammad Maifi Hasan Khan, Athanasios Bamis, Sotirios Kentros, Nhan Nguyen, and Ruhua Jiang, Designing challenge questions for location-based authentication systems: a real-life study, Human-centric Computing and Information Sciences, 2015.

[12] George Hadjidemetriou, Mario Belk, Christo Fidas, and Andreas Pitsillides,Picture Passwords in Mixed Reality: Implementation and Evaluation, CHI EA '19,CHI Conference on Human Factors in Computing Systems, pp. 1–6, 2019.

[13] Nawaf Aljohani, Joseph Shelton, Kaushik Roy, and Albert Esterline, Robust password system based on dynamic factors,6th International Conference on Information Communication and Management (ICICM), 2016.

[14] Alsuhibany, S. (2016). “Evaluating the Usability of Optimizing Text-based CAPTCHA Generation”. International Journal of Advanced Computer Science and Applications (IJACSA) 2016, Vol. 7, No. 8.

[15] Alexander Popov, Neural Network Models for Word Sense Disambiguation: An Overview,BULGARIAN ACADEMY OF SCIENCES,CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume 18, No 1 Sofia , 2018.

[16] Abrar Ullah, Hannan Xiao, and Trevor Barker, A study into the usability and security implications of text and image-based challenge questions in the context of online examination, Education and Information Technologies January 2019, Volume 24, Issue 1, pp 13–39.

[17] Harakannanavar, Sunil Swamilingappa; Renukamurthy, Prashanth Chikkanayakanahalli; and Raja, Kori Basava, ,Comprehensive Study of Biometric Authentication Systems, Challenges and Future Trends,International Journal of Advanced Networking and Applications Vol. 10, Iss. 4.

[18] Muhammad Sharif, Mudassar Raza, Jamal Hussain Shah, Mussarat Yasmin, and Steven Lawrence Fernandes, An overview of biometrics methods,,Handbook of Multimedia Information Security: Techniques and Applications pp 15-35, 2019.

[19] Amirul I Mohamad, Mohamad A Mohamed, Mokhairi Makhtar, Mustafa Mamat, Norziana Jamil,and Marina Md Din, A Framework for Experience Based User Authentication Technique for Minimizing Risk of Brute-Force Attacks, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-5S4, February 2019.

[20] Matthias Baldauf, Sebastian Steiner,Mohamed Khamis, and Sarah-Kristin Thiel, Investigating the User Experience of Smartphone Authentication Schemes -The Role of the Mobile Context,Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019.

[21] Kaur, Amanpreet A; Mustafa, Khurram K., A Critical appraisal on Password based Authentication, International Journal of Computer Network and Information Security; Vol. 11, Iss. 1, 2019.

[22] Yogesh V. Mahajan, Ganesh R. Tile, and Paresh S. Patil, Cued Click Point Graphical Authentication,National Level Conference On ”Advanced Computing and Data Processing”(ACDP 2K19).

[23] Kapile Namrata Rajesh, Bhanushali Nayan Valji, Pawase Kalpesh Dattatray, Pawar Shubham Gangaram, and Prachi S. Tambe, Voice Assistant for visually impaired person,National Level Conference On ”Advanced Computing and Data Processing”(ACDP 2K19),Vidyawarta Research Journal, 2019.

[24] Oladimeji Biodun S, Prof. Gloria Chukwudebe ,Dr. A.O Agbakwuru, and Osodeke Charles Efe, Comparative Study of Multi-Factor Authentication Systems, intternational Journal of AdvancedResearch in Science, Engineering and Technology Vol. 6, Issue 4, April 2019.

[25] Mr.Devidas S. Thosar, Mr.Narayan B. Vikhe, Ms.Rajashree R. Shinde, Ms.Prachi S.Tambe, and Ms.Priyanka S. Hase, ClickPoints:An Advanced Graphical Authentication Using Image Descrimination & Fusion,National Level Conference On ”Advanced Computing and Data Processing”(ACDP 2K19), 2019.

[26] Mudassar Ali Khan, et al., g-RAT | A Novel Graphical Randomized Authentication Technique for Consumer Smart Devices, IEEE Transactions on Consumer Electronics , Volume: 65 , Issue: 2 , May 2019.

[27] Verena Zimmermann, and Nina Gerber, The password is dead, long live the password – A laboratory study on user perceptions of authentication schemes,International Journal of Human-Computer Studies Volume 133, January 2020, Pages 26-44.

[28] Abdelrahman Abuarqoub, D-FAP: Dual-Factor Authentication Protocol for Mobile Cloud Connected Devices, journal of sensor and actuator networks, volume (9), issue (1), 2019.

[29] Abrar Ullah, Hannan Xiao, and Trevor Barker,A Dynamic Profile Questions Approach to Mitigate Impersonation in Online Examinations,Journal of Grid Computing, Volume 17, Issue 2, pp 209–223, 2019.

[30] Amanpreet Kaur and K. Mustafa, Qualitative assessment of authentication measures,3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016.

Map Reduce based REmoving Dependency on K and Initial Centroid Selection MR-REDIC Algorithm for clustering of Mixed Data

Khyati R. Nirmal[1], K.V.V Satyanarayana[2]
Department of Computer Science and Engineering,
K L Education Foundation,
Vaddeswaram -522502,
Guntur Dist., A.P, India

***Abstract*—In machine learning, clustering is recognized as widely used task to find hidden structure of data. While handling the massive amount of data, the traditional clustering algorithm degrades in performance due to size and mixed type of attributes. The Removal Dependency on K and Initial Centroid Selection (REDIC) algorithm is designed to handle mixed data with frequency based dissimilarity measurement for categorical attributes. The selection of initial centroids and prior decision for number of cluster improves the efficiency of REDIC algorithm. To deal with the large scale data, the REDIC algorithm is migrated to Map Reduce paradigm,and Map Reduce based REDIC(MR-REDIC) algorithm is proposed. The large amount of data is divided into small chunks and parallel approach is used to reduce the execution time of algorithm.The proposed algorithm inherits the feature of REDIC algorithm to cluster the data.The algorithm is implemented in Hadoop environment with three different configuration and evaluated using five bench mark data sets. Experimental results show that the Speed up value of data is gradually shifting towards linear by increasing number of data nodes from one to four. The algorithm also achieves the near to closer value for Scale up parameter, while maintaining the accuracy of algorithm.**

Keywords*—*Machine learning; clustering; similarity measurement; initial centroid selection; number of clusters; map reduce paradigm

I. Introduction

In the current era of machine learning the strategy of unsupervised algorithm to group data without prior knowledge is widely adopted in the application of data segregation. Like the applications in the field of web mining, text mining, image processing, stock prediction, signal processing, biology and other fields of science and engineering the algorithms of clustering had been applied for better results. [1] [2]

Clustering targets to find out hidden structure from underlying dataset, without any prior information, here the labels are not associated with data. As an outcome of this the clusters are formed having minimum within cluster distance and maximum between cluster distances. Here one representative of a group is chosen as cluster centroid.

A number of strategies have been proposed in last several years, for solving the clustering problems efficiently[3] The two broad categories of clustering algorithm are: Hierarchical Clustering and Partitional Clustering. In Partitional Clustering K Means Clustering has practiced the facts of simple mathematical formation, ease of implementation and fast coverage[4]. This will enlarge the application field of the algorithm.

Conversely the results generated of K Means Clustering coverage the local optimal based on initial clusters which may or may not be endure from global optimum solution. The Hamming Distance measurement is applicable for numerical dataset only, however the attributes of real world data is not restricted to numerical attributes, it consist of the categorical attributes as well. This is an additional obstacle to adopt the K Means Clustering Algorithm.

For categorical attributes the K Mode Clustering Algorithm has supplanted the Hamming distance measure with simple matching dissimilarity measurement and derived with the alternate solution. To extend the K Mode Clustering algorithm for mixed attributes the K Prototype Clustering algorithm is proposed, which is in cooperation with both the distance based measurement. [5] [6]

The K prototype Clustering algorithm is one of the approaches for clustering of mixed attributes. To enhance the performance of K prototype clustering algorithm, various provisions have been proposed in last decades. The emphasis of this paper is to augment one more approach in the same direction.

The Section 2, the REDIC K Prototype Clustering algorithm along with necessity of migrating the algorithm to MapReduce Paradigm is elaborated. The MapReduce REDIC K prototype Clustering algorithm is proposed In Section 3, the experimental setup and result of the proposed algorithm are given in Section 4.

II. Background Knowledge

In the Clustering of Mixed data using the K prototype clustering algorithm, the numerical attributes and categorical attributes are separated and functioned separately using two different similarity measurements. It implements Euclidean Distance measurement for numerical attributes and Hamming distance for categorical attributes. The Hamming distance measurement results in 0 if two categorical attributes are similar and results 1 if the results are dissimilar. This reasoning may not give the better result in many real world data set[7].

As the extension of K Prototype Clustering algorithm, The K center Algorithm has been proposed and proved that the algorithm contributes improved results by considering the frequency of attributes in consideration [8].

The effectiveness of this algorithm is contingent on the initial centroids selected, and on the other side the algorithm requires professional knowledge to choose the value for parameter K [9].

REDIC K Prototype Clustering algorithm is proposed in [10], which has precisely concerns the above stated issues, and suggested the alternate strategy.

A. REDIC K Prototype Clustering

The similarity between two categorical attributes is largely depends on the relative frequency of the common values for particular attributes.[11]. For this reason the frequency based method for similarity measurement of categorical attributes must be adopted.REmoval Dependency on K and Initial Centroid Selection (REDIC) K Prototype Clustering algorithm is proposed with a novel frequency based similarity measurement for categorical attributes. [12] Using simple furthest point heuristic (Maxmin) initialization decreases the clustering error of k-means from 15% to 6% on average [13]. This strategy is adopted in REDIC algorithm while choosing the initial centroids. (REDIC) K prototype clustering is proposed, which will have three contribution: 1) Frequency based dissimilarity measurement for the categorical attributes. 2) Select the initial centroids by calculating most significant attributes. 3) Incremental approach for deciding number of clusters. The preliminary used in the algorithm are defined as:

Preliminary 1. Similarity between two Categorical Attributes $Cdist(c_i, c_j)$
Consider any two categorical instances C_i and C_j of n instances.

$$Cdist(c_i, c_j) = FOC(c_i) - FOC(c_j) \tag{1}$$

where $FOC(c_i)$ is defined as

$$FOC(c_i) = \frac{no\ of\ occurrence\ of\ c_i}{total\ no\ of\ instances\ n} \tag{2}$$

Preliminary 2. Similarity between two Instances $Dist(I_i, I_j)$

$$\begin{aligned} &Dist(I_i,\ I_j) =\ NDist(n_i,\ n_j) + CDist(c_i,\ c_j) \\ &where \\ &NDist =\ Dis\tan ce\ of\ Numerical\ Attributes\ (n_i,\ n_j) \\ &CDist =\ Dis\tan ce\ of\ Categorical\ Attributes\ (c_i,\ c_j) \end{aligned} \tag{3}$$

Preliminary 3. Initial Centroid Selection:
The instances having minimum or maximum row factor will be consider as initial cenroids. Here out of n attributes, 1 to m are numerical attributes and m to n are categorical attributes.

$$Row\_Fact(i) = \sum_{a=1}^{m} num_i\ +\ \sum_{a=m}^{n} FOC(i) \tag{4}$$

$$Set\ of\ Initial\ Centroids\ CN =\ \{CN_1, CN_2,\ \dots\ ,\ CN_k\} \tag{5}$$

where k = number of instances having minimum or maximum value for Row Factor

Preliminary 4. Initial Value for Number of Cluster(k): The cordiality of set CN is consider as number of initial centroids.

$$k = |CN| \tag{6}$$

Preliminary 5. Decision parameter for Cluster Refinement:
Consider the Cluster C_i having cluster centroid CN_i and instances are $I_1, I_2, ..., I_q$

$$\delta_i = min(dist(I_i, C_1), ..., dist(I_i, C_k)) \tag{7}$$

Using these five preliminaries the clusters are refined and formed, here it is observed that the algorithm is computationally simple not much expensive. The evaluation results are also better than the K prototype algorithm.

REDIC Algorithm is designed for the dataset of small size, to deal with the large dataset the substitute option should be formulated. Here the Map Reduce Paradigm is preferred to speed up the REDIC algorithm.

B. Map Reduce Paradigm

To process the large scale data the Map Reduce is designed and processed. [14] Automatic parallelization and task assignment reduce the overhead while deploying the algorithm to the paradigm. Here only two phases are essential to parallelize the algorithm namely: Map and Reduce. For both of the phases the input and output is in the form of $< key, value >$ pair. Match phase accepts the input in $< key, value >$ pair and produces the intermediate list in $< key, value >$ format only. By grouping and shuffling operation this list will be rearranged as per the intermediate key value. The Intermediate list is in the form of $< key, (value1, value2, \dots, valueN) >$. Reducer accepts intermediate list as an input and produce the final values according to the algorithm. In the Figure 1, the functions of Map and Red are explained with block of data. The library of MapReduce Paradigm splits the input file in number of blocks. The copies of the input files are stored to the nodes of the paradigm. One superior copy of input files with data and metadata is maintained and referred as the master. The master node does the task allotment automatically without user interference. The rest of nodes are the slave nodes, which performs the task assigned by master node. A slave node who is assigned a map work reads the input file and convert the file into $< key, Value >$ pair. The Mapper function performs the grouping and shuffling and creates the intermediate list. The reducer function reads the intermediate list and performs the sorting operation for mapping of different task to appropriate reducer task. A slave node who is assigned a reducer work iterates over sorted data and assigns the intermediate key to appropriate output values. After successful implementation of algorithm the output is stored into different files of reducer.

III. Proposed Algorithm: Map Reduce based REDIC K Prototype Clustering

To handle the categorical data of large scale, the Map Reduce based REDIC K Prototype Clustering algorithm is proposed. In MR REDIC Algorithm the distance between two categorical instances are calculated using frequency based method and the initial centers are selected by calculating the row factor of each instance. This will eliminate user

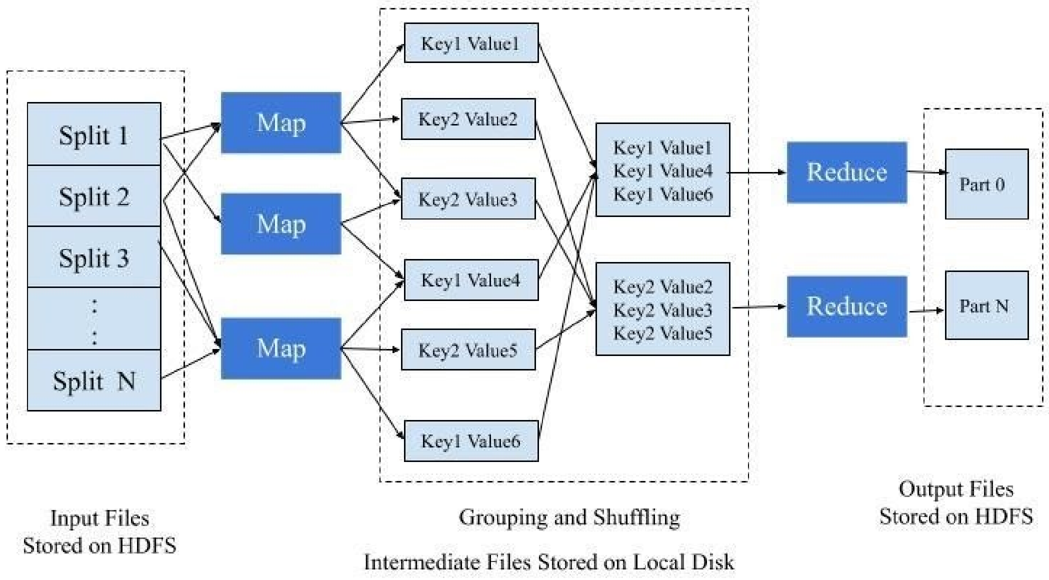


Fig. 1. Map Reduce Paradigm

dependency to choose the number of cluster priory and also improve the result in comparison with random centroid selection method. As shown in the figure 2, MR REDIC algorithm works in two Phases:

The map phase accepts the part of mixed dataset which is stored on HDFS as an input and calculates the row factor for each instance. The instance having minimum or maximum value for row factor is selected as initial centroids. This decides the number of clusters. The distance between each centroid to each instance is calculated and the instance is assigned to the cluster which have minimum distance . Here Cluster Index work as a key and the associated instances of that cluster work as a value of that key.

In reducer phase cluster refinement is done. Here the delta value is calculated for each cluster and it is compared with the distance between cluster centroid and each instance , If this value is less then only particular instance will remain in that cluster else it will be considered non promising instance. Such instances are eliminated from cluster and new cluster with that centroid will be created. This process will be continued till all the instances will be allotted to appropriate cluster and there will be no revision in cluster refinement. Here again Cluster Index work as a key and the associated instances of that cluster work as a value of that key.

The order to migrate REDIC algorithm to Map Reduce paradigm, MR REDIC algorithm is proposed. The Job class does the initialization of the job along with the directory path for Input and Output . The input dataset is stored on HDFS, which will be split and assigned to Mapper Class for further processing. The Job class of MR-REDIC constructs a global variant centers which is a null array(C) , later it will store the information about centers of the clusters. The task is distributed to the datanodes by job class using inbuilt libraries, and reducer class will refine the clusters and update the value for number of cluster parameter.

The Global Center array, offset key and sample value is assigned as an input to Mapper Class. Initially when this array is empty the Row value ϱ is calculated by using the method proposed in [12] and stored in the set R. The smallest and largest value from the R is extracted in ϱ_{min} and ϱ_{max}. The initial value for number of cluster k is set to the count of min and max value. The initial centroids are stored in CN_i variable, in next step distance from every instance to every centroid $dist(I_i, C_k)$ is calculated by using formula defined in [12] . The instance is assigned to the cluster having minimum cluster centroid distance. The intermediate $< Key, Value >$ pair is maintained where Key is the Cluster Index and Value is the value all the attributes of of particular instance.

Algorithm 1: Job Class

```
Input: I:Input path, O: Output path,W:Intermediate path
Output: a set of k Clusters with allotted instances,
        Execution Time
Method:
Input Path I;
Output Path O;
Create Job ;
Create Cluster Center Array C= $\emptyset$ ;
Set Mapper class;
Set Reducer Class ;
globalF = false ;
while globalF == True do
    Update Input path = W ;
    Start Job
end
data from W to O ;
return value of k and execution time
```

The Reducer class will refine the clusters by reallocating the

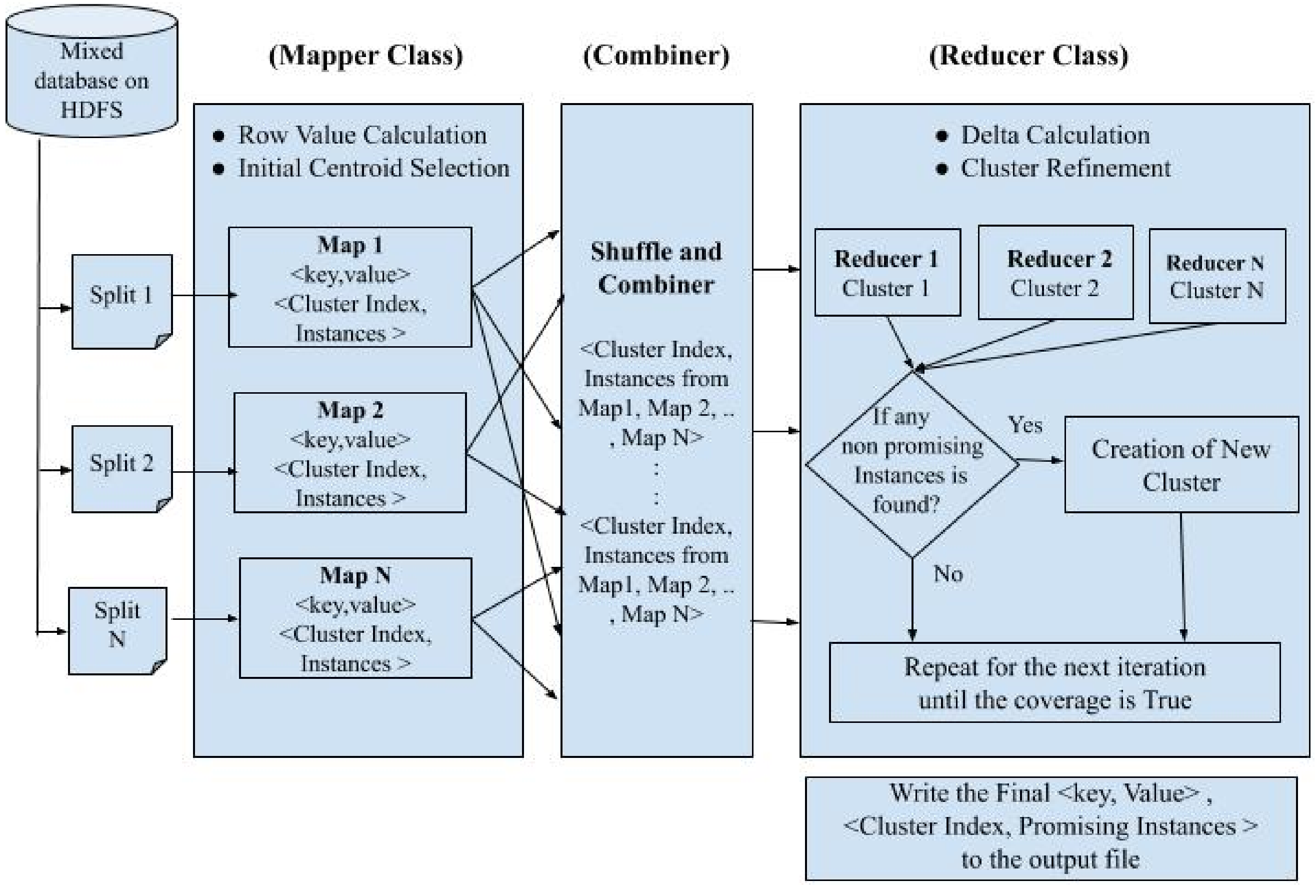

Mixed database on HDFS
(Mapper Class)
(Combiner)
(Reducer Class)
• Row Value Calculation
• Initial Centroid Selection
Split 1
Split 2
Split N
Map 1 <key,value> <Cluster Index, Instances >
Map 2 <key,value> <Cluster Index, Instances >
Map N <key,value> <Cluster Index, Instances >
Shuffle and Combiner
<Cluster Index, Instances from Map1, Map 2, .. , Map N>
:
:
<Cluster Index, Instances from Map1, Map 2, .. , Map N>
• Delta Calculation
• Cluster Refinement
Reducer 1 Cluster 1
Reducer 2 Cluster 2
Reducer N Cluster N
If any non promising Instances is found?
Yes
No
Creation of New Cluster
Repeat for the next iteration until the coverage is True
Write the Final <key, Value> , <Cluster Index, Promising Instances > to the output file

Fig. 2. Map Reduce Paradigm for MR-REDIC Algorithm

instances to appropriate clusters, and it also creates the new cluster and update the value of K also. The intermediate list produced by Mapper Class is assigned as an input to Reducer Class. For refinement of clusters the decision value parameter δ_k is calculated using the formula proposed in [12]. The values of $dist(I_i, C_k)$ will compare with δ_k. If $dist(I_i, C_k) > \delta_k$ then that particular instance I_i is not promising for cluster C_k . The instance I_i will be removed from the cluster C_k and the new cluster will be formed and centroid value of both the cluster will be updated. The final $< Key, Value >$ pair is maintained where Key is the updated Cluster Index and Value is the value all the attributes of of particular instance. Finally the results file of the Reducer Class will be stored to the Output Directory path set by Job Class.

IV. Result and Analysis

The MR-REDIC is implemented in Hadoop architecture with a single node and multi node cluster environment. The dataset used for validation are varies in size to measure out the execution time and speed up of the proposed algorithm. This section divides into three parts:The experimental setup and configuration of nodes, the details of the considered dataset, and the experimental result.

A. Experiment Setup

For implementation of the proposed algorithm using Map Reduce paradigm, Hadoop 2.6.0 and Java version 1.7.0 is considered. Operating system used is Ubuntu 14.04.The experiment was carried out on different a Hadoop cluster.

Configuration 1: Name node and Data Node on a single machine
Configuration 2: One Name Node and Two Data Node
Configuration 3: One Name node and Four Data Nodes

The nodes in the Hadoop cluster are configured with Intel $i3\alpha3.64GHZ$ processor, $2GB$of RAM for each node and 500GB of hard disk with a measured bandwidth for end-to-end TCP sockets of $100MB/s$.

B. Dataset Description

The dataset used in this experiment are downloaded from the kaggle and UCI Repository. To evaluate the performance for the execution time the dataset that of different size and mixed attributes are chosen.
Poker Hand data set: This data set is having multivariate characteristics with 1025010 instances and 11 numbers of

Algorithm 2: Mapper Class

```
Input: Global variable centers C[ ], the offset key, the
       sample value
Output: <key,value > pair,
        key: The index of the closest center point
        value: he values of all attributes of particular
        Cluster Index
Method:
Construct the sample instance from value;
if C [] is null then
    for ∀ row do
        Calculate $\varrho$ ;
        $\varrho_{min} = \min \varrho_1, \varrho_2 ... \varrho_i$ ;
        $\varrho_{max} = \max \varrho_1, \varrho_2 ... \varrho_i$ ;
        $\varrho_{min}, \varrho_{max} \in R$;
    end
    $k = |R|$
end
else
    $CN_i = \{CN_1 , CN_2 ..., CN_p\}$;
    Calculate $dist(I_i, C_k)$;
    Assign instance to the cluster $C_i$ to $CN_i$ having
      minimum cluster distance;
end
key → Cluster Index ;
value → The values of all attributes of particular Cluster
  Index ;
return <key,value > pair
```

Algorithm 3: Reduce Class

```
Input: Key is the index of the cluster, Value is the
       Instance Value
Output: <key,value > pair, where
        key: the index of the cluster
        value: The values of all attributes of new
        Cluster Index
Method:
Calculate $\delta_k$ for Cluster $C_k$ ;
if $dist(I_i, C_k) > \delta_k$ then
    Remove $I_i$ from cluster $C_k$ ;
    k= k+1;
    Create Cluster $C_{k+1}$, where $I_i \in C_{k+1}$ ;
    update $C_n$ set globalF = True;
end
Key → cluster Index ;
Value → The values of all attributes of new Cluster
  Index ;
return <key,value >
```

attributes. Each instance of dataset is an example of a hand which consist of strategy of how five playing cards are to be drawn from a deck of 52 cards. [15]

Indian Census dataset: This dataset consist of 156 attributes of mixed type. It gives the information about population based on demographic data for each district of India.

Magic Gamma Telescope Dataset The imaging technique used by telescope, captures the high energy gamma particles, as gamma rays emits radiation by charged particles in an electromagnetic shower. Every event of this radiation is described by the various parameters, like major axis of ellipse [mm] minor axis of ellipse [mm] etc,. out of which 10 , 10-log of sum of content of all pixels etc. Here 10 different numerical attributes are considered. The instances are is divided into two major categories :gammas (signal) and hadrons (background). [16]

House Sales in King County, USA [17] This dataset having the information about house sold prices for USA between May 2014 and May 2015. It contains the record of 21614 houses with 21 different properties like, area, longitude, latitude etc.

Toy Dataset: This toy dataset comprised of 150000 instance with 6 attributes. Size of the dataset is 5 MB. Here each instance is described by the features like age, gender, income, city etc. [18] The brief introduction of dataset is given in table I

C. Evaluation Parameters and Results

In these experiments three evaluation measurements are considered: Execution time. Speed up and Scale up .

1) Execution Time: The execution time of the proposed algorithm MR REDIC is compared with the REDIC algorithm, here the comparison is done with 3 different configurations. For Configuration 1, Data Node and Name Node is on single machine, for Configuration 2 One Name Node and Two Data Nodes are considered, Further in Configuration 3 One Name Node and 4 Data Nodes are considered. By observing the values of table II, it is verified that the execution time is decreasing gradually for each configuration.
The figure 3 shows the execution time is decreasing as number of data nodes will increase.

2) Speed up: : To measure out the Speed up , the number of nodes are increased in every configuration by keeping the fix size of

$$Speedup = \frac{T_1}{T_n}$$

T_1 is execution time on 1 node and T_n is execution time for the n node. [19]

The linear Speed up is the ideal case of Map Reduce paradigm. For example , the speed up of algorithm is increasing from 1 to n while splitting the task from 1 machine to n machines. the In real time it is difficult to achieve, as by increasing the number of nodes, the communication overhead will also be considered. To evaluate the Speed up parameter of the proposed algorithm, five different dataset and the 3

TABLE I. DESCRIPTION OF DATASET

Dataset	Size in MB	No of Instances	Total Attribute	Numerical Attribute	Categorical Attribute
Poker Hand	0.61	1025010	11	5	6
Indian Census	1.3	1909	156	151	5
Magic Gamma	1.5	19020	11	11	0
House Sales	2.4	21614	20	19	1
Toy	5	150000	6	3	3

TABLE II. EXECUTION TIME OF MR-REDIC

Dataset	Execution Time in seconds			
		Proposed Algorithm MR REDIC		
	REDIC	Configuration 1	Configuration 2	Configuration 3
Poker Hand	5.23	2.45	1.18	0.58
Indian Census	6.85	2.32	1.19	1.02
Magic Gamma	8.00	2.40	1.19	0.39
House Sales	11.89	6.26	6.20	4.35
Toy	29.67	8.13	7.45	4.05

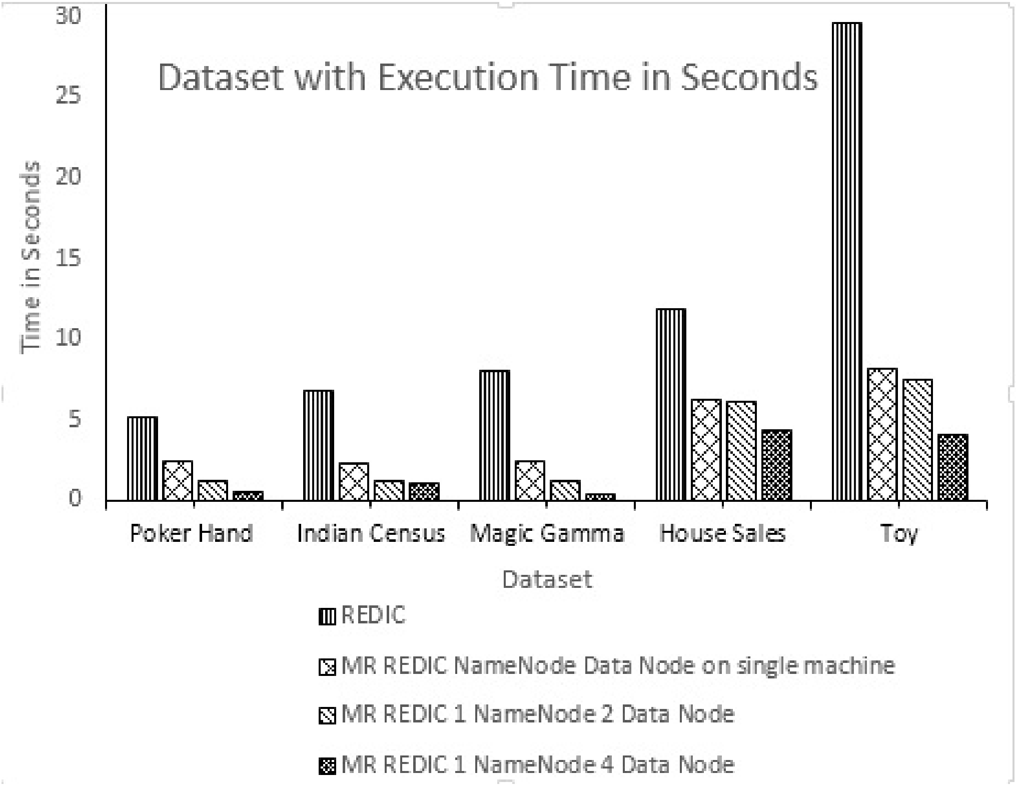


Fig. 3. Comparative Analysis of Execution time

different configuration are considered.From the below figure 4 it is observed the Speed up value is slightly drifting from linear in case of 1 Node to 2 Node, but while considering the Speed up from 2 nodes to 4 nodes it almost closer to linear.

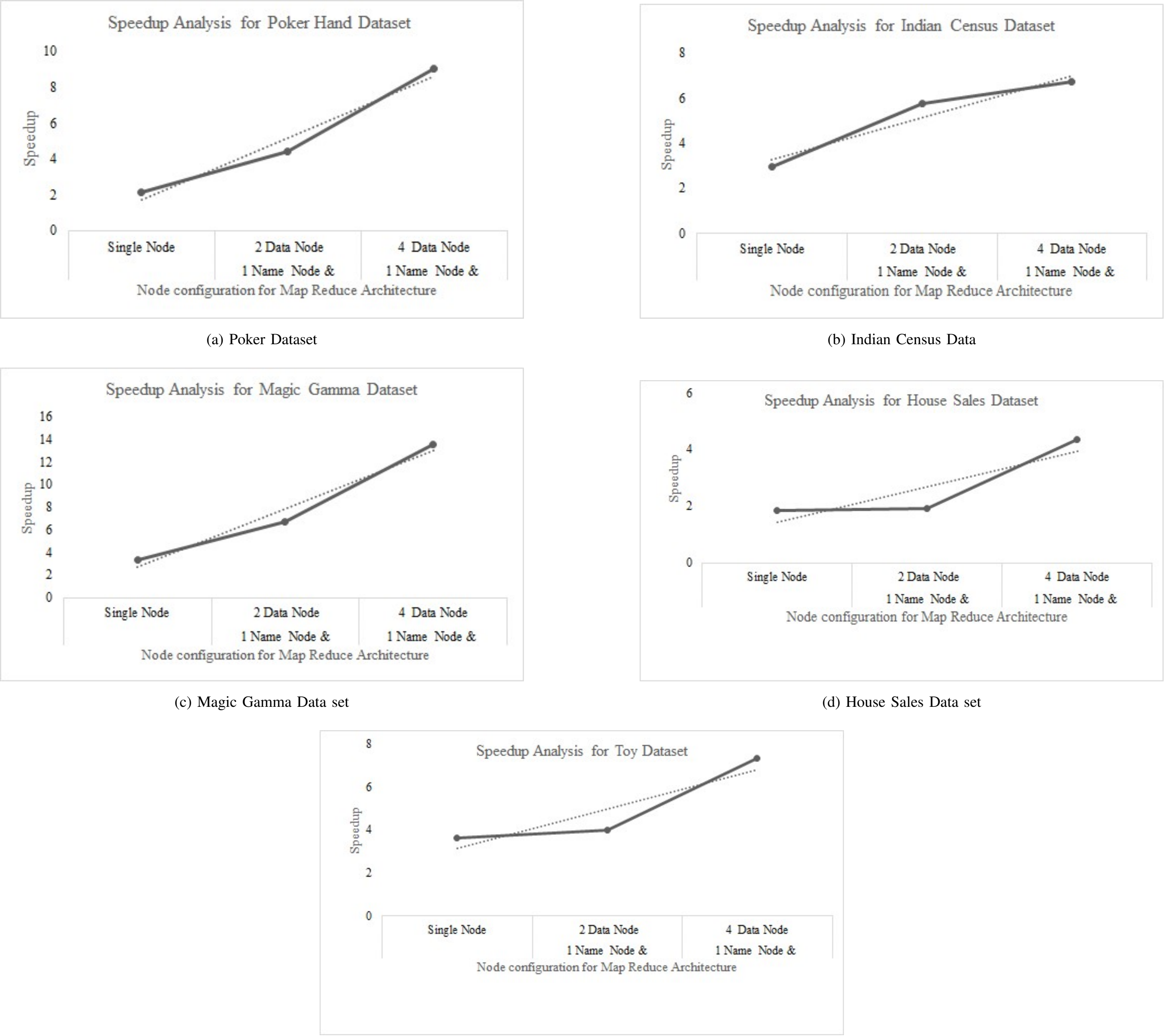
3) Scale up: To measure out the Scale up , the size of dataset is increased by keeping the fixed configuration of Data Node and Name Node. It is evaluated by using formula,

$$Scaleup = \frac{TD_n}{T2D_n}$$

TD_n is execution time of dataset of size D for n Data Nodes
$T2D_n$ is execution time of dataset of size 2D for n Data Nodes [19]

The constant value of Scale up for each size of dataset the ideal case of Map Reduce paradigm.In real time it is difficult to achieve, as execution time depends on the different values of attributes for particular instances.

The scale up analysis is calculated in table IV. To evaluate the Scale up parameter of the proposed algorithm, Toy dataset is considered. Here the three instances of dataset with 1MB,2MB, 4 MB sizes are considered. The execution time for different instances of dataset is recorded in table III. From



(a) Poker Dataset

(b) Indian Census Data

(c) Magic Gamma Data set

(d) House Sales Data set

(e) Toy Data set

Fig. 4. Speed up Analysis for different dataset

TABLE III. EXECUTION TIME FOR DIFFERENT INSTANCES OF TOY DATASET

	Execution time in seconds		
Instances of Toy DataSet	**Single Node**	**1 Name Node and 2 Data Node**	**1 Name Node and 4 Data Node**
1 MB	1.61	1.39	1.23
2 MB	3.34	2.67	2.52
4 MB	6.62	5.43	4.98

TABLE IV. SCALE UP ANALYSIS OF TOY DATASET

Measurement for Speed up Parameter	Single Node	1 Name Node and 2 Data Node	1 Name Node and 4 Data Node
1 MB to 2MB	0.48	0.52	0.48
2 M to 4MB	0.50	0.49	0.49

the table IV, it is observed that In different configurations of Map Reduce architecture with variation in size the value of Scale up parameter ranges from 0.48 to 0.52, which is nearer to constant. The graph for the same result is shown in the Figure 4.

V. Conclusion

In this paper,the MR-REDIC algorithm is proposed, in which the REDIC K-prototype algorithm is migrated to Map Reduce model for making it suitable to create cluster of mixed data having large scale. The REDIC K-Prototype Clustering is efficient due to its simple calculation and eliminating the dependency on prior parameters. It has proved its efficiency for real time mixed data-sets. But it requires more time for large scale data, so here its parallel version using Map Reduce paradigm is proposed. Experimental results were conducted with five benchmark data-sets and three different configuration of Map Reduce paradigm. In Speed up comparative analysis it has shown the near to linear increase approach, and in Scale up comparative analysis the algorithm gives almost constant value.n In future the different similarity measurement for numerical attributes can be integrated with MR REDIC algorithm.

References

[1] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*. John Wiley & Sons, Ltd, Jan. 2011. [Online]. Available: https://doi.org/10.1002/9780470977811

[2] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[3] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.

[4] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 881–892, 2002.

[5] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[6] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 5, pp. 657–668, 2005.

[7] A. Ahmad and S. S. Khan, "Survey of state-of-the-art mixed data clustering algorithms," *IEEE Access*, vol. 7, pp. 31 883–31 902, 2019.

[8] W.-D. Zhao, W.-H. Dai, and C.-B. Tang, "K-centers algorithm for clustering mixed type data," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2007, pp. 1140–1147.

[9] S. Harous, M. Al Harmoodi, and H. Biri, "A comparative study of clustering algorithms for mixed datasets," in *2019 Amity International Conference on Artificial Intelligence (AICAI)*. IEEE, 2019, pp. 484–488.

[10] K. R. Nirmal and K. Satyanarayana, "Redic k prototype clustering algorithm," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 18, pp. 5027–5036, 2018.

[11] Z. He, S. Deng, and X. Xu, "Improving k-modes algorithm considering frequencies of attribute values in mode," in *International Conference on Computational and Information Science*. Springer, 2005, pp. 157–162.

[12] K. R. Nirmal and K. Satyanarayana, "Redic k-prototype clustering algorithm for mixed data (numerical and categorical data)," *International Journal of Recent Technology and Engineering*, vol. 7, no. 6, pp. 1–6, 2019.

[13] P. Fränti and S. Sieranoja, "How much can k-means be improved by using better initialization and repeats?" *Pattern Recognition*, vol. 93, pp. 95 – 112, 2019.

[14] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[15] "Uci machine learning repository: Poker hand data set," http://archive.ics.uci.edu/ml/datasets/Poker+Hand, (Accessed on 01/12/2019).

[16] "Uci machine learning repository: Magic gamma telescope data set," https://archive.ics.uci.edu/ml/datasets/magic+gamma+telescope, (Accessed on 01/12/2019).

[17] "House sales in king county, usa — kaggle," https://www.kaggle.com/harlfoxem/housesalesprediction/data, (Accessed on 01/12/2019).

[18] "Toy dataset — kaggle," https://www.kaggle.com/carlolepelaars/toy-dataset, (Accessed on 01/12/2019).

[19] X. Yue, W. Man, J. Yue, and G. Liu, "Parallel k-medoids++ spatial clustering algorithm based on mapreduce," *arXiv preprint arXiv:1608.06861*, 2016.

A Technique for Panorama-Creation using Multiple Images

Moushumi Zaman Bonny[1] and Mohammad Shorif Uddin[2]
Department of Computer Science and Engineering, Jahangirnagar University
Dhaka, Bangladesh

***Abstract*—Image stitching, which is a process of integration of multiple images to create a panoramic image using all contents fitted into one frame, finds wide-spread applications in medical, high resolution digital map, satellite and video imaging. This paper proposes a framework to develop panorama image with multiple images. The framework is an automatic process that takes multiple images, checks correlation of the sequential images and removes overlapping area if exists and creates the panorama.We have done experimentations using different image-sets consisting multiple images with and without overlapping and got satisfactory results.**

Keywords—Panorama; image stitching; correlation; multiple images; image features

I. Introduction

Image stitching is a system for creating panoramic images. The overlapping areas of input images can make noticeable seam among the images [1]. So, it is basically a procedure for recognizing and eliminating those seams of the overlapping regions and blending [2].

Image stitching is usually applied in various applications, such as, X-ray image stitching, HDR (High Dynamic Range) image stitching, image stabilization, high resolution photo mosaics in satellite photos, medical imaging, digital maps, multiple-images with super-resolution, video stitching, microscopy image stitching, object insertion and group photographs/panoramas.

There are numerous techniques to construct panoramic images. Image stitching is generally a software-based method for making panorama. It takes several standard images to wrap up the complete viewing area and stitches collectively all those images to produce a panorama [3], [4]. Alignment or registration of images is a requirement for stitching images. It is done considering translation, scaling and rotation of images [5].

There are different types of stitching methods [6]. Direct technique is extremely efficient for stitching images which do not contain any overlapping area. Image stitching using correlation [7], [8] is one of the most primordial stitching techniques, which is mainly an intensity-based method and appropriate for images with overlapping areas.

Another method is feature-based stitching [9], which locates all corresponding feature points in every image pair. Then it evaluates all the features in one image against features in another image through feature descriptors. Feature extraction, registration and blending are different steps required for feature-based image stitching.

Two-image stitching is very common; however, multiple image stitching is somehow tricky. Therefore, the main purpose of this paper is to create a panorama using multiple images. We have prepared our paper as follows: Section II illustrates the literature review. Section III describes a concise representation of our method of multiple image stitching. Section IV presents the experimentation and assesses the performance of our method. Then, Section V comments on future works along with challenges. Finally, Section VI concludes the paper.

II. Literature Review

Many researchers have already been worked of image stitching. A comprehensive evaluation on panorama creation methods is presented below:

Adel et al. [10] presented a feature-based image stitching based on ORB (Oriented FAST (Features from Accelerated Segment Test)) and Rotated BRIEF (Binary Robust Independent Elementary Features).

Zomet et al. [11] proposed a method of seamless image stitching by minimizing false edges. In this approach, their main target was on seam removal.

Mclauchlan et al. [12] presented an image mosaicing process using sequential bundle adjustment. The newness of this approach is the transfer of photogrammetric bundle adjustment and line measurement which enables to use lines in camera-calibration.

Hua et al. [13] presented a method of image stitching based on SIFT (Scale-Invariant Feature Transform) and MVSC (Mean Value Seamless Cloning). It is basically a SIFT feature detection and matching oriented technique. Szeliski [14] presented an image alignment and stitching method. The core focus of this technique is alignment of the images.

Fatah et al. [15] proposed an automatic seamless image stitching technique. Qiu et al. [16] proposed a SIFT (Scale Invariant Feature Transform) and transformation-parameter based image stitching method.

Jia et al. [17] proposed a method of image stitching using structure deformation. Their technique is for getting the stability in image intensity. Ostiak et al. [18] proposed a totally automatic HDR (High Dynamic Range) panorama stitching method which used SIFT for the identification of the matching feature points.

Fang et al. [19] presented a feature-based stitching method of a clear/blurred image pair. Koo et al [20] proposed a feature-based image registration algorithm for image stitching applications on mobile devices.

Zhao et al. [21] presented a self-adaptive algorithm based on Harris-corner detection. Wang et al. [22] presented an automatic image stitching technique based on graph model. In this method, they have used Weighted Shortest Path algorithm and Dijkstra algorithm for multi-image stitching. Yang et al. [23] proposed a phase-correlation and Harris operator-based image-mosaicing process. This approach is a correlation and feature based hybrid method.

All techniques have some advantages and disadvantages [24]. Through these observations, in this paper we have presented methodology to stitch multiple images. Our method presents a robust stitching system where we determine the correlation among the images, if overlapping exists between the images, then we should remove overlapping area from the first image and merge the second image after first one to create panorama and assign newly created panorama as first image and another image as second.

On the other hand, if overlapping does not exist between the images, we will merge second one after first image and produce a panorama and again, we will assign this panorama as the first and another image as the second. We should repeat the above steps till nth inputted image to create the final panorama.

III. Proposed Algorithm

We have proposed a simple but efficient and robust algorithm for panorama-creation using multiple images. The proposed algorithm is as follows:

Algorithm 1: Multiple image stitching algorithm

```
Input: $n \leftarrow N, I(n), I(n-1), c(\lambda) \leftarrow 0, TH \leftarrow 0.2$
Output: $I(n)$
while $n <= N$ do
    if $n = 1$ then
        break
    else
        $c(\lambda) = correlation(I(n), I(n-1))$
        if $c(\lambda) > TH$ then
            Remove overlapping region from $I(n)$
            Merge $I(n-1)$ after $I(n)$
            $n--$
            Term the merged Image as $I(n)$
            Input image $I(n-1)$
        else
            Merge $I(n-1)$ after $I(n)$
            $n--$
            Term the merged Image as $I(n)$
            Input image $I(n-1)$
Display $I(n)$
```

In the proposed method, n represents the number of images, $c(\lambda)$ is the correlation coefficient of two consecutive images, λ represents the width of the overlapped region and TH is the limit of $c(\lambda)$ and its value set to 0.2. The steps of our proposed algorithm are detailed and explained below:

A. Image Acquisition

We must input n images $I(n), I(n-1), \ldots, I(3), I(2), I(1)$ to continue the whole process. But, at first, two consecutive images $I(n)$ and $I(n-1)$should be inputted. Then the rest of the images should be inputted sequentially for the further execution of the process depending on the criteria and conditions. Then the images are converted to gray-scale. Another important thing is all of these images must be aligned.

B. Determination of Correlation

The next step is, determining of correlation. Correlation usually used to find correspondence of two images [25]. So, we have to compute correlation coefficient $c(\lambda)$ between two images A and B using the following formula,

$$c(\lambda) = \frac{\sum_x \sum_y (A - A^{'})(B - B^{'})}{\sqrt{(\sum_x \sum_y (A - A')^2)(\sum_x \sum_y (B - B')^2)}} \tag{1}$$

Where, $A^{'} = mean(A_{xy}), B^{'} = mean(B_{xy})$.

C. Removal of Overlapping Region and Panorama Creation

If the value of correlation coefficient $c(\lambda)$ surpasses the threshold value TH, then we should remove overlapping region λ from $I(n)$ and merge $I(n-1)$ after $I(n)$ to create panorama and decrement the value of n and assign newly created panorama as $I(n)$ and input next consecutive image as $I(n-1)$ and go through the previous step.

On the other hand, if correlation does not exceed the threshold value TH which means there is no common region between the images and we don't need to remove any overlapping region. So, we will merge $I(n-1)$ after $I(n)$ to generate a panorama and then decrement the value of n. After that,we will assign the newly-produced panorama as $I(n)$ and acquire the next consecutive image $I(n-1)$. The above steps will be repeated depending on different measures and conditions till the last input image to create the ultimate panorama image.

IV. Results and Discussions

The experiments are done on ten image-sets using our technique. We have prepared input images from these ten original images using Microsoft Office Picture Manager. In seven image-sets, first 5 images contain a portion of overlapping region as well as repetition, the next 4 images does not contain any overlaps and 9th and 10th images contain a segment of overlapping area, but in the next three image-sets, all the images contain some amount of overlapping area.

Then, we used these images to generate the panorama using our stitching method. All the outputs are generated using MATLAB R2018a with Microsoft Windows platform, Intel Core i3, 3.50 GHz processor and 4.00 GB RAM. As a sample we have demonstrated the results of four image-sets here. Figure 1 depicts an image (2725×600 pixels) which is considered as a ground truth. This image is cut into ten images shown in Figure 2. Among those, the first five consecutive

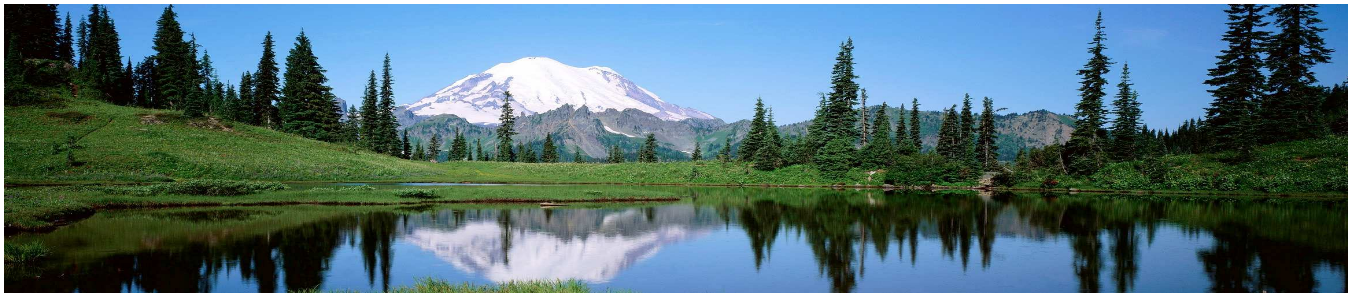


Fig. 1. Original image (treated as ground truth panorama).

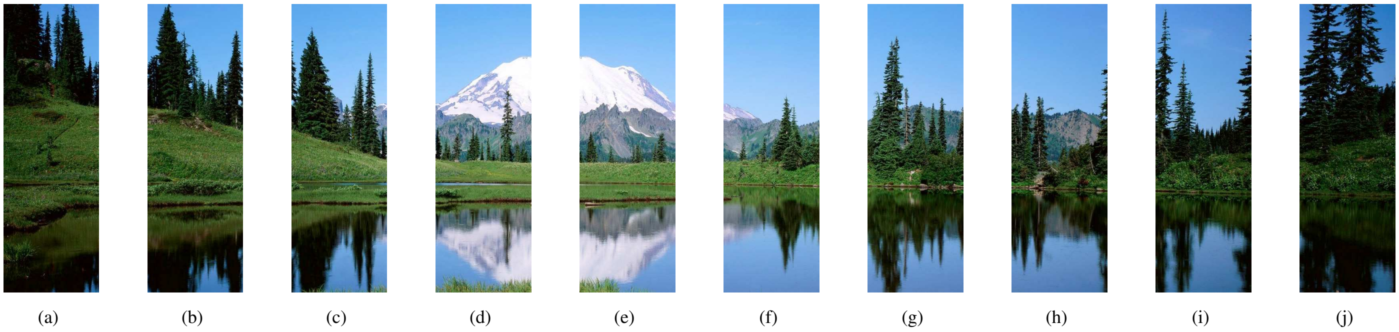


Fig. 2. Ten input images: (a) and (b) contains 5 pixels of overlapping area,(b) and (c) contains 10 pixels of overlapping area,(c) and (d) contains 15 pixels of overlapping area ,(d) and (e) contains 20 pixels of overlapping area, (e) and (f) contains no overlapping area,(f) and (g) contains no overlapping area, (g) and (h) contains no overlapping area, (h) and (i) contains no overlapping area, (i) and (j) contains 25 pixels of overlapping area.

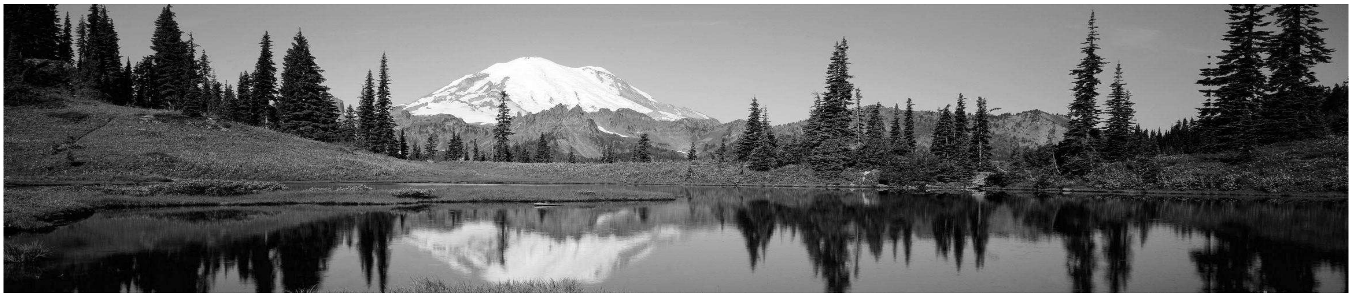


Fig. 3. Panorama using proposed method.

images contain some overlapping regions, then the next four do not have any overlaps but the last two images contain some overlaps. Then we inputted all these images consecutively for the proposed technique to create desired panorama. Here, Figure 3 shows the final panoramic image.

Similar to Figure 1, we have done the similar actions on an X-ray image(2325×650 pixels) shown in Figure 4. So, it is cut into ten images which are depicted in Figure 5 and Figure 6 shows the generated panorama using the proposed technique. Then, we have taken another image (2325×600 pixels) as the ground truth (shown in Figure 7) and cut it into ten images which are shown in Figure 8. Figure 9 demonstrates the panorama of these images using our stitching method. Similarly, Figure 10 shows another original image (2275×600 pixels), Figure 11 presents ten parts of this image and each of these parts contain some portion of overlaps. Figure 12 shows the panorama generated from these images using our proposed method.

TABLE I. Performance Evaluation of the Proposed Algorithm

Input Images	Accuracy (%)	Computation Time (sec)
Image Set 1	98.59	13.17
Image Set 2	100.00	4.95
Image Set 3	98.78	5.16
Image Set 4	99.10	5.40
Image Set 5	98.86	6.18
Image Set 6	99.34	5.58
Image Set 7	97.41	5.44
Image Set 8	98.42	4.77
Image Set 9	98.94	3.88
Image Set 10	99.03	5.24

Table I depicts the performance (accuracy and computation time) of our proposed technique. We have used the following equation to calculate the accuracy:

$$Accuracy = 1 - \frac{\sum(\sum(|(I_1 - I_2)|))}{\sum(\sum(I_1)} \qquad (2)$$

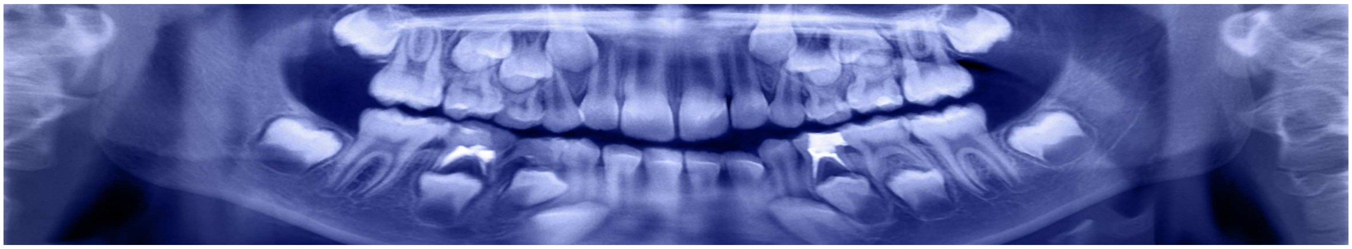


Fig. 4. Original image (treated as ground truth panorama).

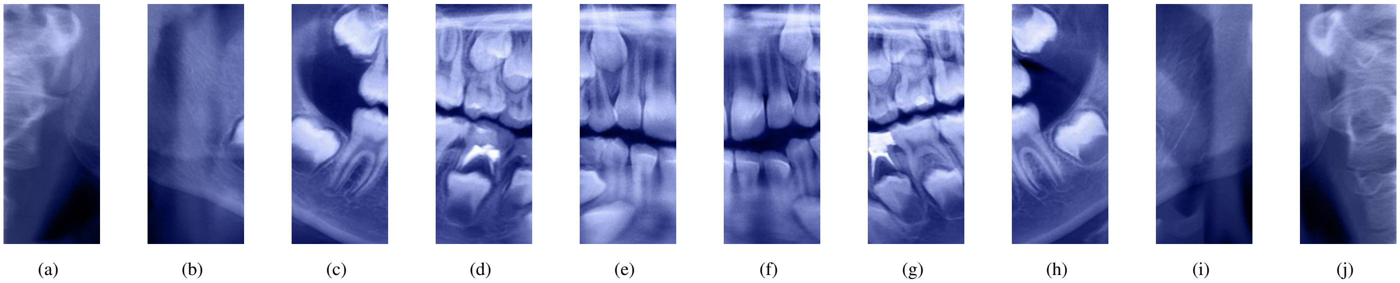


Fig. 5. Ten input images: (a) and (b) contains 5 pixels of overlapping area,(b) and (c) contains 10 pixels of overlapping area,(c) and (d) contains 15 pixels of overlapping area ,(d) and (e) contains 20 pixels of overlapping area, (e) and (f) contains no overlapping area,(f) and (g) contains no overlapping area, (g) and (h) contains no overlapping area, (h) and (i) contains no overlapping area, (i) and (j) contains 25 pixels of overlapping area.

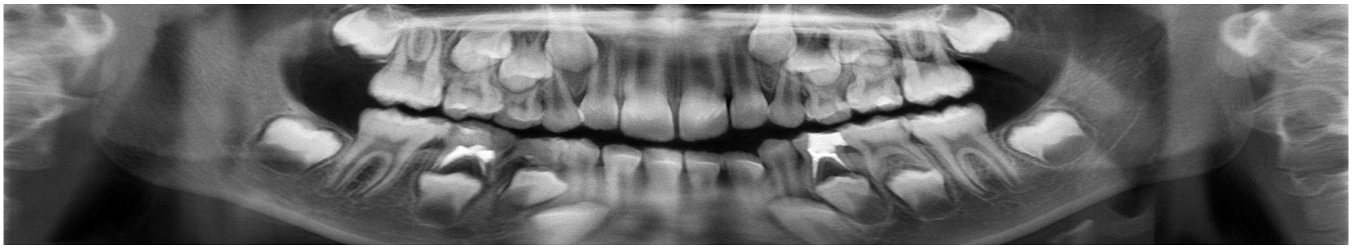


Fig. 6. Panorama using proposed method.

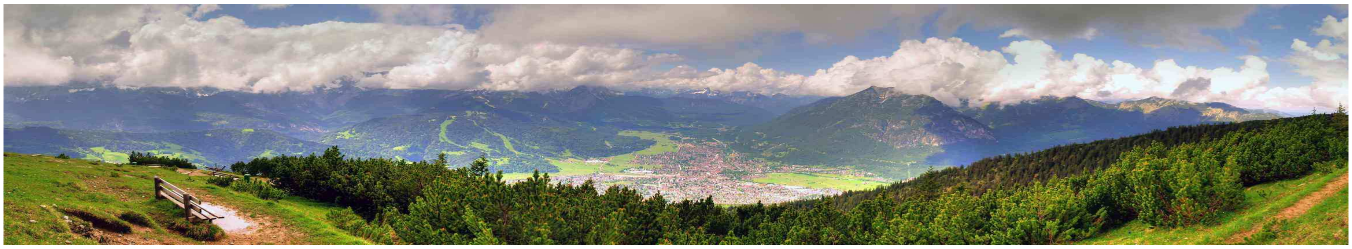


Fig. 7. Original image (treated as ground truth panorama).

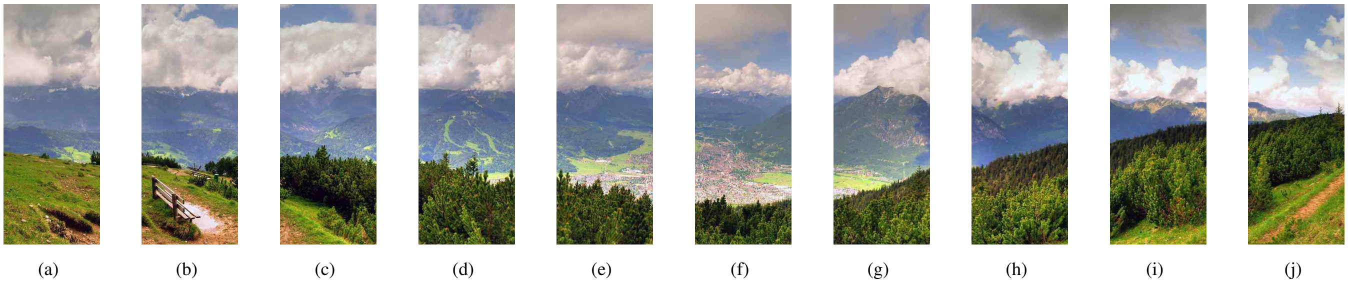


Fig. 8. Ten input images: (a) and (b) contains 5 pixels of overlapping area,(b) and (c) contains 10 pixels of overlapping area,(c) and (d) contains 15 pixels of overlapping area ,(d) and (e) contains 20 pixels of overlapping area, (e) and (f) contains no overlapping area,(f) and (g) contains no overlapping area, (g) and (h) contains no overlapping area, (h) and (i) contains no overlapping area, (i) and (j) contains 25 pixels of overlapping area.

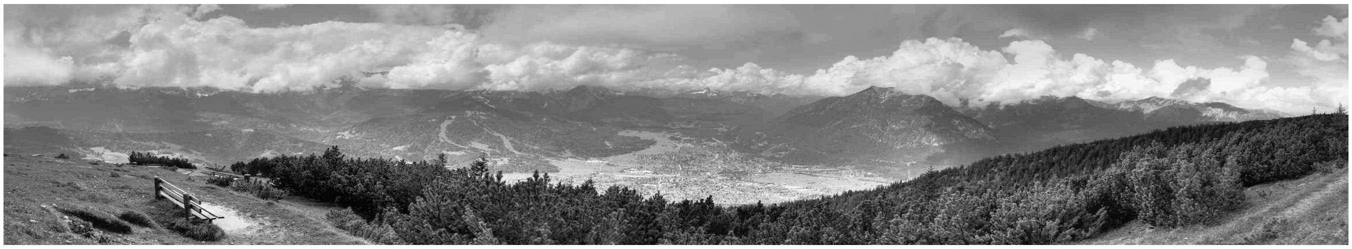


Fig. 9. Panorama using proposed method.

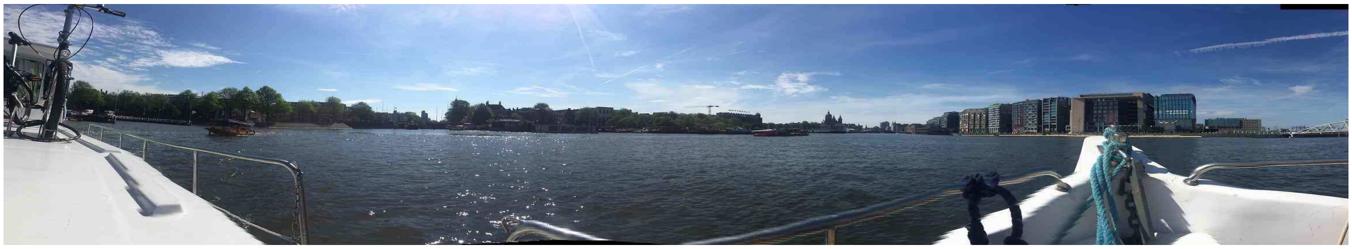
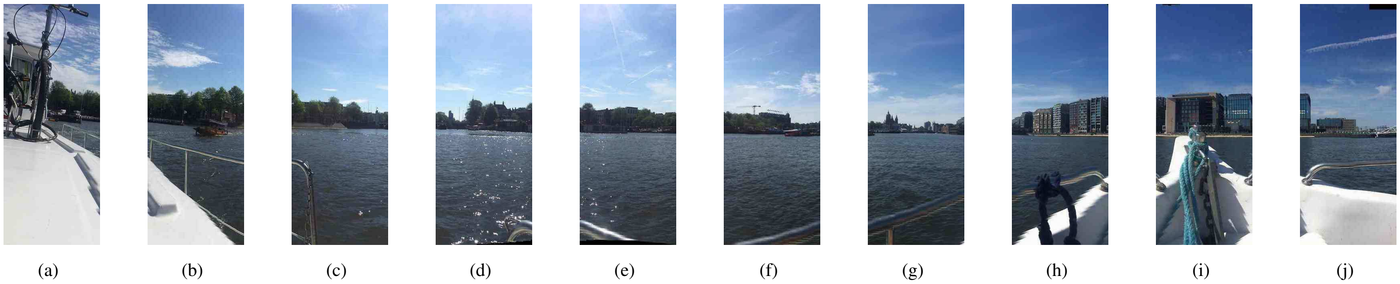


Fig. 10. Original image (treated as ground truth panorama).



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 11. Ten input images: (a) and (b) contains 5 pixels of overlapping area,(b) and (c) contains 10 pixels of overlapping area,(c) and (d) contains 15 pixels of overlapping area ,(d) and (e) contains 20 pixels of overlapping area, (e) and (f) contains 5 pixels of overlapping area,(f) and (g) contains 10 pixels of overlapping area, (g) and (h) contains 15 pixels of overlapping area, (h) and (i) contains 20 pixels of overlapping area, (i) and (j) contains 25 pixels of overlapping area.

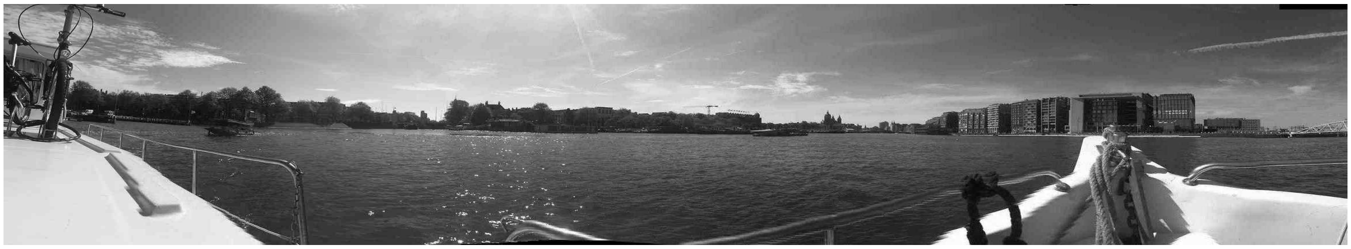


Fig. 12. Panorama using proposed method.

In Eq. (2), I_1 is the ground truth and I_2 is the output of the proposed method.

V. Future Works and Challenges

In this paper, we have experimented with natural-, HDR- and X-ray-image and got satisfactory outputs. Multiple image stitching is a challenging issue. Seam is another important issue needs to be eliminated. Different methods have been presented in the recent years and we have a vision to work on seam elimination for producing quality panorama. The experiments are done on sequential images with straight edges, but, we have a desire to work on unsequenced images with uneven edges which is another challenging matter.

VI. Conclusion

In this paper, we have publicized a structure to construct a panoramic image using sequential multiple images. Some of these images consist of overlapping regions and some images do not have any overlapping area. We have eliminated the overlapping area from the first one before joining the next image after it and assigned the newly produced image as the first one and inputted the next consecutive image as the second.

On the other hand, if overlapping does not exist between the images, we have combined the second one after the first image and assigned this image as the first one and the next consecutive image as the second. We repeated the preceding steps depending on the existence of overlap, until the last image to create the final panorama.We have simulated our method on different images and got adequate outcomes.

REFERENCES

[1] L. Juan, G. Oubang,“SURF applied in panorama image stitching, Image Processing Theory Tools and Applications (IPTA),” 2010 2nd International Conference, pp.495 – 499, Jul 2010.

[2] R.Karthik, A.Annis Fathima,V.Vaidehi, “Panoramic View Creation using Invariant Moments and SURF Features,” IEEE International Conference on Recent Trends in Information Technology(ICRTIT), pp. 376-382, July 2013.

[3] Deepak Kumar Jain, Gaurav Sexena, Vineet Kumar Singh,“Image mosaicing using corner technique,” IEEE International Conference on Communication Systems and Network Technologies,pp.79-84, may 2012.

[4] Mathew Brown, David G. Lowe, “Automatic Panoramic Image Stitching using Invariant Features,” International Journal of Computer Vision, vol. 74, pp.59–73, Aug 2007.

[5] Brown M. and Lowe, “Recognising Panoramas,” Ninth IEEE International Conference on Computer Vision, vol. 2,pp.1218-1225, 2003.

[6] Xian-Guo Li, Chang-Yun Miao and Yan Zhang, “An Algorithm for Selecting and Stitching the Conveyer Belt Joint Images Based on X-ray,” IEEE International Conference on Intelligent Computation Technology and Automation,vol. 1, pp. 474-477, May 2010.

[7] Tao-Cheng Chang, Cheng-AnChien, Jia-Hou Chang, Jiun-In Guo, “A Low-Complexity Image Stitching Algorithm Suitable for Embedded Systems,” IEEE International Conference on Consumer Electronics(ICCE),pp. 197-198, January 2011.

[8] Patrik Nyman, “Image Stitching using Watersheds and Graph Cuts,” Research Article, Lund University, Sweden.

[9] Yingen Xiong, Kari P ulli, “Fast Panorama Stitching for High-Quality Panoramic Images on Mobile P hones,” IEEE Transactions on Consumer Electronics, vol. 56, no. 2, pp. 298-306, may 2010.

[10] Ebtsam Adel, Mohammed Elmogy, Hazem Elbakry,“Image Stitching System Based on ORB Feature- Based Technique and Compensation Blending,” International Journal of Advanced computer Science and Applications, vol. 6, No. 9, 2015.

[11] A.Zomet, A.Levin, S.Peleg, Y.Weiss,“Seamless Image Stitching by minimizing false edges,” IEEE Transactions on Image Processing, vol. 15, No. 4, pp. 969-977,April 2006.

[12] Mclauchlan,P.F.,Jaenicke,A.,Xh,G.G.,“Image mosaicing using sequential bundle adjustment,” Proc. BMVC, pp.751-759, 2000.

[13] Zhen Hua, YeweiLi, Jin jiangLi,“Image Stitching Algorithm Based on SIFT and MVSC,” IEEE 7th International Conference on Fuzzy Systems and Knowledge Discovery,vol. 6, pp. 2628-2632,Aug 2010.

[14] Szeliski,“Image Alignment and Stitching,” Tech. rep., December 10,2006.

[15] Russol Abdel Fatah, Dr. HaithamOmer,“Automatic Seamless of Image Stitching,” Journal of Computer Engineering and Intelligent systems,vol. 4, No. 11, 2013.

[16] Pengrui Qiu, Ying Liang and Hui Rong,“Image Mosaics Algorithm Based on SIFT Feature Point Matching and Transformation Parameters Automatically Recognizing,” 2nd International Conference on Computer Science and Electronics Engineering , 2013.

[17] Jiaya Jia and Chi-Keung Tang,“Image St itching Using Struct ure Deformation,” IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. 30, pp. 617-631, no. 4, april 2008.

[18] Piotr Ostiak,“Implementation of HDR panorama stitching algorithm,” Semantic Scholar.

[19] Xianyong Fang,“Feature Based Stitching of a Clear/Blurred Image P air,” IEEE International Conference on Multimedia and Signal Processing, vol. 1, pp. 146-150, 2011.

[20] Hyung Il Koo and Nam Ik Cho,“Feature-based Image Regist ration Algorithm for Image St itching Applicat ions on Mobile Devices,” IEEE Transactions on Consumer Electronics, vol. 57, no. 3, pp. 1303-1310, aug2011.

[21] W. Zhao, S. Gong, C. Liu, and X. Shen,“A self-adapt ive Harris corner det ect ion algorit hm,” Computer Engineering, vol. 34, no. 10, pp. 212–214, 2008.

[22] Zhi cheng Wang, Yufei Chen,“An Automatic Panoramic Image Mosaic method based on Graph Model,” In Springer Link. vol. 75, Issue 5, pp. 2725-2740, 2016.

[23] Fan Yang, Linlin WEI, Zhiwei ZHANG, Hongmei TANG, “Image Mosaic Based on Phase Correlation and Harris Operator,” Journal of Computational Information Systems, vol. 8, pp. 2647–2655, 2012.

[24] Y. Deng and T. Zhang,“Generating Panorama Photos,” International Society for Optics and Photonics, pp. 270-279, 2003.

[25] Feng Zhao, Qingming Huang, Asn Gao, “Image Matching By Normalized Cross-Correlation,” Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China vol.60, no.1, pp.63-86, 2004.

Comparison of Accuracy between Long Short-Term Memory-Deep Learning and Multinomial Logistic Regression-Machine Learning in Sentiment Analysis on Twitter

Aries Muslim[1], Achmad Benny Mutiara[2], Rina Refianti[3], Cut Maisyarah Karyati[4], Galang Setiawan[5]
Faculty of Computer Science and Information Technology
Gunadarma University
Jl. Margonda Raya No.100, Depok 16424, Indonesia

***Abstract*—The paper is about sentiment analysis research on Twitter. In this research data with the keyword, 'Russian Hacking' concerning the 2016 US presidential election on Twitter was taken as a dataset using Twitter API with Python programming language. The first process in sentiment analysis is the cleaning phase of tweet data, then using the Lexicon-based method to produce positive, negative, and neutral sentiment values for each tweet. Data that has been cleaned and classified will be processed in the Deep learning method with Long Short-Term Memory (LSTM) algorithm and Machine learning method with Multinomial Logistic Regression (MLR) algorithm. The accuracy of these two classification methods are calculated using the confusion-matrix method. The accuracy obtained from the LSTM classification method is 93 % and the MLR classification method is 92 %. Thus, it can be concluded that LSTM is better in classifying sentiments compared to MLR.**

Keywords*—Sentiment analysis; deep learning; machine learning; Long Short-Term Memory (LSTM); Multinomial Logistic Regression (MLR)

I. Introduction

The demand for information technology needs in this era is increasing, especially in terms of communication. One form of technological progress is social media. Twitter is becoming the dominant form of social media. Twitter is a website that is a service of microblog, which is a form of a blog that limits the size of each post, which provides facilities for users to be able to write messages in Twitter updates which contain only 140 characters [21], [14]. Twitter users can express various opinions and opinions. Forms of expressions written by users on twitter are called *tweets*. The number of large tweets shared by Twitter users every second, making the collection of tweets can be processed and analyzed to find out a review or public opinion about a particular product, service, or topic.

Sentiment analysis [1], [3], [19] is a research branch of text mining that is used to analyze and classify opinions from a text document. Sentiment analysis is the process of extracting, processing and understanding textual data to get information in the form of opinions or tendencies of opinion on a problem or object by someone, whether it tends to have a negative or positive opinion or opinion. Sentiment analysis can be done with many methods, one of which is the Machine learning and Deep learning methods [4].

Based on these reasons, the authors are interested in conducting sentiment analysis research. The research conducted is to make a comparison on the application of the Deep learning classification method using the Long Short-Term Memory (LSTM) algorithm and Machine learning classification method using Multinomial Logistic Regression (MLR) algorithm. The application of the classification method is done by using data taken from Twitter with the help of the Twitter API application and the Python programming language. The analyzed topic is the case of Russian hacking that concerns the US presidential election in 2016. Data are taken via Twitter based on Hastag (#) relating to research topics. The accuracy results obtained using the LSTM classification method will be compared with the MLR classification method, so that it can be known which algorithm is capable of producing classification methods with better accuracy values.

In the rest of paper, we show briefly the literature review and related work in Section II. In Section III the research methodology is presented. The implementation and results related to our research are also shown in Section IV. The last section is conclusion and future work of our research.

II. Review of Literature

A. Sentiment Analysis

Sentiment analysis is a type of natural language processing to track people's moods about certain products or topics. Sentiment analysis, which is also called opinion mining [17], involves building a system to collect and examine opinions about products or services made in web posts, blogs, or comments on social media. Sentiment analysis is useful in many ways. For example, in marketing [7], [10], it helps the success of new advertisements or product launches, determining which version of the product or service is popular and even identifying the types of demographics that like or do not like certain features.

The basic task in sentiment analysis is to classify the text that is in a sentence or document, then determined the opinions expressed in sentence or whether the document is positive, negative or neutral. As in [15], [16], sentiment analysis can also express emotional feelings of sadness, joy, or anger.

B. Lexicon-based Method

The Lexicon-based method can [16], [20]: i) identify the sentiments of each opinion words contained in the tweet data, and ii) handle multi-opinion problems in a data. This method is an improvement from the method that cannot handle multi-opinion problems. In handling the multi-opinion problem, this method collects all sentiments from the word opinion based on the distance between the words of opinion and its features. So that finally it can be used to determine the class of opinion of each data. In this method, the data is divided into three parameters of sentiment analysis, namely positive, negative, and neutral.

C. Long Short Term Memory (LSTM)

Long Short Term Memory (LSTM) is another type of processing module for Recurrent Neural Network (RNN). LSTM was created by Hochreiter & Schmidhuber (1997)[11] and later developed and popularized by many researchers [9], [8], [2], [5], [6], [13]. Like RNN, the LSTM network also consists of modules with repetitive processing. The difference is that the modules that make up the LSTM network are LSTM modules [18], as seen in Figure 1:

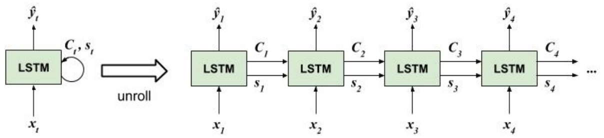


Fig. 1. LSTM Network

The LSTM module (one green box) has different processing from the regular RNN module. Another difference is the addition of a signal given from one time step to the next time step, namely the context (memory cell), represented by the symbol C_t in Figure 2.

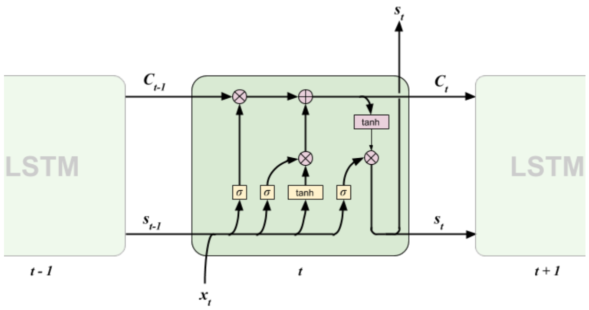


Fig. 2. Processing in LSTM

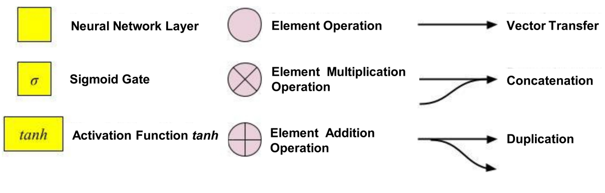


Fig. 3. Description of Processing Notation in LSTM

The diagram in Figure 3 explains that each line carries the entire vector, from the output of one node (node) to another input. The pink circle represents the operation of elements, such as the addition or multiplication of vector elements, while the yellow box is the neural network layer (containing parameters and biases) that can study. Two lines joining indicate a combination of two matrices or vectors, while the split line indicates the content is copied and the copy goes to a different node.

1) LSTM Key Mechanism: The key idea of LSTM is the path that connects the old context (C_{t-1}) to the new context (C_t) at the top of the LSTM module, as illustrated below in Figure 4.

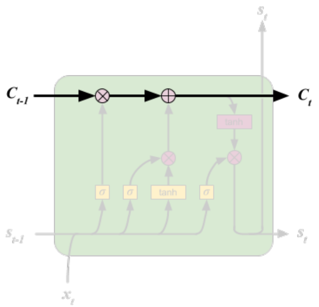


Fig. 4. LSTM Key Mechanism

C_t context is also called *cell state* or *memory cell* in several articles. With the path above, a value in the old context will easily be passed on to a new context with very little modification, if needed. Context is a vector, which we specify as the LSTM network designer. The intuition is, each element we expect to be able to record a feature of input, for example in natural language processing for English, an element recording the gender of the subject, other elements recording whether the subject is singular or plural, etc. These features will be found by LSTM alone in the training process. Another key idea is the sigmoid gate (sigmoid gate) which regulates how much information can pass. Let us see Figure 5, for an input x, the output of the sigmoid gate is $\sigma(A \cdot x + b)$, where A is a parameter, b is biased, both are studied in the training process, and σ is a sigmoid function. The gate output is a

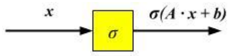


Fig. 5. Sigmoid Gate

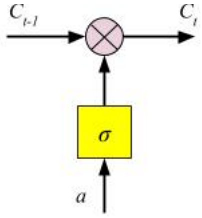


Fig. 6. Case of the Sigmoid Gate

number between zero and one; zero means that the information is totally blocked, while one means the entire information is included. The output from the sigmoid gate will be multiplied by another value to control how much the value is used.

For example, with the sigmoid gate in Figure 6, LSTM can manage how much information from C_{t-1} is included into C_t.

D. Multinomial Logistic Regression (MLR)

Multinomial logistic regression is a logistic regression used when the dependent variable has a polichotomous or multinomial scale with nominal scale response variables with three categories [12]. For regression models with a dependent variable with a nominal category of three categories, the results of Y variable variables are coded 1, 2, and 3. Y variable is parameterized into three logit functions. Logistic regression method is stated in a probability model, namely the model in which the dependent variable is the logarithm of the probability that an attribute will apply in the condition of certain independent variables. As in [12], the model used in MLR is

$$Logit[P(Y=1)] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (1)$$

Using logistical transformation, logistic functions are obtained,

$$\begin{aligned} P_1(x) &= ln\left[\frac{P(Y=1)1 \mid x}{P(Y=1)0 \mid x}\right] \\ &= \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \cdots + \beta_{1n}X_n \\ &= x^{'}\beta_1 \end{aligned} \quad (2)$$

$$\begin{aligned} P_2(x) &= ln\left[\frac{P(Y=1)2 \mid x}{P(Y=1)0 \mid x}\right] \\ &= \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \cdots + \beta_{2n}X_n \\ &= x^{'}\beta_2 \end{aligned} \quad (3)$$

Based on the two logit functions, the trichotomous logistic regression model is obtained as follows:

$$\pi_0(x) = \frac{1}{1+\exp(P_1(x))+\exp(P_2(x))} \quad (4)$$

$$\pi_1(x) = \frac{\exp(P_1(x))}{1+\exp(P_1(x))+\exp(P_2(x))} \quad (5)$$

$$\pi_2(x) = \frac{\exp(P_2(x))}{1+\exp(P_1(x))+\exp(P_2(x))} \quad (6)$$

III. Research Method

The authors conducted research with the following stages:

A. Retrieval of Data

We perform tweet data retrieval through the service provided by Twitter, namely, Twitter API using the Python programming language by specifying key words or hashtags related to the topic taken until a number of tweets are needed.

B. Pre-Processing

We perform a cleaning process (Figure 7) on the tweet that has been obtained, including eliminating the URL, deleting the hashtag (#) and mention (), changing the negation word with the negation dictionary, deleting duplication of data and classifying the tweet using the Lexicon-based method, as seen in Figure 8 and Figure 9 with the opinion lexicon owned Hu and Liu [16], who divided tweets into positive, negative and neutral classes.

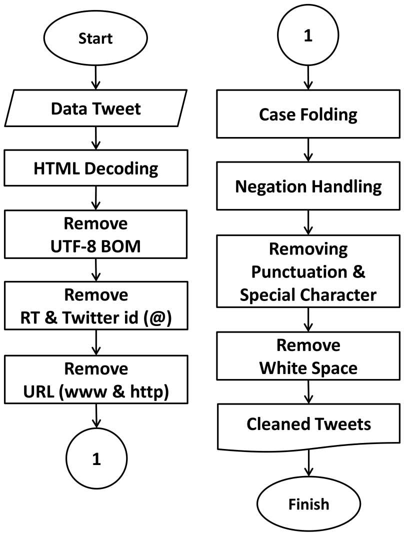


Fig. 7. Tweet Cleaning Process

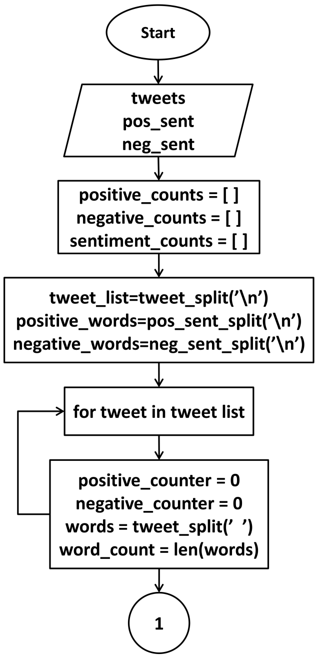


Fig. 8. Lexicon-based Flowchart

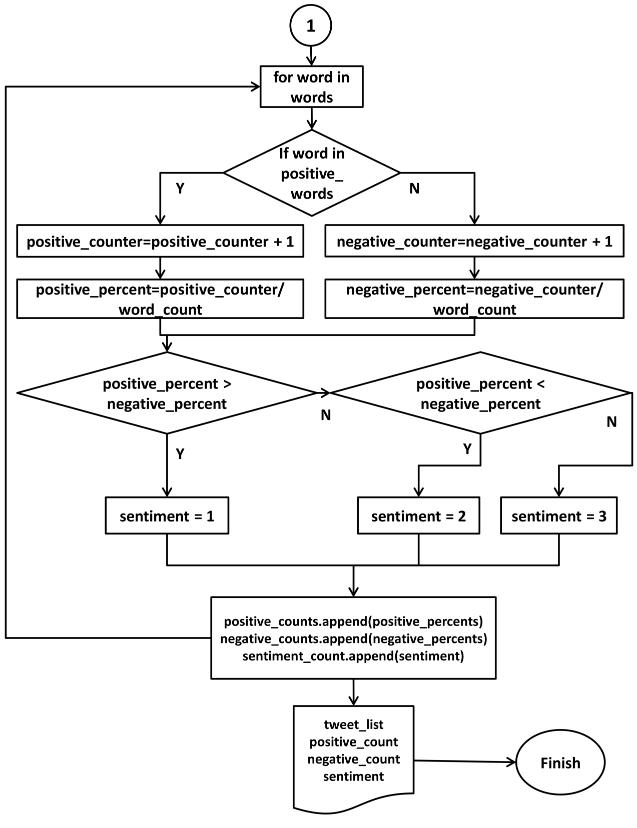

1
for word in words
If word in positive_ words
Y
N
positive_counter=positive_counter + 1
negative_counter=negative_counter + 1
positive_percent=positive_counter/ word_count
negative_percent=negative_counter/ word_count
positive_percent > negative_percent
N
positive_percent < negative_percent
N
Y
Y
sentiment = 1
sentiment = 2
sentiment = 3
positive_counts.append(positive_percents)
negative_counts.append(negative_percents)
sentiment_count.append(sentiment)
tweet_list
positive_count
negative_count
sentiment
Finish

Fig. 9. Lexicon-based Flowchart (continue)

C. Processing

LSTM and MLR algorithms are applied. Tweets that have been cleaned and classified later are included in the classification method to be built. Classification methods to be built are LSTM and MLR. The accuracy results generated by the two classification methods are then calculated using the confusion-matrix method and then compared, so it can be determined which method can produce better accuracy values.

D. Visualization

The results of the information patterns found will be visualized and displayed in a form that is easier to understand, i.e. in the form of diagrams and tables.

E. Report Preparation

Writing and documenting the research starting from the initial stage, which is taking tweets to the results of sentiment analysis and visualizing sentiment analysis data into tables, diagrams, and wordcloud

IV. Implementation and Results

Data obtained from Twitter by using access to the Twitter API that has undergone a pre-processing stage, namely cleaning and through the classification stage of the tweet using the Lexicon-based method, will be the datasets. It is processed at the processing stage. In this stage it is made classification methods by applying deep learning with LSTM algorithm and Machine-learning with MLR algorithm. In the final stage, the comparison of the two classification methods is visualized in the form of diagrams and tables.

A. Implementation of LSTM Classification Method

The built LSTM classification method is explained at this stage. This stage discusses briefly how to build a LSTM model from starting loading datasets to testing and visualization, as seen in Figure 10.

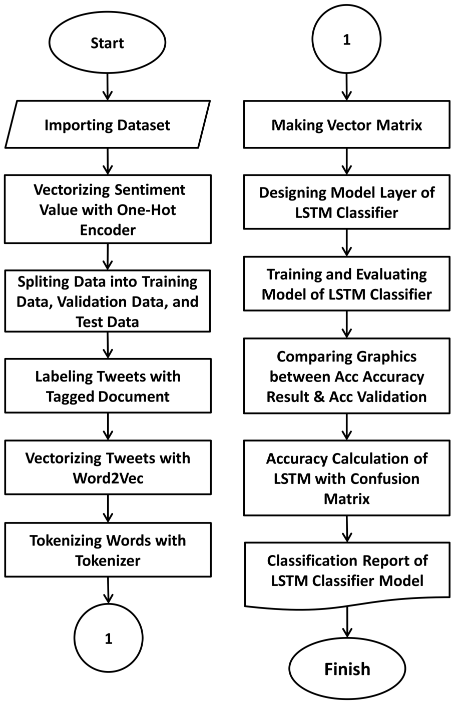

Start
Importing Dataset
Vectorizing Sentiment Value with One-Hot Encoder
Spliting Data into Training Data, Validation Data, and Test Data
Labeling Tweets with Tagged Document
Vectorizing Tweets with Word2Vec
Tokenizing Words with Tokenizer
1
1
Making Vector Matrix
Designing Model Layer of LSTM Classifier
Training and Evaluating Model of LSTM Classifier
Comparing Graphics between Acc Accuracy Result & Acc Validation
Accuracy Calculation of LSTM with Confusion Matrix
Classification Report of LSTM Classifier Model
Finish

Fig. 10. Implementation Chart of the LSTM Classification Method

1) Designing a Layer on the LSTM Classification Method: The built deep learning model has an input layer, hidden layer, and output layer. The input layer contains input data in the form of a matrix vector which is named the embedding-matrix variable, the hidden layer is an additional layer that is useful to train data repeatedly until it gets optimal accuracy or results, the output layer is the result of processing from the hidden layer. The following is the layer design on the LSTM classification method.

TABLE I. LSTM Design Layer

Layer (Type)	Output Shape	Param #
embedding_1 (Embedding)	embedding_1 (Embedding)	4584400
dropout_1 (Dropout)	(None, 35, 200)	0
lstm_1 (LSTM)	(None, 128)	168448
dense_1 (Dense)	(None, 64)	8256
dropout_2 (Dropout)	(None, 64)	0
activation_1 (Activation)	(None, 64)	0
dense_2 (Dense)	(None, 64)	4160
dropout_2 (Dropout)	(None, 64)	0
activation_2 (Activation)	(None, 64)	0
dense_3 (Dense)	(None, 3)	195
activation_3 (Activation)	(None, 3)	0

Based on the design of the LSTM classification method layer in table I, the following layer design can be specified:

1) The input layer is an embedding matrix with a number of matrix vectors of 22,922 with a vector length of 35.
2) Hidden layer 1 has 64 neurons, uses the ReLU activation function, and the dropout is 0.4.
3) Hidden layer 2 has 64 neurons, uses the ReLU activation function, and the dropout is 0.4.
4) Output Layer has 3 neurons, using the Softmax activation function.

The Figure 11 shows a graph of the workflow model of the LSTM model built:

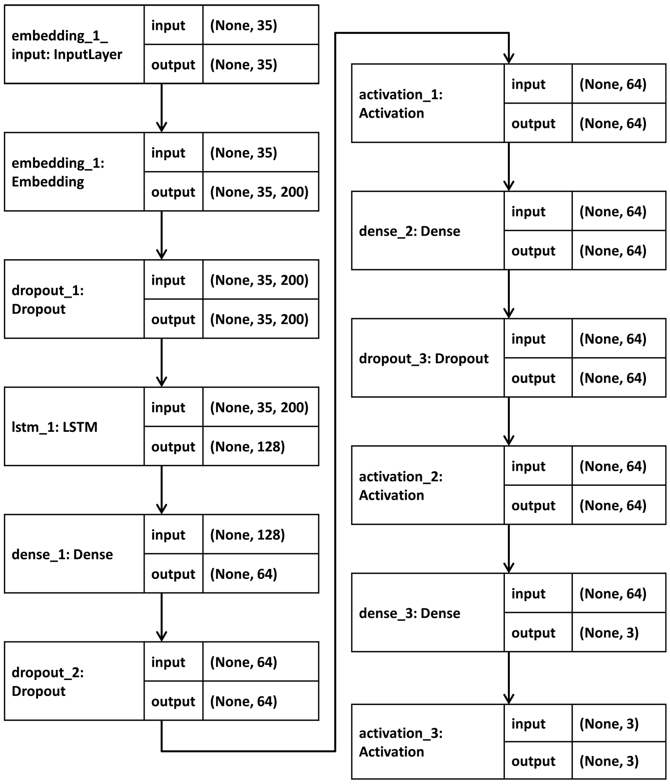


Fig. 11. LSTM Model Workflow

2) Training and Evaluation of the LSTM Classification Method: After declaration of input layer, hidden layer, and output layer, the next stage is the training data process. The training process is carried out as many as 3 epochs and the results of the training model will be stored every time there is an increase in value on the accuracy produced by each epoch. From the results of the training conducted with 3 epochs, it was found that the greatest accuracy value was produced at the 3rd epoch with an accuracy value of 0.9365 or 94 %. Visualization of the results of accuracy and loss values during the training process can be seen in Figure 12.

3) Calculation of LSTM Accuracy with Confusion-Matrix: Accuracy calculations are obtained using the confusion-matrix method. After the data training is done through the learning process, and the test data is evaluated using validation data, the next step is to enter the LSTM model into a new variable to do prediction. Table II shows the results of accuracy calculation using the confusion-matrix method in the form of classification report.

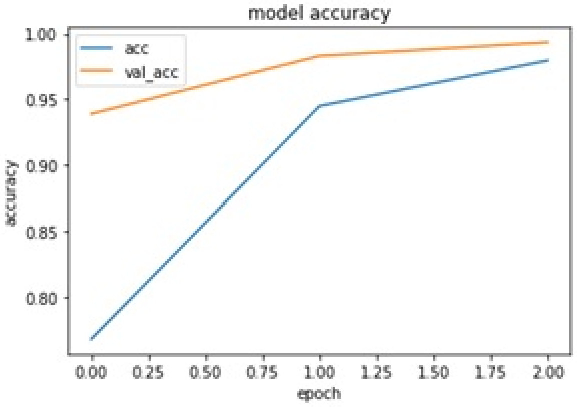


Fig. 12. Graph Comparison of Validation Accuracy against Train Accuracy

TABLE II. ACCURACY RESULT OF LSTM CLASSIFICATION

	Predicted_neg	Predicted_pos	Predicted_net
Negative	213	1	25
Positive	0	192	14
Netral	8	8	312

	Precision	Recall	f1-Score	Support
Negative	0.96	0.89	0.93	239
Positive	0.96	0.93	0.94	206
Netral	0.89	0.95	0.92	328
Avg/Total	**0.93**	**0.93**	**0.93**	**773**

B. Implementation of MLR Classification Method

The built machine-learning classification method with MLR algorithm is explained at this stage. It discusses how to build a MLR model from starting loading datasets to testing and visualization, as seen in Figure 13. Following is the implementation chart of the MLR classification method:

1) MLR Classification Training Method: The training phase is carried out on several *N-gram* models to obtain the model with the best accuracy value which will then be evaluated with the Logistic Regression algorithm. *N-gram* is a method for retrieving bits of letter characters of n from a word. *N-gram* has three types of processing models in a sentence, the type of processing includes *Unigram* for separating one word in a sentence, *Bigram* for separating two words in a sentence, and *Trigram* for separating three words in a sentence. Classifier training models are carried out on the *N-gram* model with several conditions, among others, the *Unigram* with stop words model, *Unigram* without stop words, and *Unigram* without custom stop words. Custom stop words are stop words derived from the words that most often appear on the corpus. In table III it is shown the custom stopwords list in this study.

Figure 14 shows a comparison chart of the results of using stopwords, without stopwords and with no custom stopwords.

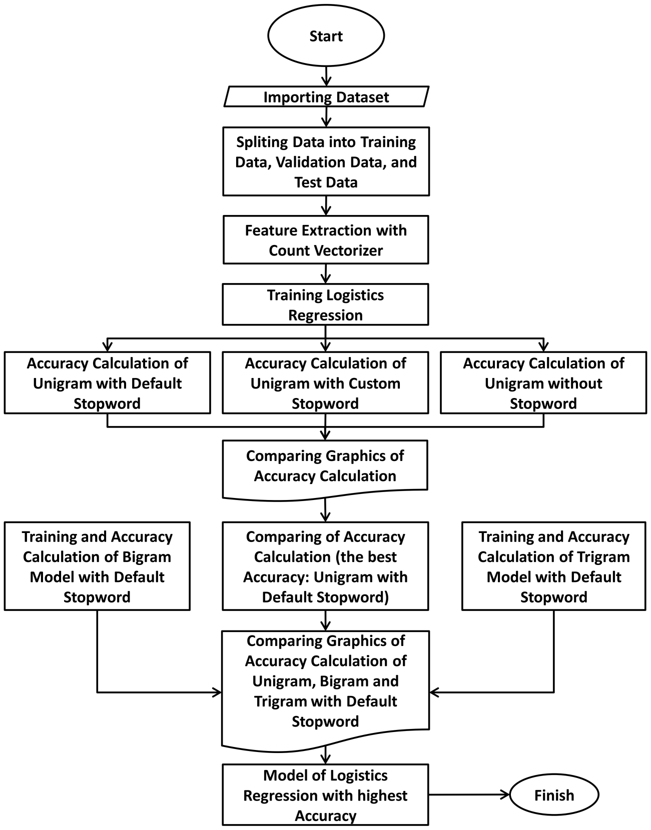

Start
Importing Dataset
Spliting Data into Training Data, Validation Data, and Test Data
Feature Extraction with Count Vectorizer
Training Logistics Regression
Accuracy Calculation of Unigram with Default Stopword
Accuracy Calculation of Unigram with Custom Stopword
Accuracy Calculation of Unigram without Stopword
Comparing Graphics of Accuracy Calculation
Training and Accuracy Calculation of Bigram Model with Default Stopword
Comparing of Accuracy Calculation (the best Accuracy: Unigram with Default Stopword)
Training and Accuracy Calculation of Trigram Model with Default Stopword
Comparing Graphics of Accuracy Calculation of Unigram, Bigram and Trigram with Default Stopword
Model of Logistics Regression with highest Accuracy
Finish

Fig. 13. Implementation Chart of MLR Classification Method

TABLE III. CUSTOM STOPWORDS LIST

	Negative	Positive	Netral	Total
the	17705	10849	7185	35740
to	10448	7522	4239	22209
is	8138	5137	3133	16408
you	8216	4450	3228	15894
and	7655	4677	2671	15003
of	7530	4609	2682	14921
trump	1673	10453	1629	13755
in	5472	3322	1974	10768
not	4613	3298	2119	10030
that	4322	3155	1716	9193
it	4069	2820	1718	8607
this	4260	2342	1765	8367
are	4434	2199	1674	8307
for	3424	2793	1352	7569
he	3386	2523	1409	7318
on	2818	1908	1104	5830
with	2565	1887	996	5448
we	2301	1509	1289	5099
do	2350	1499	1071	4920
your	2466	1201	865	4532

Based on the training process conducted using the *Unigram* word processing model with default stopwords, custom stopwords, and without stopwords, the best accuracy is obtained by 92% with default stopwords. After knowing the greatest accuracy using default stopwords, an experiment is conducted using the word processing *Bigram* and *Trigram*.

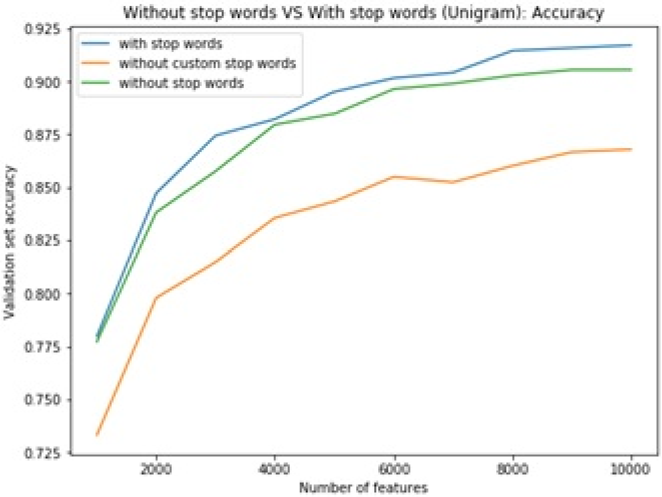


Fig. 14. Comparison Stopwords Graph

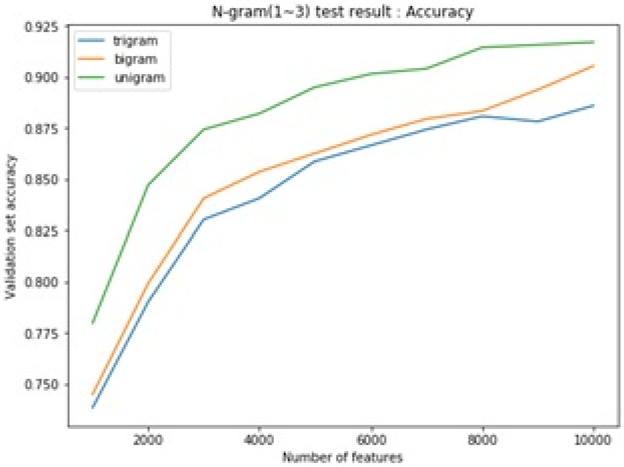


Fig. 15. Comparison N-gram Graph

Figure 15 shows the results of accuracy comparison with *Unigram*, *Bigram*, and *Trigram*. Based on the results of the accuracy obtained from the *Unigram*, *Bigram*, and *Trigram* with stopwords training and validation processes, the best accuracy values for each N-Gram model are shown in the following tabel IV:

TABLE IV. ACCURACY VALUES FOR EACH N-GRAM MODEL

	Number of Feature	Validation Accuracy (%)
Unigram	10000	92
Bigram	10000	91
Trigram	10000	89

From the training results the best accuracy of MLR classification method can be obtained through the *Unigram* word processing method with default stopwords with 92% accuracy.

2) Calculation of MLR Accuracy using Confusion-Matrix: Sentiment analysis carried out in this study is comparing the results of accuracy obtained from the method of deep learning classification using the LSTM algorithm and machine-learning classification method using MLR algorithm. Accuracy calculations are obtained using the confusion-matrix method. The table V shows the results of accuracy calculations with *Unigram* and default stopwords using the confusion-matrix method with 10,000 features in the form of classification reports:

TABLE V. ACCURACY RESULT OF MLR CLASSIFICATION

	Predicted_neg	**Predicted_pos**	**Predicted_net**	
Negative	213	0	16	
Positive	0	175	15	
Netral	15	18	320	

	Precision	**Recall**	**f1-Score**	**Support**
Negative	0.93	0.93	0.93	229
Positive	0.91	0.92	0.91	190
Netral	0.91	0.91	0.91	353
Avg/Total	**0.92**	**0.92**	**0.92**	**772**

C. Comparison of Accuracy Results for the LSTM and MLR Classification Methods

The final result in this study is to determine which classification method is better in conducting sentiment analysis. Based on the classification report obtained it can be concluded that the Deep Learning classification method with the LSTM algorithm is better in analyzing sentiments compared to the MLR classification method using *Unigram* with default stopwords. The following table VI shows the comparison results of the two classification methods tested.

TABLE VI. COMPARISON OF LSTM VS MLR CLASSIFICATION FOR PRECISION, RECALL, AND F1-SCORE

	Precision		**Recall**		**f1-Score**	
	MLR	LSTM	MLR	LSTM	MLR	LSTM
Negative	0.93	0.96	0.93	0.89	0.93	0.93
Positive	0.91	0.96	0.92	0.93	0.92	0.94
Netral	0.91	0.89	0.91	0.95	0.91	0.92
Total	0.92	**0.93**	0.92	**0.93**	0.92	**0.93**

D. Testing of LSTM and MLR Classification Methods

This test is carried out by providing input data on the two classification methods that have been built and have completed the learning process with training data. For example, the author will enter three sentences that later the LSTM and MLR classification method will provide the results of classification of sentiments. The following table VII shows three sentences that will be used as testing material:

TABLE VII. EXAMPLES OF SENTENCE TO TEST THE CLASSIFICATION METHOD

Sentence 1	12 russian programmers have hack the US presidential election system
Sentence 2	Trump won the American presidential election
Sentence 3	bob mueller has given him the information he needs to hold putin accountable for

The sentiment analysis results are stored in the variable 'LSTMpredic' in the array for the LSTM classification method and the 'MLRpredic' variable for the MLR classification method. Sentiment analysis results are stored in the 'LSTM-predic' variable. The following table VIII shows the result of sentiment classification using the LSTM classification method.

TABLE VIII. TEST RESULTS OF LSTM CLASSIFICATION METHOD FROM VARIABLE LSTMPREDIC

	0	1	2
0	**0.925665**	0.000798	0.073537
1	0.000000	**0.999739**	0.000260
2	0.004075	0.002234	**0.993691**

TABLE IX. TEST RESULTS OF MLR CLASSIFICATION METHOD FROM VARIABLE MLRPREDIC

	0	1	2
0	**0.707238**	0.025627	0.267135
1	0.000000	**0.923628**	0.076312
2	0.018730	0.027411	**0.953859**

Table IX shows the results of the sentiment classification obtained using the MLR classification method. The best accuracy results obtained from the MLR classification method are using the *Unigram* word processing model with default stopwords and 10,000 features. The test was carried out by entering the same three sentences as the tests performed on the LSTM classification model. This was done so that the accuracy of the resulting sentiment classification could be compared.

E. Comparison Results of Sentiments on Trial Data

Comparison of classification results can be seen in table X as follows:

TABLE X. COMPARISON OF LSTM AND MLR CLASSIFICATION RESULTS

	Negative(0)		**Positive(1)**		**Netral(2)**		
Sent.	**LSTM**	**MLR**	**LSTM**	**MLR**	**LSTM**	**MLR**	**Cl.**
1	**0.926**	**0.707**	0.001	0.026	0.074	0.267	0
2	0.000	0.000	**0.999**	**0.924**	0.0002	0.976	1
3	0.004	0.018	0.002	0.027	**0.994**	**0.954**	2

The conclusions obtained based on the table on the comparison of LSTM and MLR classification results are as follows:

1) The sentence "12 Russian programmers have hacked the US presidential election system" contains negative sentiment values with a weight of 0.926 (92%) for the LSTM classification method and a weight of 0.707 (70%) for the MLR classification method.
2) The sentence "Trump won the American presidential election" contains a positive sentiment value with a weight of 0.999 (99%) for the LSTM classification method and a weight of 0.924 (92%) for the MLR classification method.
3) The sentence "bob mueller has given him the information he needs to hold accountable for" contains a neutral sentiment value with a weight of 0.994 (99%) for the LSTM classification method and a weight of 0.954 (95%) for the MLR classification model.

4) Based on the sentiment class, the LSTM classification method and MLR are able to produce the same sentiment class, but the weighting of the sentiment value from the LSTM classification model is better.
5) The LSTM classification method is better than the MLR classification method in classifying sentiment classes.

V. Conclusion and Future Work

In this study sentiment analysis was carried out on the topic "Russian Hacking Cases Regarding the 2016 US Presidential Election". Sentiment analysis was carried out by testing the classification method of deep learning using LSTM algorithm and Machine-learning using MLR algorithm. Based on the results obtained from testing the LSTM- and MLR-classification method, the following conclusions can be drawn: Testing of the Deep-learning classification method using the LSTM algorithm produces an accuracy value of 93%. Testing of the Machine-learning classification method using the MLR algorithm produces an accuracy value of 92%. Deep-learning classification method using LSTM algorithm is better in doing sentiment analysis than Machine-learning classification method using MLR algorithm. Differences in accuracy generated by the classification method of LSTM Deep-learning and MLR Machine-learning are significant enough.

The sentiment analysis carried out in this study still has many shortcomings. For that it is necessary to develop the improvement of the application that has been made in this study. The suggestions of application development are as follows:

1) The research was conducted using only two classification methods, namely LSTM deep learning algorithm and MLR machine-learning algorithm. It is expected that the other classification method will be added so that more classification methods can be compared in the it's accuracy results.
2) The dataset used is still relatively small for data processing with deep learning algorithms. For further development, the dataset used is expected to be even more so that the resulting accuracy can be better.

Acknowledgment

The Authors gratefully acknowledge Gunadarma University for providing research funding and for permission in using the research facilities.

References

[1] B. Agarwal and N. Mittal, *Prominent Feature Extraction for Sentiment Analysis*. Springer, 2016.

[2] J. Bayer and C. Osendorfer, "Learning stochastic recurrent networks," in *NIPS 2014 Workshop on Advances in Variational Inference*, 2014.

[3] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, Eds., *A Practical Guide to Sentiment Analysis*. Springer, 2017.

[4] A. Chaudhuri, *Visual and Text Sentiment Analysis through Hierarchical Deep Learning Networks*. Springer, 2019.

[5] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," Dec. 2014. [Online]. Available: https://arxiv.org/pdf/1412.3555.pdf

[6] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengioy, "A recurrent latent variable model for sequential data," Jun. 2015. [Online]. Available: arxiv.org/pdf/1506.02216v3.pdf

[7] M. Dehaff, "Sentiment analysis, hard but worth it!" on-line, 2010. [Online]. Available: http://www.customerthink.com/blog/sentiment_analysis_hard_but_worth_it/

[8] F. Gers and E. Schmidhuber, "Lstm recurrent networks learn simple context-free and context-sensitive languages," *IEEE Transactions on Neural Networks*, vol. 12, no. 6, pp. 1333 – 1340, Nov. 2001.

[9] F. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 2000), Neural Computing: New Challenges and Perspectives for the New Millennium*. IEEE, Jul. 2000.

[10] A. Hepburn, "Infographic: Twitter statistics, facts & figures," on-line, 2010. [Online]. Available: http://www.digitalbuzzblog.com/infographic-twitter-statistics-facts-figures/

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[12] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. John Wiley & Sons, 2013.

[13] N. Kalchbrenner, I. Danihelka, and A. Graves, "Grid long short-term memory," Jul. 2015. [Online]. Available: arxiv.org/pdf/1507.01526v1.pdf

[14] O. Kolchyna, T. T. P. Souza, P. C. Treleaven, and T. Aste, "Twitter sentiment analysis: Lexicon method, machine learning method and their combination," On-Line, Sep. 2015. [Online]. Available: https://arxiv.org/abs/1507.00955

[15] B. Liu, *In Handbook of natural language processing*, 2nd ed., 2010, vol. 2, ch. Sentiment Analysis and Subjectivity, pp. 627–666.

[16] ——, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015.

[17] M. M. Nasr, E. M. Shaaban, and A. M. Hafez, "Building sentiment analysis model using graphlab," *International Journal of Scientific & Engineering Research*, vol. 8, no. 6, pp. 1155–1158, Jun. 2017.

[18] C. Olah, "Understanding lstm networks," on-line, Aug. 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[19] S. Poria, A. Hussain, and E. Cambria, *Multimodal Sentiment Analysis*. Springer, 2018.

[20] L. Vu and T. Le, "A lexicon-based method for sentiment analysis using social network data," in *Int'l Conf. Information and Knowledge Engineering*, 2017.

[21] D. Zarella, *The Social Media Marketing Book*. PT Serambi Ilmu Semesta Anggota IKAPI, Jakarta, 2010.

A Review of Intelligent Tutorial Systems in Computer and Web based Education

Luis Alfaro[1], Claudia Rivera[2], Elisa Castañeda[3], Jesus Zuñiga-Cueva[4], María Rivera-Chavez[5], Francisco Fialho[6]
Universidad Nacional de San Agustín de Arequipa, Arequipa - Perú[1,2,3,4,5]
Universidad Católica de Santa María, Arequipa - Perú[6]
Universidad Federal de Santa Catarina, Florianopolis-SC - Brasil[7]

***Abstract*—ITS (Intelligent Tutoring Systems) are integrated and complex systems, designed and developed using approaches and methods of artificial intelligence (AI), for the resolution of problems and requirements of the teaching/learning activities in the field of education and training of students and the workforce based in computers an web based emerging resources. These systems can establish the level of student knowledge and the learning strategies used to improve the level of knowledge to support the detection and correction of student misconceptions. Their purpose is to contribute to the process of teaching and learning in a given area of knowledge, respecting the individuality of the student. In this paper, a review of intelligent tutorial systems (ITS) is presented, from the perspective of their application and usability in modern learning concepts. The methodology used was that of bibliographical review of classic works of the printed and digital literature in relation to ITS and e-Learning systems, as well as searches in diverse databases, of theses and works in universities and digital repositories. The main weakness of the research lies in the fact that the search was limited to documents published in the English, Spanish and Portuguese.**

Keywords*—Intelligent learning systems; computer assisted learning environments; web based education

I. Introduction

The various investigations into the design and development of Intelligent Environments for computer-aided design/learning, also known as ITS, were initiated approximately in the 1970s. Some authors, such as Carbonell [1], tried to combine methods of Artificial Intelligence (AI) with Computer Aided Instruction (CAI), to propose a system that would try to create an environment that would take into consideration the diversity of the various learning styles of the students, thus adapting to the individual requirements of those who would use the system. This type of software was called Intelligent Tutor or Intelligent Tutorial Systems. IACs and ITS are apparently in a similar type of application in education.

The study of ITS is an area of research on which a large number of researchers focused, working on topics that have strong relationships with various disciplines. Researchers interested in getting started in this field may have difficulty understanding the basics and methodologies used in ITS. These difficulties include understanding the functioning of ITS and their components, their functions, the main types of ITS, the Artificial Intelligence technologies involved, learning theories and their uses, differences in terms of interaction and behaviour, the importance and contributions of ITS in education, and their effectiveness.

This paper does not aim to focus on a specific topic or dimension of ITS, nor to visualize details that may require further study. After understanding the main concepts and ITS and their behaviors, a reader can review widely detailed and conventional sources such as Woolf [2], Ma et al. [3], Nwana [4], Shute et al. [5], and Murray [6], among other authors for further research. In Section II, the fundamentals of Intelligent Tutorial Systems are described; in Section III, a review of developed intelligent tutoring systems (ITS) in education is made, in Section IV, some developments are described and analyzed, in Section V, the discussion is made, and finally, in Section VI, the conclusions are established and the corresponding comments are made.

II. Intelligent Tutorial Systems

An ITS can be described as software that involves [3], [7]:

- A computer that encodes pedagogical domains and human teacher knowledge (trainer) as a good mechanism to communicate with other humans;
- A trainee who interacts with a computer to acquire some skills in those domains.

Burns [8], emphasizes that ITS research, especially considering teaching-learning theories, should address teaching strategies, taking into consideration individual differences. This research includes well-known works referred to by P. Lach [9], such as Clance's Guidon and his later reviews, Soloway and Johnson's; PROUST, Anderson and Boyle; ACT tutors, as well as Dillenbourg ETOILE [10] and many others, which proposed the characteristics of ITS and their abilities to diagnose misconceptions of the learner during the teaching process and, based on that diagnosis, provide subsidiary teaching to students. However, many psychological issues underlying learning, teaching, and understanding have not been convincingly answered. In addition, there are enormous difficulties in accurately representing the stages of student learning and in identifying possible misconceptions, facts that contributed towards a diversification in ITS-related research.

For Dede [11], the Intelligent Tutoring Systems and Trainers, also called Intelligent Computer-Aided Instruction (ICAI), provide educational technology with the characteristics of the teacher's cognitive skills. The strategies focused on these types of educational applications are based on ideas from the field of Artificial Intelligence. ITS/ICAI applications ideally contain dynamic models of the learner in which knowledge can be communicated and discussed pedagogically. These systems establish who, what and how to teach. In a full-fledged ITS, the material to be presented to the learner is interactively treated by these dynamic models, generated by the system in real time.

A. ITS Architectures and Subsystems

ITS architectures vary from one implementation to another. Numerous systems were implemented in the mid-1970s and 1980s by Clancey in 1979, Johnson in 1985, Anderson in 1985, and Viccari in 1990, among others [12]. The architecture developed during this phase proposed that an ITS/ICAI system should include the following functional elements, described from an analytical and critical perspective.

- An explicit domain model and an expert system capable of solving problems in that domain.
- An identification model with some detail about the student's knowledge of the domain.
- A teaching model that offers instruction, presenting instructional material, capable of detecting and assisting in the resolution of learners' misconceptions [13].

The structure of a system with the above characteristics requires the development of the tutorial system composed of the modules shown in Figure 1.

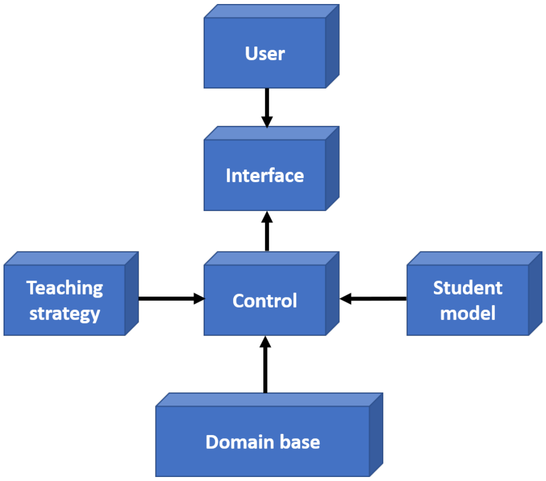


Fig. 1. General architecture of a ITS. Source: adapted from [9]

In the most recent conceptualizations of ITS [14], the emphasis on student modeling as the key to adaptive tutoring remains evident. Based on these papers and a review of the literature on ITS assessments, the following definition is adopted [3] [14]: An ITS is a computer system that for each student performs tutoring with various functions that:

- Present information for learning,
- Ask questions or assign tasks for learning,
- Make comments and/or suggestions,
- Answer questions asked by students,
- Establishes indications to provoke a cognitive, motivational or metacognitive change.

By establishing inferences from the student's responses, a persistent multidimensional model of the student's psychological states, such as subject knowledge, learning strategies, motivations or emotions, is constructed or the student's current psychological domain is established in a multidimensional domain model. The student-modeling functions identified in the previous point are used to make adaptations to one or more of the tutoring functions identified in the first point.

The different components of ITS architecture will be described below.

1) Knowledge domain base: The domain model [15], is the expert component of the tutor, the specialist who treats and manipulates the knowledge of the subject [16], constituted by the instructional material, by a systematic generation of examples, by the formulation of diagnoses and by the simulation processes. It contains the knowledge about the domain that we want to teach the student [4] [17].

This module would have the function of allowing the alteration/enlargement/adaptation of the tutor's two main knowledge-based components: the domain base and the student model. The inclusion of learning capabilities in the tutorial system [18], implies an architectural style that is not supported by bases of knowledge but based on beliefs.

Several knowledge representation models can be used in this module, among them: Semantic Networks, frames, scripts, production rules, Object Oriented Programming - OOP, among others. The selection should focus on the method that best and most easily meets the requirements of representation and manipulation of reasoning. Inadequate selection may compromise the system's performance, since this module must be able to determine, among other things, the complexity and consequently the way of presenting the concepts of the knowledge area in question. Instructional material is generally organized in a taxonomy that provides for increasing levels of complexity. The tasks are organized using a dynamically formed tree structure, according to the student's interaction and work [19]. The teaching strategy to be used will depend on the model of the defined student.

The domain knowledge base [20], is a key component of ITS, where the instructional material, i.e. the content that the tutor must provide, is represented. The fact that this content is stored in a knowledge base, and/or in a conventional database, is one of the factors that determine the difference between an ITS and a conventional CAI [21]. The knowledge base should enable the system to reason about the structure of the content to be provided, thus allowing it to take a more active role than presenting content in a linear fashion.

2) Student's model: This is the expert in teaching techniques [16], that selects the concepts, sets the levels of difficulty of the teaching activity and controls the level of the learning process. The student's model [12], contains the relevant information - from the tutor's point of view - about the student. It is the presence of this model that allows the tutor system to adapt to each student, customizing the instruction. This module represents the knowledge and cognitive skills of the student at a given moment.

It is constituted by static and dynamic data that will be of fundamental importance, so that the tutor can check the hypothesis with respect to the student [9]. It contains a representation of the student's state of knowledge at the moment that they interact with the ITS. From this model, the student manages the content to be taught and the system must be able to infer the best teaching strategy to be used next. A real model of the student implies a dynamic update as the system evaluates the student's performance. The dynamic data refers to the performance that the student has in relation to the questions formulated by the tutor and confronted with the hypotheses elaborated by the student, towards the use that the student makes of the system and towards the new knowledge that the student can receive from the tutor's teaching.

Some of the techniques that are used to build the student's model [19]:

- Include pattern recognition applied to the history of the responses provided;
- Compare the student's behavior with that of a specialist and verify the common points;
- Consider the student's preferences;
- Consider the student's particular goals;
- Things that tend to be forgotten when interacting with the tutor;
- Indication of their particular objectives.

The student model, according to Costa [22], can be represented using some description models, namely:

- Differential model: Where the student's response is compared to the knowledge base.
- Overlay model [19] The student's knowledge is represented as a subset of the tutor system's knowledge base [23] [24]. This model assumes (implicitly or explicitly) that the student's errors or abnormal behaviour are always due to the absence of some information present in the knowledge base of the domain. This psychological assumption is overly simplistic, as incorrect behaviour originates from the presence of misconceptions in the student's mind.
- The disturbance model or BUGGY model: also relates the model of the student in the knowledge base of the domain [25] [26]. The disturbance model assumes that the student's errors are derived from the misconception of some concept or lack thereof. For this case, there is the domain base and a typical error library; the student model includes elements of the domain base and error library as shown in Fig. 2.
- Simulation Model: The environment has a model of how the student can or should behave in a certain situation and through this model it is possible to foresee the future behavior of the student; in other words, the response of the student based on their behavior during the work session.
- Belief Model [27] consists of a set of beliefs that reflect the degree of understanding of the student regarding a particular concept. According to [19], it should refer to belief bases rather than knowledge bases, since the logical behavior of the student's conceptions is much more like a belief logic than a knowledge logic [28].

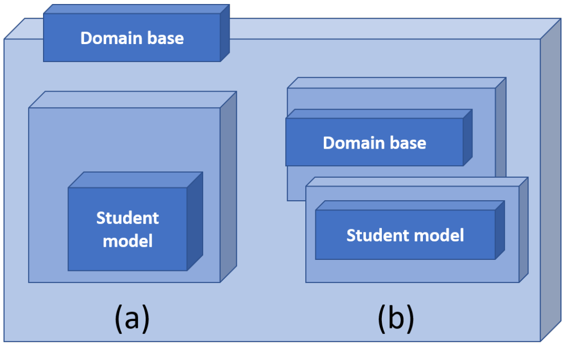


Fig. 2. Overlay Model. Source: adapted from [9]

- Agent Model: When treating the student model as a belief system [9], an important notion is implicitly used: the interaction between the student and the tutor system is an interaction between two intelligent agents (or, at least, endowed with some cognitive behavior). Considering the student as an agent, implies considering the student's model as an agent model or that it will have consequences for the model's structure.
- Constraint Based Model (CBM). This was proposed for the first time, for the modelling of students in the short term and the diagnosis of the current state of the solution. CBM uses constraints to present student mastery and knowledge [23]. The process of diagnosing the student's solution is by matching the relevant conditions of all the constraints with the student's solutions. The condition of satisfaction for all relevance conditions must also be met. The system checks every step taken by the student, diagnoses any problems, and provides feedback to the student when there is an error. The feedback informs the student that the solution is wrong, indicating the part of the solution that is wrong, and then specifies the principle of mastery that has been violated [29].
  Cognitive theories. As many researchers claim, the use of Human Plausible Reasoning (HPR) and Multiple Attribute Decision-Making (MADM) Theory for the purpose of student modeling [23] and diagnosing misconceptions, leads to the design and development of effective ITS. The cognitive theory approach helps to interpret human behavior during the learning process by trying to understand human thought and comprehension processes.
- Bayesian Networks - BN: Another well-known and established approach to representing and reasoning about uncertainty in student models is Bayesian Networks. A Bayesian network (BN) is a cyclical, directed graph containing random variables, which are represented as the network. Probabilistic relationships between variables are presented as arcs. The BN reasons about the situation it models, analyzing sequences

of actions, observations, consequences and expected outcomes [30]. With respect to the learner model, learner components such as knowledge, misconceptions, emotions, learning styles, motivation, and goals can be represented as nodes in the BN.

- Fuzzy student modeling. The student's level of knowledge is established as a fuzzy problem. One possible approach to dealing with uncertainty is fuzzy logic, introduced by Zadeh in 1956 as a methodology for computing and reasoning with words representing imprecise values rather than numerical values. Fuzzy logic is used to deal with the uncertainty of real-world problems derived from inaccurate and incomplete data, as well as from human subjectivity. Fuzzy logic uses fuzzy sets that involve variables with uncertain values. The use of fuzzy logic can improve the learning environment by allowing intelligent decisions about the learning content to be delivered to the learner, as well as personalized feedback to be given to each learner. It is a fuzzy logic channel to diagnose the level of knowledge of the student in a concept and to predict the level of knowledge for other concepts related to that concept [31]. Some authors argue that the integration of fuzzy logic in the student's model increases student satisfaction and performance, improves the adaptability of the system and contributes to more reliable decision-making. The use of fuzzy logic in student modeling is becoming popular because it overcomes computational complexity, imitating human nature.

Finally, in 1994 Nwana [4], classified the students models into the following six different types:

- Corrective in which the elimination of misconceptions in the student's knowledge is allowed.
- Collaborative, in the student's report.
- A tool that helps the tutorials to adopt the action and the performance of the student.
- Diagnostic, which identifies the errors in the most recent student's knowledge.
- Predictive, which helps to understand the answers of the most recent client's evaluation system.
- The final objective evaluates the value of the student's overall progress.

The strategies constitute the knowledge on how to teach; in other words, how to generate from the diagnostic information, monitoring and analysis, a sequence of teaching tactics, capable of successfully presenting a certain topic to a certain student. According to Breuker [19], most authors agree that a teaching strategy must question:

- When to interrupt? What reasons justify interrupting the student's course of learning?
- What to say? This question is divided into:
  - selection of the topic(s) to be presented
  - ordering of the topics, if there is more than one
- How to say it? This is probably the most difficult question. No concrete general solutions have been proposed, and many authors here point out the lack of sufficiently detailed pedagogical theories.

3) Control model: The control module manages the operation of the tutor system. Its execution cycle can be characterized as follows [13]:

- Selection of a teaching strategy from the strategy bank;
- Based on the teaching strategy, select an instructional material from the domain's knowledge base;
- Present the material to the student through the interface module (which may include the presentation of exercises and solution of proposed exercises)

From the student's responses, diagnose their behavior and monitor their progress, reading/updating the student's model and restarting the cycle. This is the module responsible for the general coordination of the tutor regarding the functions, natural language interfaces, information exchange between modules and communication with other utility programs through the operating system.

Communication between the tutor modules consists of saving or reading files, keeping a historical file of the learning session, activating and deactivating the databases which can be conceived as "worlds" created from the interaction between the tutor and the student.

The ICAIs have a learning capacity in what refers to the alterations made in the tutor's role, resulting from the interaction process with the student. In some tutors, the initial core is not altered at the end of the session, restarting in the same way for any new user, while in other more refined models, each interaction or initial database is altered, so that the system evolves learning with each user and applying this new knowledge to each student.

4) Interface: This is the expert who interprets natural language [16]. It is established that a good interface is vital for the success of any interactive system and ITS are no exception. On the contrary, it can be said that the quality of the interaction grows in importance in this kind of system, because it is during the interaction that the tutor system exerts two of its main functions [13]:

- The presentation of the instructional material;
- The monitoring of the student's progress through the reception of the student's response.

From these two functions, some objectives can be derived to be fulfilled by the interface module [19]:

- It is necessary to avoid that the student perceive that the session is tedious; that is, the wealth of resources in the presentation of the instructional material is necessary;
- It is desirable that there be facilities for change in the dialogue initiatives: the student must be able to intervene easily in the tutor's discourse, and vice versa;

- The response time must obviously remain within acceptable limits;
- Monitoring must be carried out as far as possible in the background, so as not to burden the student with questionnaires that could increase their workload, and also respecting the barrier of response time.

Since the incorporation to the systems of the resources offered by hypertext and hypermedia, the possibilities of quality in the presentation of the instructional material have received significant progress, being improved even further, with the incorporation of the emerging resources associated to technologies based on the Web. The variety of resources, associated with the possibility of reviewing the material in a way linked to the semantics of the content, makes hypermedia systems a high potential tool for the presentation of instructional material in ITS.

Monitoring of student progress occurs at two levels:

- At the level of historical student analysis, that is, from one session to the next.
- At the level of diagnosis limited to one session.

III. Current development in ITSs in Computer and Web based Education

Over the past few years, several effective and successful ITS have been proposed and mentioned in various parts of this paper. In this section, a review will be made of the different areas of research on which the research is focusing.

A. Intelligent Adaptive Educational Systems

Adaptive Intelligent Web-Based Systems (AIWBES) or adaptive hypermedia, are an alternative to the traditional approach that only provides the development of web-based educational courses. These systems offer a high degree of adaptability in terms of objectives, preferences, learning styles [32] [33] [34], and individual student knowledge during interaction with the system [35] [36] [37].

In the area of research, the first research was focused on adaptive educational hypermedia, which is the result of the incorporation of resources that combine educational hypermedia with ITS, providing greater flexibility and functionality than traditional static educational hypermedia [38]. Several systems have been developed under the AIWBES category, focusing on distance learning not only to provide support with textbook course material, but also to provide elements for problem solving. Adaptive navigation through the material was implemented to support individual student learning. An important attribute is that the system classifies the content of a page so that it is ready to be learned, or even if it is not ready to be learned because the prerequisites have not yet been worked with [39]. Additionally, links are ordered according to relevance to the current state of the student, so that students know which situations are most similar or which web pages are most relevant. When the student enters a page that contains a portion of prerequisite and knowledge to be learned, the system alerts the student to the prerequisite and suggests additional links to the textbook and textbook pages about them. Empirical studies have shown that hypermedia systems in conjunction with tutoring tools can be useful in supporting self-learning [40]. Some intelligent adaptive hypermedia systems that have been used by many students include Interbook [41] and the AHA! [42], which were designed to help students learn better and in less time [43].

B. Cultural Awareness in Education

In recent years, special attention has been paid to problems arising in the context of education in a globalized society [44]. Researchers are concerned about how learning technology systems can be adapted through cultural diversity. Nyein [45] addressed the barriers faced by ITS in the developing world. Barriers such as lack of computer skills of students, problems arising from multiple languages and cultures, etc. were presented along with existing solutions. Ogan et al. [44], conducted an analysis of learners' help-seeking behaviors regarding ITS in different cultures. Behavioral models of help-seeking behavior during learning have been developed based on data sets of students from different countries: Costa Rica, the Philippines and the United States. M. Mohan [46] takes the first step to avoid focusing on this problem. The root of this study provides students with some control over their cultural preferences, including the description of the problem, feedback, and the presentation of images and suggestions. The deployment of such systems has provided researchers with the opportunity to experiment with the phenomena surrounding the social acceptability of the use of non-dominant language in education, and has effects that should be further investigated.

C. Collaborative Learning

Research in contemporary education suggests that collaborative learning or group learning improves group learning performance as well as individual learning outcomes [47] [48]. In a collaborative learning environment, students learn in groups through interactions among themselves by asking questions, explaining and justifying their opinions, explaining their reasoning, and presenting to their peers topics of their knowledge [49]. Several researchers have pointed out the importance of having a group learning environment and how significant it can be in terms of improving learning [50]. Now-a-days, interesting implementations of collaborative learning in tutoring systems are emerging to show the benefits obtained from the interaction between students during problem solving, such as in the domain of engineering [20], in which teams of two or more students work on the same task when solving a problem. Also, a series of experiments were carried out to investigate the effectiveness of collaborative learning and how to involve students more deeply in the conversations of the instructional process with tutors, using teaching techniques such as attention focusing, question and answer and social interaction strategies. Students working in pairs were found to learn better than students working individually [47] [48].

D. ITS with a Playful Focus

ITS and their interactive components can be interesting when used for a short period of time (e.g., a few hours), but can become monotonous and boring or even annoying when a learner needs to interact with ITS for weeks or even months [51]. The idea behind play-based learning is that learners learn best when they find the teaching/learning activities fun by getting involved and participating in the learning process in an active way, and by motivating them to use the system longer [52]. Some researchers argue that the principles of ITS development approaches maximize learning and that play technologies maximize motivation. Instead of learning a subject with a conventional, traditional approach, students play an educational game that successfully integrates game strategies with curriculum-based content. Although there is no overwhelming evidence to support the effectiveness of educational systems that incorporate playful aspects which can take the form of games associated with computer tutors, educational games have been found to have advantages over traditional tutoring approaches [53] [54]. Finally, several studies found a strong relationship between learning outcomes, problem solving in play and increased engagement [55].

IV. Some Developments in ITS's

In this section, various proposals for ITS architectures are reviewed and analyzed, including reviewing the results and evidence reported by the different research groups that proposed them.

Badaracco and Martinez [56], propose an ITS which has the objective of dynamically customising teaching processes according to the profile and activities of the student, using artificial intelligence techniques. In the proposal, the pedagogical model is crucial, because the complexity of the ITS will depend on their scope, which can be defined as specific or generic. This work proposes an architecture that uses a pedagogical model of learning based on competences, with the purpose of managing the complexity and facilitating its understanding, along with a diagnostic process for this system.

H. Al Rekhawi and S. Naser [57], propose the design of a web-based ITS system for teaching application development in an Android environment to help overcome the difficulties they face. The system introduces the topic of application development in Android and manages automatically generated problems for students to solve. In addition, automatic adaptations are made in relation to execution times, considering the individual growth of the student. The system provides support for the realization of constructions through adaptive demonstrations. An initial evaluation was conducted to examine the effect of the use of ITS on the performance of students in the development of applications for smart phones, obtaining results that showed a positive impact on evaluations.

M. Hamed. S. and Abu Naser [58], proposed an ITS system to facilitate the learning of science subjects at school level since, according to these authors, students face some learning problems. The system supports the understanding of these issues through analysis and explanation in a systematic way. The design of the ITS described is focused on the teaching of science for the 7th grade, supporting students in the knowledge of characteristics of living beings, providing in addition all the topics related to living beings and generating some questions for each topic, to which the students must answer correctly, in order to pass to the next level. The authors carried out an evaluation to verify the satisfaction of the students and teachers who use their proposal. The results obtained in relation to the usefulness and usability of the system were satisfactory.

N. Gon, S. Bisw. and L. Ba But [59], report the development of an ITS for Multimedia Virtual Power Laboratory (VPL) which can simulate an electric machine laboratory. In the VPL architecture, which consists of instructional design and implementation of an ITS, the virtual lab is supervised by a virtual Smart Tutor who can track the students' progress and monitor their actions, on the virtual lab platform. The VPL offers a virtual experimental environment with 2D graphics, 3D animations, audio guidance, simulations, knowledge concept bases and virtual experiments, functionalities on which the ITS is designed. It can process user actions and existing resources in the lab, as well as track student progress by answering questions, monitoring their actions and, if necessary, guiding students through the contents of the previous material required for the mastery of the subject.

W. Yu-Ying [60], propose an ITS to support students in improving their English skills. The aim of the system is to provide a tutorial environment where teachers and students do not need to prepare much teaching and learning support material by teaching or learning English in that environment. An interesting element of the proposal is the verification method of the intelligent tutoring system, using Petri dishes. The ITS was developed in an Augmented Reality (AR) environment, a Text to Speech (TTS) and Speech Recognition (SR) system. The system is divided into two parts: one for teachers and one for students. The reported experimental results show that the use of Petri dish networks can support users in the verification of the intelligent tutoring system for better learning performance and correct operation.

The authors W. Ma, O. J. Nesbit, Q. Liu. O and Adesope [3], conducted a meta-analysis regarding research that compared the learning outcomes of students who learned using ITS with those who learned from learning environments that did not include ITS. The meta-analysis examined how the magnitudes of effects varied according to the type of ITS, the type of comparative treatment received by the learners, the learning outcome, and whether the knowledge to be learned was procedural or declarative, as well as other factors. The use of ITS was associated with higher performance compared to teacher-led instruction for large groups of students, non-computerized ITS-based instruction, and textbooks or workbooks. There were no significant differences between learning using ITS and learning through individualized human tutoring or small group instruction. Significant positive average effect sizes were found regardless of whether ITS were used as the primary means of instruction, a supplement to teacher-led instruction, an integral component of teacher-led instruction, or a homework aid. Significant positive effect measures were found, at all levels of education, in almost all subject domains assessed, and whether or not ITS provided erroneous student feedback or models. The research claim made in this paper that ITS are relatively effective tools for learning is consistent with the studies and analyses conducted by these authors.

Likewise, the research carried out evidences that the ITS developed at the moment constitute effective tools for the activities of teaching learning, due to which they count on a high degree of adaptation to the demands and interests of the students, this in part because they consider elements like the styles of learning, the individual differences, the levels of advance in the learning, among other attributes and also to the use of techniques of Artificial Intelligence and emergent resources associated to the Virtual Reality, Augmented Reality and Multimedia, among others.

V. Discussion

ITS were designed with the objective of having a degree of adaptation to the characteristics and special requirements that students demand, individually. The ability to track the individual cognitive states of students in order to provide an appropriate response is what differentiates ITS from other educational systems. Likewise, the ITS focused the attention of researchers from various disciplines such as Psychology, Cognitive Sciences, Educational Sciences and especially Computer Science. These systems aim to achieve the possibility of imitating expert human tutors in the way they teach and interact with students. Throughout the last decades, ITS have demonstrated their pedagogical effectiveness, contributing with the improvement of students' learning results. It is likely that their systems will be useful for adults or children with special needs in the pursuit of their learning goals, as they can be effective and helpful in supporting the teaching of people with cognitive disabilities such as dyslexia, attention deficit disorder and dyscalculia, among others.

Also, many benefits can be obtained from research in this line of investigation, since there are hundreds of ITS software applications for a variety of subject domains, and these applications can be extended to other disciplines.

Research and findings in this direction could increase the popularity of ITS as a new educational tool approach in order to support students in their decisions regarding which majors to choose.

Finally, they can also provide subsidiary support over time to significantly improve students' competencies and skills and prepare them for other stages of their lives.

VI. Conclusion and Comments

It is important to note that despite important developments in ITS, the differences between human tutors and ITS are numerous. The role of ITS is therefore to support learners and human tutors in teaching and learning activities.

Various models were developed for the representation of knowledge, teaching styles and student knowledge. The different models have advantages and disadvantages. The reviews conducted concluded that in some circumstances, the use of ITS allows better results than traditional classroom instruction and study in printed materials.

Hybrid models were designed to improve and strengthen traditional models. There are many unanswered research questions regarding the principles of human thinking and learning, such as those from approaches in Cognitive Biology, Psychology, Educational Sciences, Neurosciences, and others. Many of these approaches, their methodologies and techniques, have been incorporated into ITS, and have been implemented and tested with certain levels of success.

ITS may become a competitive alternative for humans in the future, considering aspects of cost, time and levels of application at scale. ITS promise to standardize and implement aspects of human learning as much as possible, but they still have many limitations to overcome.

The convergence of ITS with AI and psychology in research teams, promises continued progress in the development of ITS research.

Acknowledgment

The authors would like to thank Saint Augustin National University (UNSA) – Arequipa – Peru, Catolic University of Santa María - Peru and the Federal University of Santa Catarina – Florianópolis - Brasil

References

[1] J. Carbonell, *AI in CAI: An artificial intelligence approach to computer assisted instruction*. IEEE Transactions on Man-Machine Systems,vol.11, no.4, pp.190–202, 1970.

[2] B. Woolf, *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-Learning*. Morgan Kaufmann, 2010.

[3] W. Ma and O. Adesope and J. Quing and Q. Liu, *Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis*. Journal of Educational Psychology 2014 American Psychological Association 2014, Vol. 106, No. 4, 901–918.

[4] H.S. Nwana, *Intelligent tutoring systems: an overview*. Artificial Intelligence Review, vol.4, no.4, pp.251–277, 1990.

[5] V. Shute and J. Psotka, *Intelligent tutoring systems: Past, present, andfuture*. DTIC. Document, Tech.Rep., 1994.

[6] T. Murray, *An overview of intelligent tutoring system authoring tools: Updated analysis of the state of the art*. in Authoring tools for advanced technology learning environments. Springer, 2003, pp.491–544.

[7] E. Aimeur and C. Frasson, *Analizing a new Learning strategy according to different konowledge levels*. Computers Education. Vol. 27, No. 2.

[8] L. Alfaro, *Contribuições para a modelagem de um ambiente inteligente de educação baseado em realidade virtual*. Doutorado. Programa de Pós-Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina, Brasil, 1999. Brasil.

[9] L.M.M. Giraffa and M.A. Nunes and R.M. Viccari, *Multi-Ecological: an Intelligent Learning Environment using Multi-Agent architecture*. MASTA'97: Multi-Agent System: Theory and Applications. Proceedings. Coimbra: DE-Universidade de Coimbra, 1997.

[10] T. Dillenbourg, *Intelligent Learning Environments*. Carouge - Switzerland. TECFA (Technologies de Formation et Apprentissage). Faculté de Psychologie et des Sciences de l'Education. University of Geneva (Switzerland), 1993. 34 p.

[11] C. Dede and M. Salzman and C. Loftin and R. Bowen, *ScienceSpace: Virtual realities for learning complex and abstract scientific concepts*. 1995. http://www.virtual.qmu.edu/vriaspdf.htm

[12] L. Giraffa, *Selecting teaching strategies using pedagogical agents*. Porto Alegre, 1998. Proposta de tese (Doutorado em Ciência da Computação. Instituto de Informática) Universidade Federal do Rio Grande do Sul. 77 P.

[13] F. Oliveira, *Critérios de equilibração para sistemas tutores*. Porto Alegre, 1994. Tese (Doutorado em Ciencia da Computação. Instituto de Informática) Universidade Federal do Rio Grande do Sul. Brasil. 68 p.

[14] R. Sottilare and A. Graesser and X. Hu and H. Holden (Eds.), *Design recommendations for Intelligent Tutoring Systems*. Orlando, FL: U.S. Army Research Laboratory. 2013. ISBN: 978-0-9893923-0-3.

[15] R. Mizoguchi, *Student modeling in ITS*. Emerging Technologies in Education, vol.8, pp. 35–48, 1995. 68 p.

[16] M. Corredor, *Sistemas tutoriales inteligentes*. Boletín de Informática Educativa. Colombia. Proyecto SIIE. Vol. 2, Nº 1. 1989.

[17] E. Sierra, *Towards a methodology for the design of intelligent tutoring systems*. Research in Computing Science Journal, vol. 20, pp.181–189, 2006.

[18] R. Nkambou, *Modeling the domain: An introduction to the expert module*. in Advances in Intelligent Tutoring Systems, ser. Studies in Computational Intelligence, R. Nkambou, J. Bourdeau, and R. Mizoguchi, Eds. Springer Berlin Heidelberg, 2010, no. 308,pp. 15—32.

[19] L. Giraffa, *Seleção e adoção de Estratégias de Ensino em Sistemas Tutores Inteligentes*. Porto Alegre, 1997. Exame de Qualificação (Doutorado em Ciência da Computação. Instituto de Informática) Universidade Federal do Rio Grande do Sul. 127 P.

[20] F.S. Gharehchopogh and Z.A. Khalifelu, *Using intelligent tutoring systems in instruction and education*. in 2nd International Conference on Education and Management Technology, vol. 13. IACSIT Press Singapore, 2011, pp. 250–254.

[21] F.M. Oliveira and R.M. Viccari, *Are learning systems distributed or social systems?*. EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE IN EDUCATION, I, Lisboa - Portugal. 1996.

[22] E.B. Costa, *Um modelo de Ambiente Interativo de Ensino-Aprendizagem baseado numa Arquitetura Multi-Agentes*. Campina Grande, 1997. Exame de Qualificação (Doutorado CPGEE), UFPA. SP - Brasil.

[23] K. Chrysafiadiand and M. Virvou, *Student modeling approaches: A literature review for the last decade*. Expert Systems with Applications, vol.40, no.11, pp. 4715–4729, 2013

[24] C. Carmona and R. Cornejo, *A learner model in a distributed environment*. in Adaptive Hypermedia and Adaptive Web-Based Systems. Springer, 2004, pp. 353–359.

[25] A. Martins and L.C. Faria and E. Carrapatoso, *User modeling in adaptive hypermedia educational systems*. Journal of Educational Technology & Society, vol.11, no.1 ,pp. 194–207, 2008

[26] L. Nguyenand and P. Do, *Learner modeling adaptive learning*. World Academy of Science, Engineering and Technology, vol. 45, pp. 395–400, 2008.

[27] E. Rich, *Stereotypes and user modeling in User Models*. in Dialog Systems, ser. Symbolic Computation, A.Kobsaand W.Wahlster, Eds. Springer Berlin Heidelberg, 1989, pp.35–51.

[28] J. Kay, *Stereotypes, student models and scrutabilit*. in Intelligent Tutoring Systems, Lecture Notes in Computer Science, G. Gauthier, C. Frasson, and K.V. Lehn, Eds. Springer Berlin, no.1839, pp.19–30.

[29] A. Mitrovic, *Fifteen year sof constraint-based tutors: what we have achieved and where weare going*. UserModeling and User-Adapted Interaction, vol.22, no.1-2, pp.39–72, 2012.

[30] E. Millan and T. Lobo and J. Perez-de-la-Cruz, *Bayesian networks for student model engineering*. Computers & Education, vol. 55, no.4, pp.1663–1683, 2010.

[31] M. Danaparamita and F. Gaol, *Comparing Student Model Accuracy with Bayesian Network and Fuzzy Logic*. in Predicting Student Knowledge Level. International Journal of Multimedia and Ubiquitous Engineering, vol.9, no.4, pp. 109–120, 2014.

[32] L. Alfaro and C. Rivera and J. Luna-Urquizo, *Using Project-based Learning in a Hybrid e-Learning System Model*. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 10, 2019.

[33] L. Alfaro and C. Rivera and J. Luna-Urquizo, *Utilization of a Neuro Fuzzy Model for the Online Detection of Learning Styles in Adaptive e-Learning Systems*. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 12, 2018.

[34] L. Alfaro and E. Apaza and J. Luna-Urquizo and C. Rivera, *Identification of Learning Styles and Automatic Assignment of Projects in an Adaptive e-Learning Environment using Project based Learning*. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 11, 2019.

[35] S. Blessing and S. Gilbert and S. Ourada and S. Ritter, *Authoring model tracing cognitive tutors*. International Journal of Artificial Intelligence in Education, vol. 19, no.2, p.189, 2009.

[36] A. Flores and L. Alfaro and J. Herrera and E. Hinojosa, *Proposal Models for Personalization of e-Learning based on Flow Theory and Artificial Intelligence*. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019.

[37] A. Flores and L. Alfaro and J. Herrera, *Proposal model for e-learning based on Case Based Reasoning and Reinforcement Learning.*. In IEEE World Conference on Engineering Education (EDUNINE), 2019.

[38] P. Brusilovsky, *Adaptive hypermedia: From intelligent tutoring systems to web-based education*. in Intelligent Tutoring Systems, ser. Lecture Notes in Computer Science, G. Gauthier, C. Frasson, and K.Van Lehn,Eds. Springer Berlin Heidelberg, 2000, no.1839, pp.1–7, 2000.

[39] P. De-Bra, *Adaptive educational hypermedia on the web*. Communications of the ACM, vol.45, no.5, pp.60–61, 2002.

[40] J. Eklund and P. Brusilovsky, *The value of adaptivity in hypermedia learning environments: A short review of empirical evidence*. in Proceedings of 2nd Adaptive Hypertext and Hypermedia Workshop at the 9th ACM International Hypertext Conference Hypertext, vol.98, 1998, pp. 11–17.

[41] J. Eklund and P. Brusilovsky, *Interbook: an adaptive tutoring system*. Uni Serve Science News, vol.12, no.3, pp.8–13, 1999.

[42] D. Braand and L. Calvi, *Aha! an open adaptive hypermedia architecture*. New Review of Hypermedia and Multimedia, vol.4, no.1, pp.115–139, 1998.

[43] P. Brusilovsky, *Adaptive hypermedia for education and training, iaptive Technologies for Training and Education*. Cambridge University Press, 2012, pp.46–68.

[44] A. Ogan and W. Johnson, *Preface for t the especial issue on culturall y aware educational ntechnologies*. Artif Intell Educ, vol. 25, pp.173–176, 2015.

[45] B.D. Nye, *Intelligent tutoring systems by and for the developing world: a review of trends and approaches for educational technology in aglobal context*. International Journal of Artificial Intelligence in Education, vol.25, no.2, pp.177–203, 2015.

[46] P. Mohammed and P. Mohan, *Dynamic cultural contextualization of educational content in intelligent learning environments using icon*. International Journal of Artificial Intelligence in Education, vol.25, no.2, pp.249–270, 2015.

[47] N. Dowell, *What works: Creating adaptive and intelligent systems for collaborative learning support*. International Journal of Artificial Intelligence in Education, vol.25, no.2, bpp.177–203, 2015.

[48] Y. Lou and P.C. Abrami and S. Apollonia, *Small group and individual learning with technology: A meta-analysis*. Review of educational research, vol.71, no.3, pp.449–521, 2001.

[49] A. Soller, *Supporting social interaction in an intelligent collaborative learning system*. International Journal of Artificial Intelligence in Education (IJAIED), vol.12, pp.40–62, 2001.

[50] Y. Hayashi, *Togetherness: Multiple pedagogical conversational agents as companions in collaborative learning*. in Intelligent Tutoring Systems, ser. Lecture Notes in Computer Science, S.Trausan Matu, K.E.Boyer, M.Crosby, and K. Panourgia, Eds. Springer International Publishing, 2014 , no. 8474, pp. 114–123.

[51] B. Kyun and B. Young, *Gaming for Classroom-Based Learning: Digital Role Playing as a Motivator of Study*. IGI Global, 2010.

[52] A. Raut and S.D.A. Uroojussama and U. Farheen and A.Anwari, *Game based intelligent tutoring system*. International Journal of Engineering Research and General Science, vol.3, no.2, 2015.

[53] M. Easterday and V. Aleven and R. Scheines and S.Carver, *Using tutors to improve educational games*. in Artificial Intelligence in Education, Lecture Notes in Computer Science, G. Biswas, S. Bull, J. Kay, and A. Mitrovic, Eds. Springer Berlin 2011 , no. 6738, p p. 63–71.

[54] G. Jackson and D. McNamara, *Motivational impacts of a game-based intelligent tutoring system*. in 24th International FLAIRS Conference, 2011.

[55] J. Lester, *Serious games get smart: Intelligent game-based learning environments*. AI Magazine, vol.34, no.4, pp. 31–45, 2013.

[56] M. Badaracco and L. Martínez, *An Intelligent Tutoring System Architecture for Competency-Based Learning*. A. König et al. (Eds.): KES 2011, Part II, LNAI 6882, pp. 124–133, Springer-Verlag Berlin, 2011

[57] H. Al-Rekhawi and S.A. Naser, *An Intelligent Tutoring System for Learning Android Applications UI Development*. International Journal of Engineering and Information Systems (IJEAIS) ISSN: 2000-000X Vol. 2 Issue 1, January – 2018, Pages: 1-14

[58] M. Hamed and S.A. Naser, *An intelligent tutoring system for teaching the 7 characteristics for living things*. International Journal of Advanced Research and Development ISSN: 2455-4030, Volume 2; Issue 1; January 2017; Page No. 31-35

[59] N. Gon and S. Biswa and L. Ba, *An Intelligent Tutoring System for Multimedia Virtual Power Laboratory*. ASEE's 123rd Annual Conference & Exposition. New Orleans, LA. June 26-29, 2016

[60] W. Yu-Ying, *Modeling and verification of an intelligent tutoring system based on Petri net theory*. Mathematical Biosciences and Engineering Volume 16, Issue 5, 4947–4975.

An Algorithmic Approach for Maritime Transportation

Dr. Peri Pinakpani[1]
HBS, GITAM University
Hyderabad, TS – 502329

Dr. Aruna Polisetty[2]
IoM, GITAM, Rushikonda
Visakhapatnam, AP – 530045

G Bhaskar N Rao[3]
DoC, MAHE, Manipal University
Manipal – 576104

Prof. Harrison Sunil D[4]
College of Business and Economics
Blue Hora University, Ethiopia

Dr. B Mohan Kumar[5]
Professor, Aurora PG College
Hyderabad

Dandamudi Deepthi[6]
Faculty, The Princeton Review
Hyderabad, TS – 500072

Aneesh Sidhireddy[7]
ECE – IV Year, Roll No:
16BEC0690
VIT University, Vellore, Tamil Nadu,
India - 632 014

***Abstract*—Starting from the 3rd millennium BC, Indian maritime trade has augmented the life of a common man and businesses alike. This study, finds that India can leverage on the 7,500 long coast line and derive holistic development in terms of interconnected ports with hinterland connectivity and realize lower expenditure coupled with reduced carbon emission. This research analyzed a decade of cargo data from origination to destination and found that around 82.95 per cent (953 MMTPA in 2017–18) of road based consignments in India comprised of Fertilizers, Hydrocarbons, Coal, Lubricants and Oil. Essentially, a quantum of this i.e. 78.39 per cent of MMTPA cargo consignments (State Owned Hydrocarbons) traverses on Indian roads. The study drew parameters of this transportation paradigm and modeled the same using Artificial Intelligence to depict a monumental opportunity to rationalize costs, improve efficiency and reduce carbon emission to strengthen the argument for the employment of Multimodal Logistics in the Maritime Sector. Subsequent to model derivation the same set of parameters are plotted as an efficient transit map of Interstate transit lines connecting 16 major hubs which now handle bulk cargo shipped by all modes of transport. For the pollution segment a collaborative game theoretic approach i.e., Shapley value is proposed for improved decision making. This study presents data driven and compelling research evidence to portray the benefits of collaboration between firms in terms of time and cost. The study also proposes the need and method to improve hinterland connectivity using a scalable greedy algorithm which is tested with real time data of Coal and Bulk Cargo. As a scientific value addition, this study presents a mathematical model that can be implemented across geographies seamlessly using Information Communication Technology.**

Keywords*—*Maritime transport; multimodal logistics; game theory; greedy algorithm; freight management; intermodal transportation

I. INTRODUCTION

Leading National newspaper “Business Standard” in the morning of 16-May-17, shared with the Indian transportation and logistics fraternity that with an expenditure budget of Rs. 33,000 crores, the process to build 15 Multimodal parks across strategic coordinates has been initiated, reflecting a monumental infrastructural building regime for the next decade or more. A glimpse into the details of this announcement led to the discovery of an all-encompassing ginormous transportation paradigm and a grand vision to fuel growth by “Port Led Development”.

With the enactment of systematic programs for the revamp of the Logistics sector, now will be a right time where the Nation can deliberate on this infrastructural solution. Only one-third of the transportation system had been constructed in the past, thus giving a scope of two-third of infrastructure to be constructed in future. From past experiences and adopting international practices, India can draft a novel Maritime strategy that can assist to minimize investment, maximize cost efficiency and reduce different types of losses by the creation of better transportation infrastructure.

A primary goal of this research is to contribute to the National Transportation policy framework by way of designing a scalable and sustainable model in order to transform the Nation’s logistics infrastructure. It tries to portray the scenario of Indian logistics in year 2020. It identifies the facts about the Indian logistics infrastructure and emphasizes the need for growth and development in order to satisfy the demand of the large Indian population.

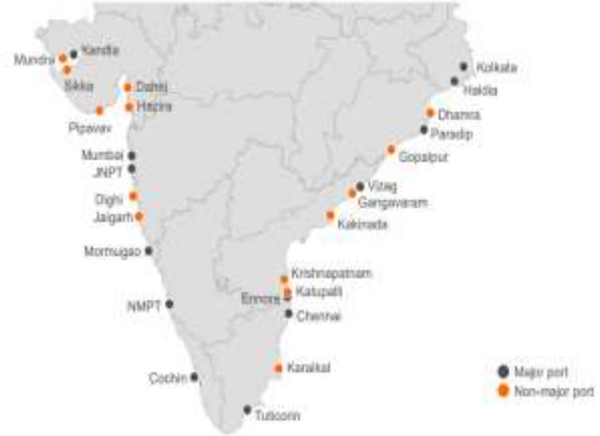
A. Research Objectives

This manuscript seeks to objectively analyze the fundamental mechanism of the Indian Logistics Sector encompassing all modes of transport. This study also endeavors to act as an enabler for seamless implementation of *a)* policy guidance and Plan perspective,

b) coordinate planning protocols for a multitude of products, and

c) structure and create building blocks for implementation by global stakeholders.

To addresses the need to ameliorate exports and optimize pricing by enabling micro, small, medium enterprises inside the port complex(s) thus rationalize time and expenditure with 100 per cent compliance to norms in terms of shipping bulk cargo and containers across international waters, this study as a first objective seeks to explore if there can be a collaboration between the participants of road, rail and sea to possibly shift from the present national dependence on road transportation to the new resurgent Maritime sector. The second objective seeks to analyze strategically the scale of economies and provide resources for capacity augmentation in a scalable and sustainable procedure for upgradation of existing ports (i.e. 12 Major Ports and 200 Non-Major Ports) as per Fig. 1 below.



Source: DG-Shipping, India

Fig. 1. Prominent Ports in India (7,500 Kilometers of Coast).

Third Objective, orients towards an economic perspective; and seeks to portray the financial advantage gained by shifting from road transport (per truck load capacity price) to the maritime sector for movement of freight.

II. Significance of the Study

In the Transportation and Logistics sector, a basic and important criterion is to find solutions to the question "what drives future value creation?". Given that today's scenario of the transportation and logistics sector are getting shaped by the dynamic and ever changing global mega trends for a better future [1]. India relies more on road transportation as per Table I and takes top place as compared to many other high freight nations. India relies three times more on road transport despite the fact that India's freight traffic is comprised largely of bulky materials that move over longer distances, which could be served more economically by Rails and Waterways when compared to neighboring China [23].

Further, from Table I, it can be ascertained that India is highly dependent on road transportation for long distances which affect(s) adversely and abuses our environment with harmful emissions viz., SO2, CO2, CO, NO2, etc. Recent studies endorse that emission of CO2 is 84g per ton-km in roadways, 28 gm per ton-km in railways and 15 gm per ton-km in waterways. Despite this fact India uses road network predominantly for bulk transport. Only If India can shift moderately from roadways to railways it could be able to save about 5.71 per cent of its total energy consumption [27].

TABLE I. India's Freight Transport is More Road Oriented

(Weight-distance)				
#	**In Billions Ton-Per Km (s)**			**Emmission per ton-kmg CO2 equivalent**
	5,183	**5930**	**1325**	
Air	<1	<1	<1	>1000
Water	30 (Squared)	13.73	5.66	14.47
Freight Rail	46.35	47.95	35.38	27.4
By Road	22.81	36.32	51.55	63.63
	China	**USA**	**India**	

(Source: World Economic Forum; China Statistic Yearbook; Planning Commission India; NHAI; Indian Railways; DG Shipping; Bureau of Transportation Statistics US; McKinsey)

The recent interdisciplinary and proven methodology associated to Logistics, Supply Chain Management, Warehousing, Freight Management and Containers handling is "Multimodal Logistics". From this perspective enablement of Multimodal Logistics was mooted in the 12th Five Year Plan; [12] National (Government) Transport Development Policy Committee (NTDPC - 2016-17); Other reports i.e., Report on Indian Transport (Moving India to 2032), Internationally acclaimed "Annual (2016-17) Logistics Report of KPMG", the policy manual "Industry Outlook for 2017-18 by Price Waterhouse Coopers" (PwC) and "McKinsey's", Projects and Infrastructure Team's report for the year 2016-17 seek and advocate the deployment of a "Multimodal Structure" as a possible solution to the Indian transportation sector [32], [53].

This set of literature specifically attempted to isolate primary challenges as:

- Insignificant Kutcha roads and limited connectivity among network.
- Halting of vehicles indiscriminately at check posts of state border [7] (reason for delay – statistically 38.19 per cent of transit time delay is associated to these unscheduled halting).
- Porous system and weak regulations for starting a trucking business.
- Freight prices being subjected to surge pricing.
- Absence of tracking systems for rail freight, for distributed cargo management.
- Low infrastructure for disconnected rail & road connectedness.

Inefficient berthing of ship and unusual delay in time for loading and unloading that lead to abnormally high turnaround time of vessels [47].

At a national scale, this study calls for and portrays a systematic model (inclusive of gained profits and time) for the integration of freight modal mix, specific trade and economic zones, interconnectivity towards and inside the port, and improvement or building suitable infrastructure for scaling of operations [2]. This study asserts that the possible solution can be inclusive of,

a) rationalization of Operational costs,

b) enablement of seamless Cold Chain(s),

c) systematic Containerization,

d) establishment of ancillaries like ICD's and CFS's and

e) operationalization of dry ports and possibly forward link them to integrators like logistics parks.

There is still a very high scope and infrastructure requirement in India [4]. Therefore, the required infrastructure could be re-designed to address the rapidly burgeoning demand. For this India needs to adopt a positive outlook, and cross-integrate each mode of transportation (air, water and land). It should be matched and optimized to the needs and available resources from origination to destination. In particular, India will have to move from roadways to seaways, and should also realize the potential scope of its waterways [24].

The coalescent approach adopted by this study can assist India to increase its Maritime transport share to 42.91 per cent [20].

If India delays to register this paradigm shift, the pollution caused from present {not-up-to-the mark logistics framework} would be very high viz., from USD $ 44.16 billion (equal to 3.92 per cent of Indian Gross Domestic Product) to USD $ 13942 Million or even higher (can reach a total of 4.91 per cent of India's GDP) by 2020.

Therefore, this study asserts the serious requirement to tackle this situation by integrating and coordinating of different modes. This can help India to reduce this waste generation by half and can also lead to reduction in fuel consumption by 15-20 per cent [18] and [48].

III. Theoretical Overview

This study seeks to derive the benefits of collaboration by way of Game Theory applications of tariff and profit (as single parameters).

Transport tariff, cost structure and determination, marginal costs and Shapley value along with semi-proportional transport costs. Profit maximization, and simulation results were assessed to identify relevant theoretical foundations.

For reflecting on the possible advantage of optimizing modes of transport, algorithms are employed; a greedy algorithm is employed as the decision making is dynamic and transient when the demand for merging cross functional entities and multimodal logistics are modeling using dynamic programing for scheduling. Different mode of transportation in India is a heritage handed over by British empire during its colonial rule over India.

The present framework is what the British have ideated about two centuries ago; even in 2018, most of Indian cargo moves on the same network. In lieu of this Indian transportation network is not properly designed and cannot handle increasing freight. Growth and development of Indian economy [50] will lead to increase more pressure on the existing logistics infrastructure. Four Dimensional components as per Table II outlined below characterize the network of Indian logistics [44], [13].

The study attempts to integrate the following six key success factors (as per Fig. 2) to propose a multimodal system as a plausible solution to some of the challenges India address as on date [6].

A. *Constructing and Optimizing Multimodality*

Multi modal transportation or multimodal logistics park is a facility that provides a singular access to all transportation. It is a complex facilitation comprising of earmarked spaces for all operations required from a transportation perspective. Container Terminals, Stowage Facilities, Warehouses, [26]. Access to Rail Network (Freight), Financial Centers, 3rd Party Logistic providers and inter-modal transport [40]. The key components of Logistics Park are: (a) Transportation, (b) Storage Facilities, and (c) 360 0 degree service operations as a single window clearance.

B. *Multimodal Logistics Parks (MMLP)*

Multimodal Logistics Park can augment Indian transportation infrastructure to rationalize expenditure on transportation [28]. The most important characteristic; it helps to reduce the overall transit time [25].

TABLE II. Dimensions and Applications Profile

#	Dimensions		Application area
1	Integrated Framework	Components in Networked Roads	Maritime Frieght Network
			Golden Quadrilateral
			Interlinking Road and Rail
2	Enablers	Enabler to support Structure	Logistics Parks
3	Payback	Proposed Change	Automation of e-Pass
4	Budgets	New Focus	Seaways

Source: Author Collated

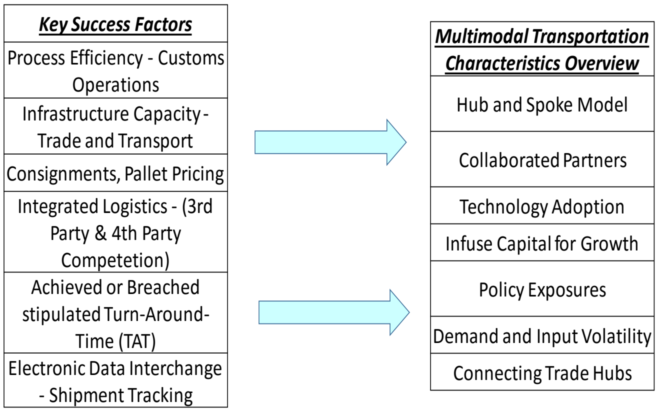

Key Success Factors
Process Efficiency - Customs Operations
Infrastructure Capacity - Trade and Transport
Consignments, Pallet Pricing
Integrated Logistics - (3rd Party & 4th Party Competetion)
Achieved or Breached stipulated Turn-Around-Time (TAT)
Electronic Data Interchange - Shipment Tracking
Multimodal Transportation Characteristics Overview
Hub and Spoke Model
Collaborated Partners
Technology Adoption
Infuse Capital for Growth
Policy Exposures
Demand and Input Volatility
Connecting Trade Hubs

Fig. 2. Key Success Factors of Multimodal Logistics.

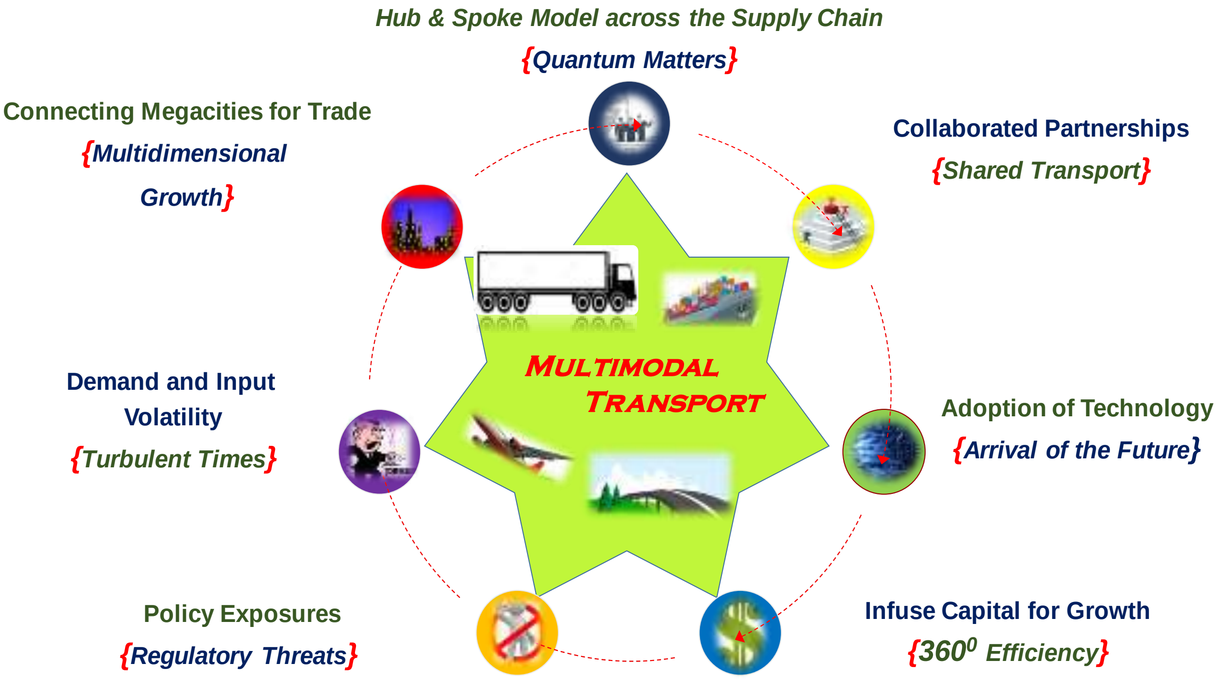


Fig. 3. The all-encompassing Multimodal Logistics.

In-turn this helps to decrease inventory carrying cost for both, the ultimate customer as well as the logistics operator. It also helps to utilize all the resources optimally with decreased transit time [41]. MMLP contributes in equitable growth of all modes of transportation, proper utilization of pooled assets like railways.

Proper implementation of green technology can reduce greenhouse gases emission and help to decrease dependency on fossil fuels such as crude oil, coal etc. The advantages are:

1) Automating container transportation system, improved hinterland and inter modal connectivity.

2) Implementing intelligent transport system, use of Global positioning software to track movement of freight and manage interconnected transport. Implementation of environmentally designed hybrid trains.

Depicted in Fig. 3 is a master model of the all-encompassing Multimodal Logistics Park.

IV. Literature Review

The study explores the paradigm of connecting hinterland through port based multimodal logistics deployment via published research work s as part of the literature and identify associative affairs detailing integration with a view so as to recognize the procedure prior to integration [35]. For the achievement of the objective to ensure hinterland connectivity, it is important to analyze the aspect of "what gets transported by roads", which is contributing to the high transportation cost as compared to other modes of transport [3].

It is quite prominent that, in India, [45] "FIVE" commodities add-up to 79.62 per cent of entire import-export shipments in India with dynamic origination and destination coordinates across the nation.

These are: (a) Coking Coal, (b) Petrol & Diesel, (c) Processed Oil & Lubricants (d) Ore of Iron (e) Agricultural Grade Fertilizers and bulk movement related "Container(s)" [9].

The literature which deals with the factors that determine partnerships between institutions and corporates who operate in the transportation space for these products can give insights in terms of forming the [37], fundamental structure of multimodal framework establishment that can yield in tangible results which can be specific to these commodities.

This study seeks to analyze four dimensions which are being modelled for possible enablement of Multimodal Logistics.

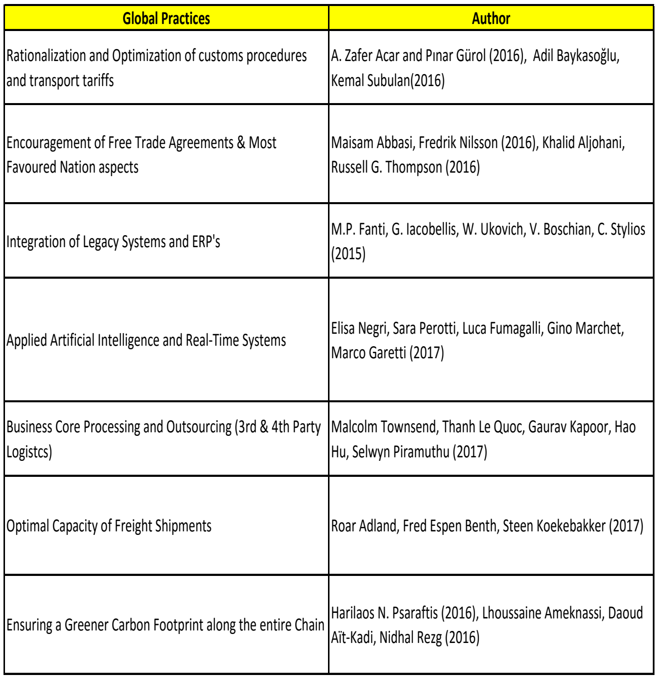
a) Incorporate a methodical change in the way Indian Logistics operates as compared to Global Practices.

b) Incorporate the Variables of Transportation Paradigm being modeled which can build enablers for realizing port led transportation and development.

c) Depict multiple key Success Factors for MMLF that build's Institutions which operate within the Multimodal Logistics Framework.

d) Present model characteristics which enable technology creation to assist and sustain new greenfield projects. In stage-1; the study seeks to understand the multifaceted and inter connected structures including methodology [52], [54].

Global Research practices aligned to the research work for the transportation segment are presented in Fig. 4.

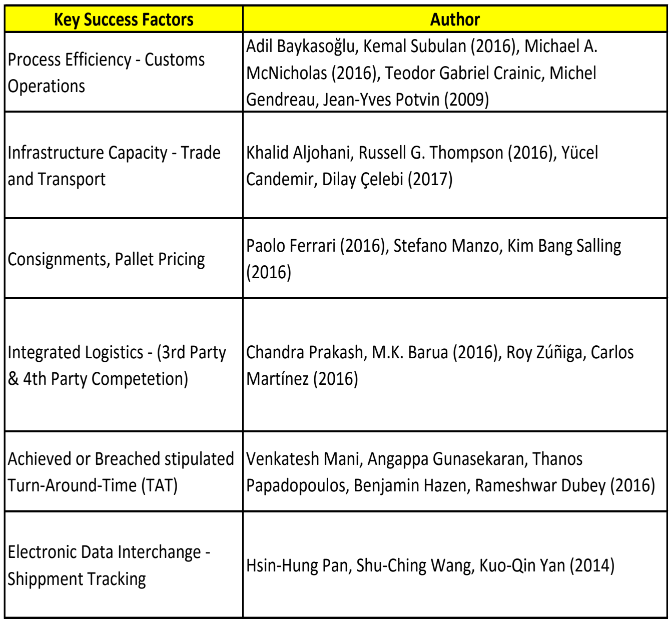


Global Practices	Author
Rationalization and Optimization of customs procedures and transport tariffs	A. Zafer Acar and Pınar Gürol (2016), Adil Baykasoğlu, Kemal Subulan(2016)
Encouragement of Free Trade Agreements & Most Favoured Nation aspects	Maisam Abbasi, Fredrik Nilsson (2016), Khalid Aljohani, Russell G. Thompson (2016)
Integration of Legacy Systems and ERP's	M.P. Fanti, G. Iacobellis, W. Ukovich, V. Boschian, C. Stylios (2015)
Applied Artificial Intelligence and Real-Time Systems	Elisa Negri, Sara Perotti, Luca Fumagalli, Gino Marchet, Marco Garetti (2017)
Business Core Processing and Outsourcing (3rd & 4th Party Logistcs)	Malcolm Townsend, Thanh Le Quoc, Gaurav Kapoor, Hao Hu, Selwyn Piramuthu (2017)
Optimal Capacity of Freight Shipments	Roar Adland, Fred Espen Benth, Steen Koekebakker (2017)
Ensuring a Greener Carbon Footprint along the entire Chain	Harilaos N. Psaraftis (2016), Lhoussaine Ameknassi, Daoud Aït-Kadi, Nidhal Rezg (2016)

Fig. 4. Literature Review of Global Practices [14], [17].

A. Profit Considerations and Optimal Inventory

As a replacement for altering the decision responsibility in higher association arrangements, harmonization among constituent firms can share profits. While in unconditional terms, government spends very less for the growth and development of logistics industry while the cost of operations and maintenance is quite high of this industry in India because of inefficiencies. India spends 12.21 per cent of Gross Domestic Product on transportation and is higher than the expenses incurred by that of North America (9.46 per cent) and Germany (7.91 per cent) [33], [36]. Other key Success Factors are presented in Fig. 5 (below).



Key Success Factors	Author
Process Efficiency - Customs Operations	Adil Baykasoğlu, Kemal Subulan (2016), Michael A. McNicholas (2016), Teodor Gabriel Crainic, Michel Gendreau, Jean-Yves Potvin (2009)
Infrastructure Capacity - Trade and Transport	Khalid Aljohani, Russell G. Thompson (2016), Yücel Candemir, Dilay Çelebi (2017)
Consignments, Pallet Pricing	Paolo Ferrari (2016), Stefano Manzo, Kim Bang Salling (2016)
Integrated Logistics - (3rd Party & 4th Party Competetion)	Chandra Prakash, M.K. Barua (2016), Roy Zúñiga, Carlos Martínez (2016)
Achieved or Breached stipulated Turn-Around-Time (TAT)	Venkatesh Mani, Angappa Gunasekaran, Thanos Papadopoulos, Benjamin Hazen, Rameshwar Dubey (2016)
Electronic Data Interchange - Shippment Tracking	Hsin-Hung Pan, Shu-Ching Wang, Kuo-Qin Yan (2014)

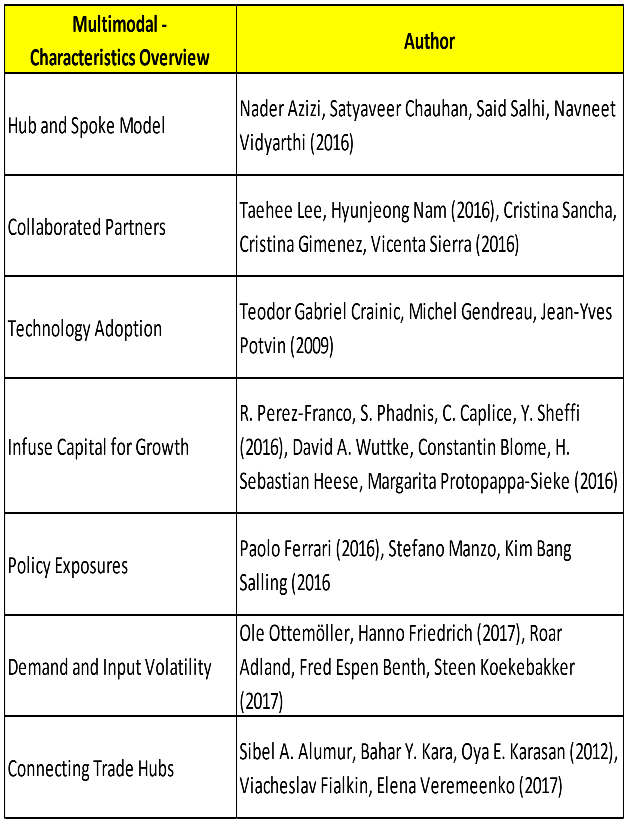
Fig. 5. Literature Review of Key Success Factors for MMLF [38].

North American Infrastructure Organizations consider and believe that Indian Logistics Infrastructure is very poor and inefficient. For example, the expenses for Freight by rail and the maritime transport are approximately 68.29 per cent more than that of their expenses for all modes of employed transport in the USA. Similarly, the road cost is also high by 30.05 per cent in India compared to US. This leads to increase the prices as well as lower the rate of competency. It also hampers the economic growth of the country [30], [34]. The research suggests, poor logistics infrastructure cost an extra of 45 billion USD to one's economy i.e. 4.3 percent of GDP every year. One unknown fact is that two-third of these costs are hidden from outside world [43], [19].

B. Possible Technology for New Green Field Projects

If above mentioned shifts as per Fig. 6 are Implemented, India would be able to bring down its logistics cost by almost one-third of its logistics waste USD 100 billion by 2020. Further it could be lowered to USD 7127 Million (THREE percent of Indian Gross Domestic Product) [21], [49]. If government could increase the investment on this industry to USD 700 billion. It would result to lower commercial deployment of energy in excess of 1.25 per cent.

This calls for an integrated plan and policy which needs to target on improved energy efficiency, reducing economic waste and to have greater share of rail. Such plan will require to enable a multitude of programs such as coastal freight corridor, road maintenance, technology adoption, last mile roads, last mile rail, dedicated rail freight corridor, skill development and equipment and service standards [42], [8].



Multimodal - Characteristics Overview	Author
Hub and Spoke Model	Nader Azizi, Satyaveer Chauhan, Said Salhi, Navneet Vidyarthi (2016)
Collaborated Partners	Taehee Lee, Hyunjeong Nam (2016), Cristina Sancha, Cristina Gimenez, Vicenta Sierra (2016)
Technology Adoption	Teodor Gabriel Crainic, Michel Gendreau, Jean-Yves Potvin (2009)
Infuse Capital for Growth	R. Perez-Franco, S. Phadnis, C. Caplice, Y. Sheffi (2016), David A. Wuttke, Constantin Blome, H. Sebastian Heese, Margarita Protopappa-Sieke (2016)
Policy Exposures	Paolo Ferrari (2016), Stefano Manzo, Kim Bang Salling (2016
Demand and Input Volatility	Ole Ottemöller, Hanno Friedrich (2017), Roar Adland, Fred Espen Benth, Steen Koekebakker (2017)
Connecting Trade Hubs	Sibel A. Alumur, Bahar Y. Kara, Oya E. Karasan (2012), Viacheslav Fialkin, Elena Veremeenko (2017)

Fig. 6. Literature Review of Modal Characteristics [55].

V. Model Discussions: Methodology

For enabling multi-modal logistics, the below participants and their characteristics are pivotal:

- Business Processes in the Freight and Containers Segment.
- The Routing and geo-mapping of the National expressways integrated to Ports.
- Parameters associated to the Interconnectedness of roads with adoption of Technology.
- Introduction of the Shapley Value and its constructs.

A. Analyzing Collaboration in Supply Chains

From Game Theory, the Shapley value as a proven methodology is adopted for this study and illustrates the incremental gains that are related to each participant in the market place. It determines the gains that can be derived from each level of collaboration extended by both players on an individual basis. The summation of the gains can be evaluated prior to executing a market strategy. The weighted costs and gains are determined for each sequence of actions that the players can perform [29].

For the interaction of variables associated to the partnership formation for facilitation of inter-modal and multimodal logistics for the Indian context this mathematical model can be applied by Shapley allocation [10], [46]. Shapley method is the most optimal technique in the Logistics paradigm derived from economics and is a chosen because of the multitude of option for variable definitions and building constructs. This method renders a path to parametrize the created value by collaboration to its respective input participants in the Indian Transportation Sector. Historically, the Shapley values for non-dependent input (Martin Shubik, 1978) variables are tabulated for understanding variance as a key parameter. The application of Shapley values for this structure of inputs modeled as dependent variables. This study addresses only the basic constructs and the associated appropriateness of the established Shapley Method to primary dependent clusters of variable(s), and not to iterate or computational methodology [11] [39]. Shapley value is a mono variate outcome in cooperative games, as postulated by Shapley S Lloyd way back-in 1953. It helps analyze incremental contribution of each participant to the partnership and all permutations of the games as desired or modeled. Carbon output as per Shapley value is derived using the principle Shapley function as below:

$$y_j = \sum_{s\subseteq R/j} P\left(M\right)\left(d\left(M\cup\{j\}\right)-d\left(M\right)\right) \quad . \quad ..A$$

where '*j'* is a random participant either as an originator or as a factor load, y_j is the associated CO$_2$ of participant '*j'* , $P\left(M\right)$ is the probability of the participated outcome M , $d\left(M\right)$ is attributed to the carbon output of the partnership as a yield function of participation outcome 'M ' and $d\left(M\cup\{j\}\right)-d\left(M\right)$ is attributed to the incremental emission of CO$_2$ induced by adding participant '*j'* into the partnership '*M*' .

Based on the above primary functionality, below derivation describes how costs can be formally realized and described by marginal procedure of partnerships. The impact of the last arrived order being of size n is

$$t_{|Q|}(Q,n) \quad . \qquad (1)$$

The average of $t_{|Q|}(Q,n)$is taken for the cases where $n_{|Q|}=n$ to arrive at the rate

$T(n)$ is associated to a consignment of quantum n ,

$$T(n)=E(t_{|Q|}(Q,n)/n_{|Q|}=n) \quad . \qquad (2)$$

For an order pool

$$Q=(n_1,...,n_{|Q|}) \quad , \qquad (3)$$

The following recursions are taken into consideration, for tabulating (α_i) , this is termed as allocations, and iterative variables (p_i) , termed as reminders,

$$\alpha_i = \min\{\omega_i\rho_{i-1},\#(n_i)\} \qquad (4)$$

and $\rho_i = \rho_{i-1} - \alpha_i$ (5)

With p_0 some given positive number and weights by

$$\omega_i = n_i/(n_i+...+n_{|Q|}) \qquad (6)$$

Note that $\omega_{|Q|}=1$.The allocation α_i is such that if $n_i = n_{i+1}$ then

$$\alpha_i = \alpha_{i+1} \quad . \qquad (7)$$

Future tabulations can associate α_i to yield n_i as an effective collaboration and is sorted according to decreasing size. Now the desired quantum:

$$Q=(n_1,...,n_{|Q|}) \qquad (8)$$

is termed

$$n_1 \geq n_2 \geq ...n_{|Q|} \quad . \qquad (9)$$

The set of instructions can be iterated only if any $n\in Q$. The yield set can comprise of $\#(n)=1$ for all n that comprises of all instructions associated to collaboration for cost and time advantage among competing associates.

B. *Algorithmic Approach for Multimodal Logistics*

A greedy algorithm adopts a solution path and arrives at a local optimality at each stage and can scale exponentially to arrive at a global optimum across various activities interlinking of activities at the same time. In the current study, a greedy heuristic is determined to yield desired results in a quick time and is scalable across the parties willing to collaborate for deriving mutual benefits. Each iteration or collaboration can be subjected to the Huffman Method of deduction and arrival at optimality. Huffman Code depends on the greedy-choice characteristics and the aspect of optimal substructure [16]. To prove relatedness of this algorithm to Multimodal Logistics, instead of demonstrations, the pseudocode is derived initially.

This will assist in clarification that the choice will follow the optimality of a Greedy algorithm property. Allocating that P would comprise of a set of k elements and each element is $p \in P$ with bounded frequency of $f(p)$.

The heuristic constructs the loop A to an optimality from the end point. The loop begins with the set $|P|$ and ends with a sequence of $|P|-1$ creating a loop-tree illustration.

A lesser priority queue L, iterated as a function, is applied to merge together two of the lesser frequented objects. The yield is a new characteristic and the associated frequency would be the summation of the merged entities (Fu-Sheng Chang, May 2014).

Let $\mu = \{b_1, b_2, b_3 .... b_k\}$ of k proposed activities which seek for a common mode of transport like a Cape Size Vessel that can process only one consignment at a time. Each decision activity function of q_1 has an initial instance time s_i and a closing time r_1, where $0 \leq r_1 \prec s_i \prec \infty$.

The below is a constituent option of one activity and one selection between the multimodality and logistic operators. The linking and de-linking of the nodes would assist us in terms of a dynamic understanding of where to interlink so that optimality may be achieved as a function prior to executing and interlinking operations as per Fig. 7.

Greedy interlinking of sub-nodes of the transport mode and the rational chosen choice problem:

$$\mu_{a,b} = \{P_k \in \mu : f_a \leq r_l \prec f_b \prec s_m\} \quad \ldots(10)$$

$\mu_{a,b}$ would be the subset of all initiated function. First in first out procedure is adopted and functionally,

$$a_0 \leq a_1 \leq a_2 \leq a_3 \leq ... \leq a_n \leq a_{n+1} \quad \ldots\ (11)$$

i	1	2	3	4	5	6	7	8	9	10	11
r_i	2	4	1	6	4	6	7	9	9	3	13
s_i	5	6	7	8	9	10	11	12	13	14	15

Fig. 7. Illustration of advantages by enabling Multimodal Logistics

Subsets comprising of maximum functional elements are compatible mutually as an optimal choice:

$$b_{j,l} = b_{j,k} \cup \{r_k\} \cup b_{k,l} \tag{12}$$

1) A greedy recursive solution: The constant C value is the optimal choice comprising of both cost and time. This optimality is the quantum that we seek for at real time. There are $p-i-1$ profit values for C, to derive in $C = i+1, ... p-1$. The best possible choice to form a multimodal solution is the maximum subset $\mu_{a,b}$. The optimal modal is the choice of C. The complete recursive recurrence yield would be

$$r[a,b] = \begin{cases} 0 \\ \max\limits_{\substack{a \prec k \prec b \\ s_k \in \mu_{a,b}}} \{r[a,k] + r[k,b] + 1\} \end{cases}$$

$$\begin{aligned} &if\, \mu_{a,b} = \varnothing \\ &if\, \mu_{a,b} \neq \varnothing \end{aligned} \tag{13}$$

2) Transformation to a greedy response from a dynamic – program

Lemma 1.1

Initiated activity parameter $\mu_{a,b}$ is chosen; let r_i be the interlink with the most optimal choice and time greedy

$$f_i = \min\{f_c : p_k \in \mu_{a,b}\} \tag{14}$$

Then

a) Modal Function r_i is applied as mutually interlinked activity for cost and time optimization for activities defined as per $\mu_{a,b}$.

b) The complete Supply Chain modal-optimization $\mu_{a,b}$ is defined as NULL, as to opting for r_i would ensure that the previous set of nodes used for the multimodal of $\mu_{a,b}$ is not an empty optimality.

The interlinked activity-solution r_i needs to be chosen with the least time variation that is (Paweł B, January 2018-Accepted Manuscript) applicable from an inter-modal group of logistic functions. The chosen modality is termed as a "greedy" choice as it renders the saved time quotient for possible programming of other variables that determine the modal mix in the unscheduled remainder of time quotient [51].

3) The recursive greedy algorithm: On the Maritime transport front, this study asserts that, an algorithm functioning in a pure greedy top-down approach is termed as a *"RECURSIVE LOGISTIC FUNCTION"* decision as per Fig. 8. The initial sequence and completion duration can be collated as arrays of r and s, and the functional indices of l and m that can design a solution for the sub-modal, $S_{l,m+1}$. This function will qualify a maximum quantum of interconnected nodes that are internally connected. For a single activity selector that is integrated with a forward function, increases the completion time as per [15].

$$a_0 \le a_1 \le a_2 \le a_3 \le ... \le a_n \le a_{n+1}. \quad [15]$$

This study seeks to segregate these modal participants into $R(m, \log, m)$ which randomly connects loops which give maximum reduction in terms of time and cost parameters.

"RECURSIVE LOGISTIC FUNCTION" (r, f, i, s)

```
1. $a \leftarrow k+1$
2. while $a \le b$ and $r_a \prec f_i$ ▷ find the first activity in
       $\mu_{a,b+1}$
3.         do $a \leftarrow a+1$
4. if $a \le b$
5. then return $\{r_i\} \cup$ RECURSIVE LOGISTIC
       FUNCTION" $(r, f, a, b)$
6. else return $\varnothing \leftarrow$ Object location root of the
   summation.
```

4) An iterative greedy algorithm: Mathematically, it is simple to transform a recursive method to an iterative procedure. *RECURSIVE LOGISTIC FUNCTION"* (r, f, a, b) ends with a recursive condition only to be succeeded by a Set Union operator [5]. The functional model is depicted in Fig. 8 (below).

```
1. $l \leftarrow length[a]$
2. $R \leftarrow \{s_1\}$
3. $i \leftarrow 1$
4. for $r \leftarrow 2$ to $l$
5.     do if $a_r \ge f_i$
6.         then $R \leftarrow R \cup \{r_i\}$
7. $i \leftarrow n$
8. return $R$
```

The procedure works as follows. The logistic component *a* associates with newer counterparts of transport modes R yields activity a_r as a recursive component. As the functions are [22] incrementally profit and time optimization mandates the end delivery R. That is,

$$f_i = \max\{f_k : n_k \in R\} \quad (16)$$

Similar to the recursive iteration, [31] GREEDY-LOGISTICAL-OPERATOR schedules a functional activity set of *l* activities in $\lambda(l)$ in shortest duration assuming that the collaboration is pre-fixed and time and cost parameters are arrived at.

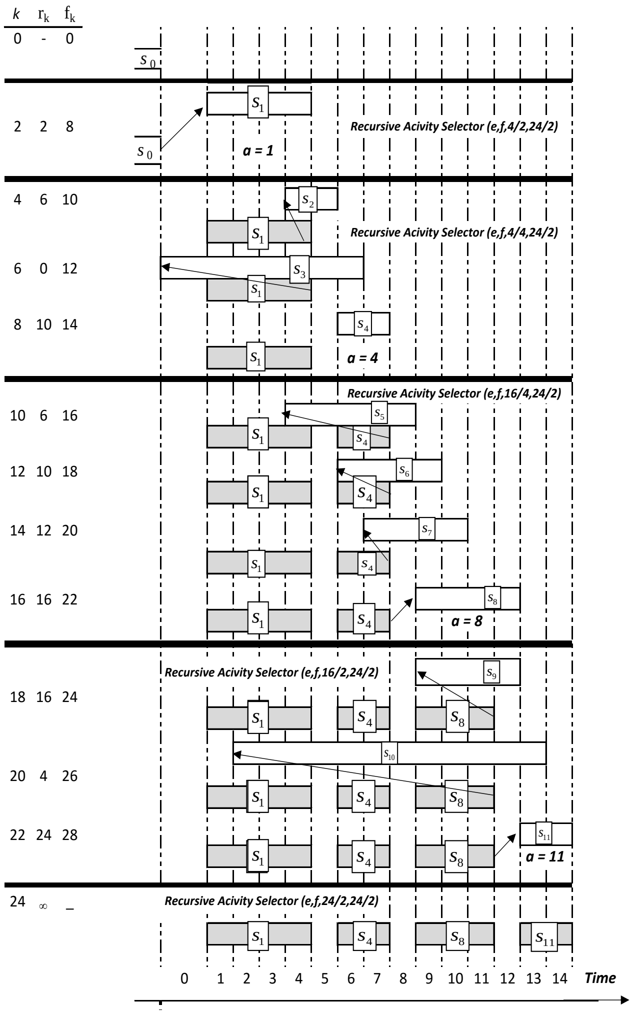

k rk fk
0 - 0
2 2 8
4 6 10
6 0 12
8 10 14
10 6 16
12 10 18
14 12 20
16 16 22
18 16 24
20 4 26
22 24 28
24 ∞ –
Recursive Acivity Selector (e,f,4/2,24/2)
a = 1
Recursive Acivity Selector (e,f,4/4,24/2)
a = 4
Recursive Acivity Selector (e,f,16/4,24/2)
a = 8
Recursive Acivity Selector (e,f,16/2,24/2)
a = 11
Recursive Acivity Selector (e,f,24/2,24/2)
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 Time

Fig. 8. Recursive Logistic Function as start (s) with variables (a), Greedy Logistic Function (a, b).

VI. Data Analysis, Interpretation and Suggestions

India imports crude predominantly at only four of the major ports across its large 7,500 kms of its coastal corridor. This study has chosen the segment of Crude oil receipt versus its transport on road which impacts both the cost and time parameters coupled with the aspect of pollution.

The above derived Algorithm can be applied to rationalize both variables of cost and travel duration. A comparison of the transported consignments for thermal coal, oil and lubricants (POL), fertilizers, containers and iron ore is presented which averages around 78.53 per cent of total freight quantum volumes of 1372 MMTPA in 2016–17) currently handled from the major ports in India.

The study asserts that should the initiative of multimodal logistics be commissioned from the Eight major ports, logistics cost–saving opportunity could be around INR 34,000-38,000 Crores per annum, by optimizing freight transportation. Four key initiatives could drive these savings:

- Inland water shipping can handle about 221-241 MMTPA from the incremental capacity in the next 5-6 years across the five commodities as above and estimated cost reduction would be around 19,500-24,500 Crores by 2026.

- Bulk Shipments from Major Ports Cement & Fertilizers of 78-88 MMTPA estimated INR 4,200-6,200 Crores saving by 2026.
- Transit time reduction in the container shipment segment by 120 hours can be estimated INR 4,300-5,700 Crores saving by 2026.

A transition to rail transport for containers from roadways from current 17.6 percent by 2026 can reduce expenditure by 1,800-2,800 INR crores.

A. *Coastal Shipping for Existing/Planned Capacities*

1) Coal: In 2016–17, around 1,317 MMTPA of thermal coal was shipped by Indian railways alone. Around 48 MMTPA moved through waterways given the accounted price of INR 0.19 per tonne km vs. INR 1.19 to 1.37 per tonne km. Given the 1/6th price of transport via rail, Indian can for the 392 thermal plants deploy waterways to interconnect and ferry 95 to 118 MMTPA coal and ease the undue stress on rail based transportation and reduce expenditure by INR 12,370 by 2025. The routes are presented in Fig. 9.

B. *Bulk commodities: Iron Ore, Grains and Cement*

Bulk Freight Stations as per below Fig. 10 have always been set-up next to the natural reserves of raw material. 76 per cent of capacity outlay follows this structure. Multimodal Logistics, on the other hand, offers transportation cost optimization, ease of raw material flow, and improved linkages with international markets.

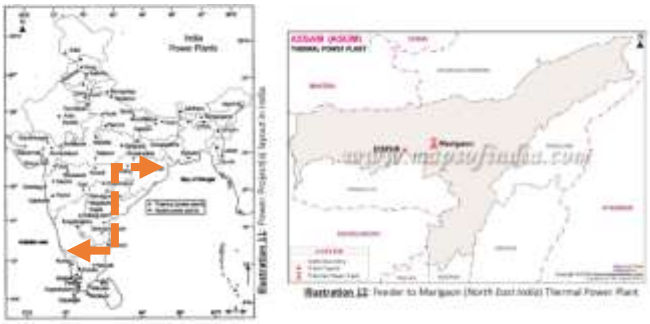

India
Power Plants
ASSAM (ASOM)
THERMAL POWER PLANT
www.mapsofindia.com
Illustration 12: Feeder to Marigaon (North East India) Thermal Power Plant

Fig. 9. Shipment of Coal in the Indian Subcontinent.

Fig. 10. Transhipment of Bulk in the Southern Part of the Country.

Grains and Cement, the other two commodities analyzed estimates that a potential of approximately 84–94 MMTPA (~42 MMTPA for Cement & ~42 MMTPA for Dry Bulk) can be improved at existing costs by 2026.

Multimodal Logistics routing raw material from mine to coast as a fundamental structure as derived by the Greedy Algorithm asserts that, an average transportation expenditure of INR 670 to INR 940 per tonne can be optimized as per Fig. 10 above.

As portrayed in Table III major bulk freight is routed through identified locations for North Andhra Pradesh and Tamil Nadu, Odisha, connecting Telangana and Southern Maharashtra, a distance of approximately 1650 Kilometres.

TABLE III. SOURCING QUANTUM FROM GOVERNMENT WAREHOUSE TO DESTINATION CONNECTING NORTH AND SOUTH INDIA MMTA

#	**2013-14**	**2014-15**	**2015-16**	**2016-17**	**2017-18**
Wheat	268	271	282	299	308
Soyabean	93	102	106	112	132
Total Grains	**361**	**373**	**388**	**411**	**440**
Dry Bulk	3857	4205	4373	4635	4812
% Grains / Total Dry Bulk	**9.36**	**8.87**	**8.87**	**8.87**	**9.14**

Source: National Bulk Handling Corporation, 2018 (***Q_2***)

This study asserts that Logistic Parks with ICD's could be established close to distribution centres like Krishnapatnam which is well connected through inland waterways.

For Limestone, the study proposes Vijayawada in Andhra Pradesh and the Gujarat clusters from Vadodara based on the reserve of raw material for limestone.

C. *Reduce Time to Export by Five Days*

Hinterland container shipment in India averages 30 days, which is about 25 in other parts of South East Asia which houses five large ports for the same distance. This unwanted transit time compels exporters to earmark higher buffer duration (Xu, 2018). This research proposes three initiatives for optimizing container transit time by 80-100 hours:

a) Interconnect Highways with Ports with Logistics Parks as the convergence points: The Bharatmala initiative and the Golden Quadrilateral can be earmarked with dedicated freight-friendly corridors and establish custom houses at the logistics parks, with RFID enabled EXIM container sealing which reduces inspection time and reduce unwanted halting of containers by adopting exclusive pre-paid toll tags across all modes of transport. A summation is presented in Fig. 11.

b) Simplification of customs reforms: The automated Customs Clearance currently deployed at Mumbai (M/s JNPT), Gujarat (M/s Mundra) and AP (M/s Krishnapatnam), associated to the EXIM license to generate unique routing numbers to permit single-window document validation extended to 24 X 7 and 365 days a year for participants of import and export.

c) Redeploy containers to railways transport: Which is otherwise skewed in favour of road transport in India. This can reduce crude imports by 1.15 Mn KL. The greedy algorithm proposed by this study identifies EIGHT priority routes for road to rail deployment that currently deploys 2.19 Mn TEU from highways but are capacitated to ferry 3.09 Mn TEU by railways.

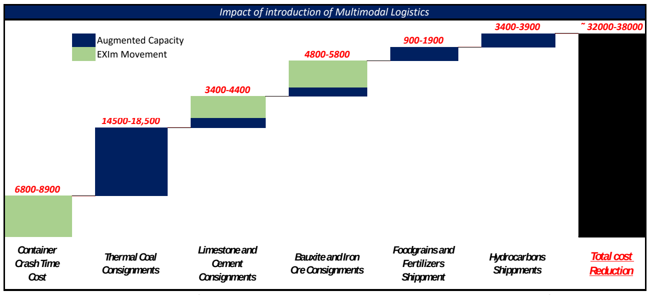
d) Enable Direct Freight Corridor within Western India and Eastern India: Exclusive railway lines to interlink ports with the warehouses at Pipavav, Hazira, Mundra, Kandla. On the Eastern front, Chennai, Krishnapatnam, Ennore, Kakinada and Visakhapatnam.

e) Enable Inland Container Depot: The Greedy Algorithm as proposed has identified an opportunity in that, city Tughlakabad gets 13 rail rakes a day, as against higher transport demands from trade centers of Bhopal and Agra receive less than 1 rail-rake per day. The study proposes, an exclusive Milk-run from Gujarat through the Inland water ways to other parts of India.

D. Bulk Cargo Transport

Major ports in India have handled around 1318 MMTPA of bulk cargo in 2016–17. This study estimates that by 2026, this segment can go up to 1,975 MMTPA. EXIM bulk can go higher by 3.5 per cent to reach 1,030 MMTPA. The port based bulk freight is poised to improve by 20 per cent to breach 421 million tonnes by 2026. This demands enabling dedicated logistics parks with multimodal capabilities at specific ports to manage 12.73 Mn TEU container traffic in 2016–17. Container shipments have witnessed a SEVEN per cent over the last five years as has the extent of containerization from 52 per cent in 2015–16 to 28.75 per cent in 2016–17. This study estimates that container traffic will register a 6.45 per cent rate to attain 22.75 Mn TEU by 2026. The following infrastructure for Multimodal transportation will need to be installed to address this increased traffic.

All INR in Crores



Source: Compiled from Various Sources and Extrapolation from the Algorithm derived

Fig. 11. Possible Reduction in Expenditure with the implementation of Multimodal Logistics.

1) Transhipment facility at a Southernmost tip connecting Indian Ocean and Bay of Bengal with capacity of 11-14 Mn TEU.

2) Increase capacity of Western port M/s Mundra Port and M/s Navasheva Port by 1.95-2.45 Mn TEU and on the East Ennore, Paradip, Krishnapatnam and Visakhapatnam to be ameliorated.

3) New dedicated freight corridor for transhipment from West Bengal to Andhra Pradesh connecting Odisha with capacity of 1.2 Mn TEU.

VII. Summary And Conclusions

In the marine transportation paradigm, integration of stakeholders services and product lines as a collaboration improves competitiveness. Collaboration can occur in terms of an ICT enabled shared logistic design to enable reduction of empty kilometers across larger transit lines, freight management in terms of shared load and warehouses, possible implementation of multimodal logistics (Raut, Gardas, Jha, & Priyadarshinee, 2017). The study reflected on ways to operationalize "Maritime" as the spearhead to harness development as depicted in Fig. 12.

1) This can be realized by application of analytical methods for transportation intensive units and is representative of expenses, time duration, as primary variables to be optimized for better profitability. Game theory applications can ascertain the equilibrium point for such kind of collaboration and Shapley values derive the maximum benefit from dependent variables as discussed. This model can be seamlessly extrapolated and multiple variables can be added as filters to derive needed results. The manuscript portrays the positive impact of collaboration in terms of costs and time rationalization. The implementation of the transportation space by way of waterways would be possible by integration of non-major and major ports on the Southern and Western Coast of India.

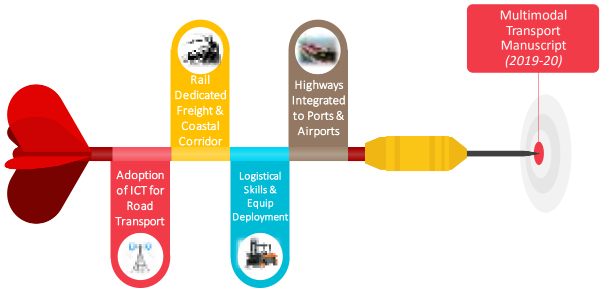

Multimodal Transport Manuscript (2019-20)
Rail Dedicated Freight & Coastal Corridor
Highways Integrated to Ports & Airports
Adoption of ICT for Road Transport
Logistical Skills & Equip Deployment

Fig. 12. The roadmap to arrive at the Multimodal Structure.

2) The identified variables of this study are attempted to be modeled using a greedy algorithm which is best suited when participants are multiple and decision making need to be dynamic. The study has identified variables to optimize the operations of all participants in the transportation segment, shared information aspects are pooled as a dependent variable, and as the other participant, being In-Time associated information, variables of Electronic Data Interfaces with market mechanism and operationalization of methodology, quantifiable messages, authenticity of information.

3) For the aspect of recursive frameworks in dynamic programming, the applicable paradigm is identified in terms of Information Communication Technology framework, quantum metric experience (recorded digitally). The two constructs time-span and market capitalization are taken as exogenous variables. The aspects of Mutual understanding, Long-Term relation & viability, leadership exchange facilitation, economic characteristics coupled with Market Conditions, supplier's market competitiveness, market and product characteristics, augmenting low demand and volume, network breakages, fluctuating demand constructs are taken as a third cluster.

4) These three clusters are integral for the future of Indian Supply Chain networks especially from a multi-modal perspective. It is very crucial to understand the end user financing avenues as this will shed light as to what problems or advantages the MSMEs face when acquiring finance. The methods and models applied in the transportation space gives an inside view of the macro process taking place in the back end. As a measurable output the following can mooted to initiate the process of a Port-Led Development as called upon by the national leaders.

5) The enablement of transportations warehouses and strategic centers as a key component of a connectivity plan around each operational port to facilitate cargo labeling and processing using scientific methods, based on the quantum numbers for the last five years in terms of cost and time. Once the optimality is arrived, design the same for each stakeholder and possibly integrate them using multimodal logistics. For smooth and seamless movement of freight, the existing industrial corridors need to be scaled and newer avenues identified. Possibly strategic enablement of warehouses along the Golden Quadrilateral road network to handle bulk logistics through the maritime sector.

6) The present legislation of multiple authorities and approvals for similar cargo; this system which replicates from each state as and when cargo moves needs to be reduced and can lead to reduction in transit durations. Possible geo-tagging of trucks and consignments can be looked into apart from reducing documentation for export and import containers initially and then scaled to cargo in a progressive manner. Real time systems and artificial intelligence can be harnessed to ensure that all stakeholders inclusive of the government departments interact with each other with improved efficiency and effectiveness.

References

[1] Adil Baykasoğlu, Kemal Subulan; "A multi-objective sustainable load planning model for intermodal transportation networks with a real-life application", Transportation Research Part E: Logistics and Transportation Review, Volume 95, November 2016, Pages 207-247.

[2] Agata Mesjasz Lech; "Urban Air Pollution Challenge for Green Logistics", Transportation Research Procedia, Volume 16, 2016, Pages 355-365.

[3] Allan Woodburn; "An empirical study of the variability in the composition of British freight trains", Journal of Rail Transport Planning & Management, Volume 5, Issue 4, December 2015, Pages 294-308.

[4] Ayse Sena Eruguz, Tarkan Tan, Geert-Jan van Houtum; "A survey of maintenance and service logistics management: Classification and research agenda from a maritime sector perspective", Computers & Operations Research, Volume 85, September 2017, Pages 184-205.

[5] Bilel Marzouki, Olfa Belkahla Driss, Khaled Ghédira; "Multi Agent model based on Chemical Reaction Optimization with Greedy algorithm for Flexible Job shop Scheduling Problem", Procedia Computer Science, Volume 112, 2017, Pages 81-90.

[6] Boban Melović, Slavica Mitrović, Arton Djokaj, Nikolai Vatin; "Logistics in the Function of Customer Service – Relevance for the Engineering Management", Procedia Engineering, Volume 117, 2015, Pages 802-807.

[7] Carlo Vaghi, Luca Lucietti; "Costs and Benefits of Speeding up Reporting Formalities in Maritime Transport", Transportation Research Procedia, Volume 14, 2016, Pages 213-222.

[8] Chandra Prakash, M.K. Barua; "A combined MCDM approach for evaluation and selection of third-party reverse logistics partner for Indian electronics industry", Sustainable Production and Consumption, Volume 7, July 2016, Pages 66-78.

[9] Chaug-Ing Hsu, Hsien-Hung Shih, Wei-Che Wang; "Applying RFID to reduce delay in import cargo customs clearance process", Computers & Industrial Engineering, Volume 57, Issue 2, September 2009, Pages 506-519.

[10] Cristina Sancha, Cristina Gimenez, Vicenta Sierra; "Achieving a socially responsible supply chain through assessment and collaboration"; Journal of Cleaner Production, Volume 112, Part 3, 20 January 2016, Pages 1934-1947; Data Analytics for Intelligent Transportation Systems, 2017, Pages 1-29.

[11] David A. Wuttke, Constantin Blome, H. Sebastian Heese, Margarita Protopappa-Sieke; "Supply chain finance: Optimal introduction and adoption decisions"; International Journal of Production Economics, Volume 178, August 2016, Pages 72-81.

[12] David Gillen, Hamed Hasheminia; "Measuring reliability of transportation networks using snapshots of movements in the network – An analytical and empirical study", Transportation Research Part B: Methodological, Volume 93, Part B, November 2016, Pages 808-824.

[13] Dezhi Zhang, Qingwen Zhan, Yuche Chen, Shuangyan Li; "Joint optimization of logistics infrastructure investments and subsidies in a regional logistics network with CO2 emission reduction targets", Transportation Research Part D: Transport and Environment, In press, corrected proof, Available online 14 March 2016.

[14] Elisa Negri, Sara Perotti, Luca Fumagalli, Gino Marchet, Marco Garetti; "Modelling internal logistics systems through ontologies", Computers in Industry, Volume 88, June 2017, Pages 19-34.

[15] Ellen Kenia Fraga Coelho, Geraldo Robson Mateus; "A capacitated plant location model for Reverse Logistics Activities", Journal of Cleaner Production, Volume 167, 20 November 2017, Pages 1165-1176.

[16] Fu-Sheng Chang, Jain-Shing Wu, Chung-Nan Lee, Hung-Che Shen; "Greedy-search-based multi-objective genetic algorithm for emergency logistics scheduling", Expert Systems with Applications, Volume 41, Issue 6, May 2014, Pages 2947-2956.

[17] Hangtian Xu, Hidekazu Itoh; "Density economies and transport geography: Evidence from the container shipping industry", Journal of Urban Economics, Volume 105, May 2018, Pages 121-132

[18] Harilaos N. Psaraftis; "Green Maritime Logistics: The Quest for Win-win Solutions", Transportation Research Procedia, Volume 14, 2016, Pages 133-142.

[19] Hsin-Hung Pan, Shu-Ching Wang, Kuo-Qin Yan; "An integrated data exchange platform for Intelligent Transportation Systems"; Computer Standards & Interfaces, Volume 36, Issue 3, March 2014, Pages 657-671.

[20] Hyun Jung Nam, Yo Han An; "Default Risk and Firm Value of Shipping & Logistics Firms in Korea" The Asian Journal of Shipping and Logistics, Volume 33, Issue 2, July 2017, Pages 61-65.

[21] Jens Ehm, Michael Freitag, Enzo M. Frazzon; "A Heuristic Optimisation Approach for the Scheduling of Integrated Manufacturing and Distribution Systems"; Procedia CIRP, Volume 57, 2016, Pages 357-361.

[22] Jonas Volland, Andreas Fügener, Jan Schoenfelder, Jens O. Brunner; "Material logistics in hospitals: A literature review", Omega, Volume 69, June 2017, Pages 82-101.

[23] Khalid Aljohani, Russell G. Thompson; "Impacts of logistics sprawl on the urban environment and logistics: Taxonomy and review of literature", Journal of Transport Geography, Volume 57, December 2016, Pages 255-263.

[24] Lhoussaine Ameknassi, Daoud Aït-Kadi, Nidhal Rezg; "Integration of logistics outsourcing decisions in a green supply chain design: A stochastic multi-objective multi-period multi-product programming

model", International Journal of Production Economics, Volume 182, December 2016, Pages 165-184.

[25] M. Grazia Speranza; "Trends in transportation and logistics", European Journal of Operational Research, Volume 264, Issue 3, 1 February 2018, Pages 830-836.

[26] M.P. Fanti, G. Iacobellis, W. Ukovich, V. Boschian, C. Stylios; "A simulation based Decision Support System for logistics management", Journal of Computational Science, Volume 10, September 2015, Pages 86-96.

[27] Maisam Abbasi, Fredrik Nilsson; "Developing environmentally sustainable logistics: Exploring themes and challenges from a logistics service providers' perspective", Transportation Research Part D: Transport and Environment, Volume 46, July 2016, Pages 273-283

[28] Malcolm Townsend, Thanh Le Quoc, Gaurav Kapoor, Hao Hu, Selwyn Piramuthu; "Real-Time business data acquisition: How frequent is frequent enough?", Information & Management, In press, corrected proof, Available online 18 October 2017.

[29] Martin Shubik; "Game Theory: Economic Applications," in W. Kruskal and J.M. Tanur, ed., International Encyclopedia of Statistics, 1978, v. 2, pp. 372–78.

[30] Megan Thomas, Nicky Westwood; "Student experience of hub and spoke model of placement allocation - An evaluative study", Nurse Education Today, Volume 46, November 2016, Pages 24-28

[31] Meiyan Lin, Kwai-Sang Chin, Chao Fu, Kwok-Leung Tsui; "An effective greedy method for the Meals-On-Wheels service districting problem", Computers & Industrial Engineering, Volume 106, April 2017, Pages 1-19.

[32] Michael A. Mc.Nicholas; "International and U.S. Maritime Security Regulations and Programs", Maritime Security (Second Edition), 2016, Pages 91-135.

[33] Mike Brison, Yann LeTallec; "Transforming cold chain performance and management in lower-income countries", Vaccine, Volume 35, Issue 17, 19 April 2017, Pages 2107-2109.

[34] Nader Azizi, Satyaveer Chauhan, Said Salhi, Navneet Vidyarthi; "The impact of hub failure in hub-and-spoke networks: Mathematical formulations and solution techniques", Computers & Operations Research, Volume 65, January 2016, Pages 174-188.

[35] Ole Ottemöller, Hanno Friedrich; "Modelling change in supply-chain-structures and its effect on freight transport demand", Transportation Research Part E: Logistics and Transportation Review, In press, corrected proof, Available online 4 September 2017.

[36] Oludaisi Adekomaya, Tamba Jamiru, Rotimi Sadiku, Zhongjie Huan; "Sustaining the shelf life of fresh food in cold chain – A burden on the environment", Alexandria Engineering Journal, Volume 55, Issue 2, June 2016, Pages 1359-1365.

[37] Paolo Ferrari; "Instability and dynamic cost elasticities in freight transport systems", Transport Policy, Volume 49, July 2016, Pages 226-233.

[38] Paweł B. Myszkowski, Łukasz P. Olech, Maciej Laszczyk, Marek E. Skowroński; "Hybrid Differential Evolution and Greedy Algorithm (DEGR) for solving Multi-Skill Resource-Constrained Project Scheduling Problem", Applied Soft Computing, Volume 62, January 2018, Pages 1-14.

[39] R. Perez-Franco, S. Phadnis, C. Caplice, Y. Sheffi; "Rethinking supply chain strategy as a conceptual system", International Journal of Production Economics, Volume 182, December 2016, Pages 384-396.

[40] Rafik Makhloufi, Diego Cattaruzza, Frédéric Meunier, Nabil Absi, Dominique Feillet; "Real Time Systems in Logistics, Simulation of Mutualized Urban Logistics Systems with Real-time Management", Transportation Research Procedia, Volume 6, 2015, Pages 365-376.

[41] Roar Adland, Fred Espen Benth, Steen Koekebakker; "Multivariate modeling and analysis of regional ocean freight rates", Transportation Research Part E: Logistics and Transportation Review, In press, corrected proof, Available online 3 November 2017.

[42] Roy Zúñiga, Carlos Martínez; "A third-party logistics provider: To be or not to be a highly reliable organization", Journal of Business Research, Volume 69, Issue 10, October 2016, Pages 4435-4453.

[43] Sakib M. Khan, Mizanur Rahman, Amy Apon, Mashrur Chowdhury; "Chapter 1: Characteristics of Intelligent Transportation Systems and Its Relationship with Data Analytics",

[44] Sibel A. Alumur, Bahar Y. Kara, Oya E. Karasan; "Multimodal hub location and hub network design", Omega, Volume 40, Issue 6, December 2012, Pages 927-939.

[45] Stefano Manzo, Kim Bang Salling; "Integrating Life-cycle Assessment into Transport Cost-benefit Analysis", Transportation Research Procedia, Volume 14, 2016, Pages 273-282.

[46] Taehee Lee, Hyunjeong Nam; "An Empirical Study on the Impact of Individual and Organizational Supply Chain Orientation on Supply Chain Management", The Asian Journal of Shipping and Logistics, Volume 32, Issue 4, December 2016, Pages 249-255.

[47] Teodor Gabriel Crainic, Michel Gendreau, Jean-Yves Potvin; "Intelligent freight-transportation systems: Assessment and the contribution of operations research", Transportation Research Part C: Emerging Technologies, Volume 17, Issue 6, December 2009, Pages 541-557.

[48] Thomas Poulsen, Rasmus Lema; "Is the supply chain ready for the green transformation? The case of offshore wind logistics", Renewable and Sustainable Energy Reviews, Volume 73, June 2017, Pages 758-771.

[49] Venkatesh Mani, Angappa Gunasekaran, Thanos Papadopoulos, Benjamin Hazen, Rameshwar Dubey; "Supply chain social sustainability for developing nations: Evidence from India, Resources", Conservation and Recycling, Volume 111, August 2016, Pages 42-52.

[50] Viacheslav Fialkin, Elena Veremeenko; "Characteristics of Traffic Flow Management in Multimodal Transport Hub (by the Example of the Seaport)", Transportation Research Procedia, Volume 20, 2017, Pages 205-211.

[51] Weishi Shao, Dechang Pi, Zhongshi Shao; "Optimization of makespan for the distributed no-wait flow shop scheduling problem with iterated greedy algorithms", Knowledge-Based Systems, Volume 137, 1 December 2017, Pages 163-181.

[52] Xinqing Xiao, Zhigang Li, Maja Matetic, Marija Brkic Bakaric, Xiaoshuan Zhang; "Energy-efficient sensing method for table grapes cold chain management", Journal of Cleaner Production, Volume 152, 20 May 2017, Pages 77-87.

[53] Yücel Candemir, Dilay Çelebi; "An inquiry into the analysis of the Transport & Logistics Sectors' Role in Economic Development", Transportation Research Procedia, Volume 25, 2017, Pages 4692-4707.

[54] Yunlong Yu, Tiaojun Xiao; "Pricing and cold-chain service level decisions in a fresh agri-products supply chain with logistics outsourcing", Computers & Industrial Engineering, Volume 111, September 2017, Pages 56-66.

[55] Zafer A, Acar, Pınar Gürol; "An Innovative Solution for Transportation among Caspian Region", Procedia - Social and Behavioral Sciences, Volume 229, 19 August 2016, Pages 78-87.

Opportunistic use of Spectral Holes in Karachi using Convolutional Neural Networks

Aamir Zeb Shaikh[1], Shabbar Naqvi[2], Minaal Ali[3], Yamna Iqbal[4], Abdul Rahim[5], Saima Khadim[6], Talat Altaf[7]
Department of Electronic Engineering, NED University of Engineering and Technology Karachi, Pakistan[1, 3, 4, 5]
Department of Computer Systems Engineering, Balochistan University of Engineering and Technology Khuzdar, Pakistan[2]
Department of Telecommunications Engg, Dawood University of Engineering and Technology, Karachi. Pakistan[6]
Department of Electrical Engineering, Sir Syed University of Engineering and Technology Karachi, Pakistan[7]

***Abstract*—Wireless services appearing in the next generation wireless standard i.e. 6G include Internet of Everything (IoE), Holographic communications, smart transportation and smart cities require exponential rise in the bandwidth in addition to other requirements. The current static spectrum allocation policy does not allow any new entrant to exploit already grid-locked Radio Frequency (RF) spectrum. Hence, quest for larger bandwidth can be fulfilled through other technologies. These include exploiting sub-Terahertz band, Visible Light Communication and Cognitive Radio scheme or exploiting of RF bands in opportunistic fashion. Cognitive Radio is one of those engines to exploit the RF spectrum in secondary style. Cognitive Radio can use artificial intelligence driven algorithms to complete the task. Several intelligent algorithms can be used for better forecasting of spectral holes. Convolutional Neural Network (CNN) is a Deep Learning algorithm that can be used to predict the presence of a spectral hole that can be opportunistically exploited for efficient utilization of RF spectrum in secondary fashion. This paper investigates the performance of CNN for metropolitan Karachi city of Pakistan so that the users can be provided with uninterrupted access to the network even under busy hours. Dataset for the proposed setup is collected for 1805 MHz frequency band through NI 2901 Universal Software Radio Peripheral (USRP) devices. The root mean square error (RMSE) for the predicted results using CNN appears to be 81.02 at epoch of 200 and mini-batch loss of 3281.8. Based on the predicted results, it was concluded that CNN can be useful for investigating the possible opportunistic usage of RF spectrum; however, further investigation is required with different datasets.**

Keywords*—Cognitive radio; spectral hole; deep learning; Convolutional neural network (CNN)

I. Introduction

6G wireless communication standard and services discussion and investigation is already initiated. The proposed network initiatives promises to provide traffic capacity in the range of 1-10 Gbps/m3 as compared to 10 Gbps/m3 availability in 5G networks. It is assumed that the technologies that will play the role of enablers may include sub-Terahertz, visible light communication and Artificial Intelligence enabled cognitive communications. Additionally, greater use of multi-RAT and multi-link schemes will also be required to rectify the issues arising from higher frequency propagation and providing higher reliability communication links [1].

The higher capacity requirement of the future technologies is a great challenge towards ubiquitous connectivity of wireless devices. Cognitive Radio is a possible option to play key role towards successful exploitation of Radio Frequency (RF) spectrum in opportunistic fashion. Thus, producing enough bandwidth to provide connectivity to wireless devices even in the congestion time. Cognitive Radio is a novel concept to incorporate artificial intelligence enabled techniques to exploit spectrum in opportunistic fashion. The major requirement towards a successful exploitation of RF band is to use primary bands in such a fashion that the usage does not produce harmful interference to the primary users. The cognitive radio is implemented through cognitive cycle [2]. The process of cognitive cycle starts with gathering RF spectrum monitoring [3]–[5]. This can be done through spectrum sensing process [6] and geo-location databases [7]. Spectrum monitoring combined with Radio Environmental Maps (REM) can also produce a better method of identifying unused Spectral bands [8]. Additionally, in [8] it is presented that the geo-location based database scheme typically results in an underutilized spectrum sharing scenario hence, better spectrum measurement scheme shall be incorporated that includes joint spectrum monitoring network and REM schemes[8] [9].

In the Artificial Intelligence enabled spectrum sensing radios, radios measure the Signal to Noise Ratio (SNR) and Energy levels of the primary radio transmission and this data is fed to the training algorithms. These algorithms predict the future nature of the RF spectrum bands regarding their usage in future i.e. empty or occupied. For this activity, different algorithms can be involved. As the spectrum assignment is a completely random assignment so a linear algorithm like Logistic regression may not produce useful results in all the channel conditions with different available active and secondary users [10]. While in most of the nonlinear environments, different algorithms use many other algorithms due to their specific use such as Support Vector Machines, K-Nearest Neighbor (KNN), Linear discriminant analysis (LDA) and Decision Trees [11]. In comparison to the aforementioned machine learning algorithms [12]–[14], deep learning algorithms typically perform much better.

The efficiency of Deep Learning algorithms is far better than the machine learning algorithms in general. However, this efficiency is resulted on the basis of available dataset and the nonlinear nature of the outputs. For example in case of linear functions, the logistic regression is preferred over the nonlinear algorithms because not only the regression produces simpler results but also predicts the given function in perfect fashion. However, when the available function is nonlinear and we have luxury to collect the 5000 samples of the data set we choose deep learning based algorithms in comparison to machine learning algorithms. Typical applications of deep learning algorithms include the real time test cases when the logistic regression don't produce the required level of accuracy. Additionally, there are complex patterns that can only better identified through these complex algorithms. Typically, audio and video data come into the picture of the said category [15].

Convolutional Neural Networks (CNNs) find their way to the area of deep learning based problems at enormously higher rate. These algorithms find their applications into the domain of machine translation, sentence classification, and sentiment analysis. Similar to the classic case of slide window algorithms, these algorithms slide over the given sequence of characters [15].

In the current work, CNN is chosen to predict the data for the opportunistic use of spectral holes in a large metropolitan city of Karachi Pakistan. For this purpose, a dataset is trained using RMSprop, which is a fast learning algorithm. The CNN has been used as a high performance classifier. The operation of convolution strengthens and reduces interference with the original signal features and have better tolerance to noise. The training parameters in CNN are less than in a fully connected network [16]. Two factors including Root Mean Square Error (RMSE) and loss function have been used to interpret the results.

The rest of the paper is divided in such a way that Section-II describes the related work. It is followed by Section-III which provides details of the proposed system. Section-IV explains the results obtained. Section-V is the last section providing Conclusion and future directions.

II. Related Work

In the literature, Deep Neural Network based prediction algorithms are used by various authors to produce the data of spectral holes so that maximum number of users could be accommodated in available RF spectral bands [17]. Authors in [8] have presented a novel framework for spectrum monitoring purpose. The authors add CNN based algorithm to the spectrum measurement devices i.e. spectrum detectors that help the devices to identify the presence or absence of primary user signals (in this case, Radar signals are considered as primary users) even if the signals are found overlapped with other secondary users (in this case, WLAN and LTE are considered as secondary users). Furthermore, a large dataset containing various signal waveforms such as Radar pulses, LTE, WLAN and thermal noise waveforms is also developed that can be utilized by other researchers. A novel pre-processing scheme is also implemented that produces samples of amplitudes and phase shift of the collected waveforms. The results produced by the proposed scheme are excellent i.e. 99.6% accuracy is achieved when the proposed algorithm is run on testing dataset. This is in addition to the robustness of the proposed algorithm to noise. The proposed algorithm is also run on various SNR regimes that shows an improved performance of the selected algorithm in comparison to other models such as spectrogram-based CNN algorithms.

In [18], authors model the spectrum sensing issue as a classification problem by using CNN based processing. The received signal samples are initially normalized to overcome the effect of noise and uncertainties embedded in the received signal samples. The proposed model utilizes following different signaling schemes along with their variants i.e. PSK, FSK, QAM and AM. This encompasses reasonably wide variety of signals that are tested under the scenario considered in the paper. The maximum possible number of real world scenarios is considered for the training setup of proposed scheme. This feature will help the proposed scheme to perform better in the undefined wireless environments such as real-world wireless channels. The performance of proposed model is also compared with two algorithms maximum-minimum Eigen Value Ratio and frequency-domain entropy based methods. The results show better performance against its competitors even in the presence of colored noise. That shows the proposed scheme may find better use in the real-world environments. Additionally, to tackle the real world wireless issues, the proposed schemes is added with transfer learning technique to improve the performance. The data utilized in the proposed scheme is simulated except the experimentation into the real-world wireless environment testing.

Detection of RF signals under low SNR regime poses a great problem towards successful utilization of RF spectrum by opportunistic users [19]. In [16], authors attempt to solve the pressing issue (of detection under low SNR) by using CNN algorithm. The data collection for the proposed experimentation is performed by using Cyclostationary feature detector (An algorithm that converts the received signal into its signatures) and energy detection (An algorithm that measures the energy of the received signals). The performance of the proposed scheme is also compared with classic cyclostationary feature based detection schemes. The results show an improvement.

In classic spectrum sensing algorithms, typically model based approach is applied to sense the wireless environment i.e. energy and other methods, however, to devise better results the authors in [20] recommend a data driven approach to solve the problem of spectrum sensing through Deep convolutional network based spectrum monitoring approach. The proposed algorithm uses data to train itself. The sample covariance matrix is used as an input to the CNN algorithm. Maximum Aposterior Probability (MAP) based technique is used to devise the cost function. In offline method, the results of the proposed scheme show excellent performance by showing improved performance even under the correlated samples of primary user with correlation coefficient of 0.7 and uncorrelated primary user. Additionally, the proposed scheme works well than the conventional scheme i.e. Eigen value detection algorithm by approximately 7.5 times at an SNR level of -14 dB.

CNN is a branch of deep neural network in deep learning which is becoming popular tool for analyzing visual images. It includes image recognition, classification, and detection tasks [21]. They are referred to as convolutional neural networks because they use mathematical operation of convolution which is a specific type of linear operation [22]. In terms of Neural Networks, CNNs fall under category of generalized multilayer perceptron (MLPs). The major difference between CNN and MLP is that in MLP each input element is connected to each neuron in the hidden layer causing full connectivity whereas in CNN only a limited number of input elements called the receptive field is connected to only part of the hidden layer. CNN reduces the risk of over fitting of data in this way as compared to MLPs [23].

CNN is a type of deep learning which has been found to be very successful in areas of objects and image recognition, detection and segmentation challenges. CNNs applications in the various fields are on the rise. A number of reviews on applications of CNN have been done in the literature. A few recent reviews are shown in Table I. It can be seen that CNN applications are found in health informatics and in other image classification schemes.

A generalized CNN consists of input layer, an output layer and a number of hidden layers [8]. A generic CNN architecture has been shown in Fig. 1. CNN algorithm uses three layers to convert an input signal to output. The layers are: input, feature extraction e.g. learning and output layer. The input layer passes the data to a series of kernels i.e. convolutional layers with filters, pooling and fully connected (FC) layers. Then the Softmax function is used to decide the output data with a value of probability. Thus, it can be deduced that the convolutional layers works as an engine to extract features of input signal.

TABLE. I. YEAR WISE EXAMPLES OF RECENT REVIEWS DONE ON CNN

S.No	Paper Title	Year of Publication
1	Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review [24]	2019
2	Learning image-based spatial transformations via convolutional neural networks: A review [25]	2019
3	Medical Image Analysis using Convolutional Neural Networks: A Review [26]	2018
4	Skin Cancer Classification Using Convolutional Neural Networks: systematic review [27]	2018
5	Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review [21]	2017
6	Convolutional Neural Networks for Inverse Problems in Imaging: A Review [28]	2017

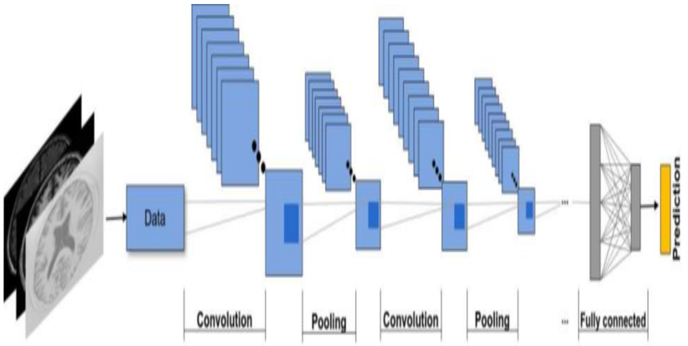


Fig. 1. Generic Architecture of Convolutional Neural Network [24].

It is also represented through mathematical operation of convolution. Convolutional layer works in conjunction with ReLU function. It is also called Rectified Linear Unit. There are many nonlinear functions that can be utilized such as Sigmoid; however, performance of ReLU under real environments is better than its competitors. The output of the layer is connected with pooling function. Pooling function is used to reduce the layer number of input parameters. Hence, it is also known as subsampling or down sampling unit. It is used to reduce the dimensionality of the input map; however, the significant information is always retained to make it highly useful function of the defined algorithm. It is generally implemented through max, average and sum pooling.

In this paper, CNN is used to predict the RF Usage in Karachi so that the available bands can be identified for further opportunistic usage. For this purpose, a dataset is also developed by using NI 2901 Universal Software Radio Peripheral (USRP) devices. CNN is applied to this dataset for prediction of the available RF bands. The RF band selected for the purpose is 1805 MHz. The RMSE for the proposed case comes out to be 81.02 at 200 epoch. The next section explains the details of the experimentation and analysis of the simulation results.

III. PROPOSED SYSTEM

Fig. 2 shows the structure of proposed system. In system, NI USRP 2901 is used to sense the presence of useful signals under real channel. A dataset is s developed using spectrum sensing algorithm. The Labview graphs are ported onto the Excel and Matlab to train the given algorithm. RMSprop is used for training purpose for the selected CNN Algorithm. Performance of RMSprop is closer to Momentum Gradient Descent Algorithm. It is used to control the learning rate thus taking larger steps produces convergence more quickly in horizontal direction. Two parameters including RMSE and Loss Function were used for investigation of the results. RMSE can be defined as standard deviation of residuals (prediction errors).

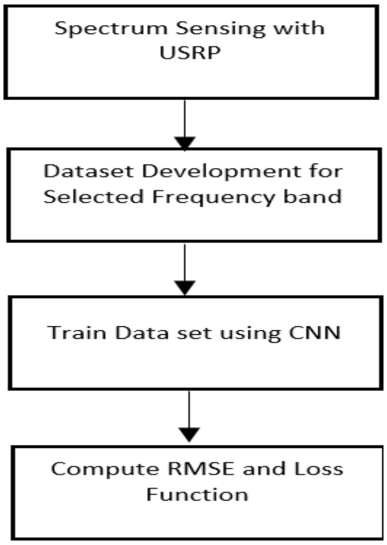
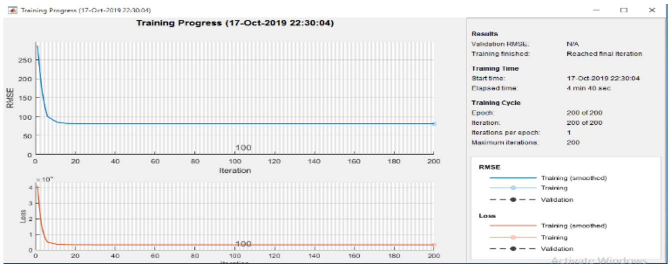

Spectrum Sensing with USRP
Dataset Development for Selected Frequency band
Train Data set using CNN
Compute RMSE and Loss Function

Fig. 2. Structure of Proposed System.

In other words, it shows the difference between the regression line and spread of the prediction errors. Machines work typically through loss function parameter. Loss function is used to determine the output of the given algorithm and the given target value. However, under circumstances when predicted values deviate too much from loss functions, the estimation error algorithms uses optimization functions.

IV. Results

Fig. 3 shows two parameters for the proposed setup. These include RMSE and Loss function. Total time to generate the results from proposed setup is 4 minutes and 40 seconds. One iteration is assumed for each epoch. Total epochs and iterations are taken to be 200. For first epoch and iteration, time requirement is 1 second. Mini- batch RMSE is 287 min-batch loss comes out to be 41185.5. Base Learning rate is 0.0010. Learning rate is used to determine how quickly CNN model learns the model. For 200 epoch and iteration, time requirement is 4 minutes and 40 seconds, whereas mini-batch RMSE is 81.02 and Mini-batch Loss is 3281.8 and Base learning Rate is 5.4976 e-31. Thus, as the iterations are increased RMSE and Loss function decreases but also Base Learning Rate decreases in larger amount. It means that chances of reaching a global optimal solution are high but there is also risk of getting stuck at a sub optimal solution. The training statistics for the experiments conducted is also shown in Table II for 200 iterations as till that time the result becomes consistent in terms of RMSE and loss function. Results indicate that for the dataset of Karachi, after few iterations, RMSE and Loss function become consistent, however rapid decrease in learning rate is an area which needs to be further investigated. It is worth mentioning that the proposed setup is implemented using a single CPU. Hence, for large amount of data, GPUs are recommended to use.



Results
Validation RMSE: N/A
Training finished: Reached final iteration

Training Time
Start time: 17-Oct-2019 22:30:04
Elapsed time: 4 min 40 sec

Training Cycle
Epoch: 200 of 200
Iteration: 200 of 200
Iterations per epoch: 1
Maximum iterations: 200

Fig. 3. Shows RMSE and Loss Functions for the Proposed Algorithm.

TABLE. II. Training Statistics for the Proposed Algorithm

EPOCH	Iteration	Time Elapsed (hh:mm:ss)	Mini Batch RMSE	Mini Batch LOSS	Base Learning Rate
1	1	00:00:01	287.00	41185.5	0.0010
50	50	00:01:15	81.02	3281.8	5.1200e-10
100	100	00:02:24	81.02	3281.8	5.2429e-17
150	150	00:03:32	81.02	3281.8	5.3678e-24
200	200	00:04:40	81.02	3281.8	5.4976e-31

V. Conclusion and Future Work

In this work, deep learning using CNN has been used to predict the spectral spaces for Karachi city. The algorithm takes longer time to train as compared to machine learning algorithms. The proposed setup is implemented using a single CPU machine. The results show that the best possible RMSE and Loss function are achieved at 50th iterations. Even after increasing the iterations, no significant progress on these parameters is found. However, the Base Learning Rate decreases significantly after running the machine for 200 iterations and needs to be further investigated and compared with other deep learning algorithms. This work was based on data set for metropolitan city of Karachi. In future, we aim at investigating the rural area datasets to find out any significant differences while changing datasets with CNN and other deep learning methods for investigating opportunistic use of spectral holes with the aid of artificial intelligence techniques.

Acknowledgment

The authors would like to thank the Administration of NED University of Engineering & Technology, Karachi, Pakistan, for providing resources to complete this research.

References

[1] E. C. Strinati, S. Barbarossa, J. L. Gonzalez-Jimenez, D. Kténas, N. Cassiau, and C. Dehos, “6G: The Next Frontier,” pp. 1–16, 2019.

[2] N. Savage, “Cognitive radio,” Technol. Rev., vol. 109, no. 1, pp. 61–62, 2006, doi: 10.15373/2249555x/apr2012/28.

[3] A. Z. Shaikh and T. Altaf, “Collaborative Spectrum Sensing under Suburban Environments,” Int. J. Adv. Comput. Sci. Appl., vol. 4, 2013.

[4] A. Z. Shaikh and L. Tamil, “Cognitive radio enabled telemedicine system,” Wirel. Pers. Commun., vol. 83, no. 1, pp. 765–778, 2015.

[5] A. A. Khan, A. Z. Shaikh, S. Naqvi, and T. Altaf, “A Novel Cognitive Radio Enabled IoT System for Smart Irrigation,” J. Inform. Math. Sci., vol. 9, no. 1, pp. 129–136, 2017.

[6] A. Ahmed, A. Zeb, S. Naqvi, and T. Altaf, “Implementation of Cooperative Spectrum Sensing Algorithm using Raspberry Pi,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 12, pp. 363–367, 2016, doi: 10.14569/ijacsa.2016.071247.

[7] H. Yilmaz Birkan, T. Tugcu, F. Alagöz, and S. Bayhan, “Radio environment map as enabler for practical cognitive radio networks,” IEEE Commun. Mag., vol. 51, no. 12, pp. 162–169, 2013, doi: 10.1109/MCOM.2013.6685772.

[8] A. Selim, F. Paisana, J. A. Arokkiam, Y. Zhang, L. Doyle, and L. A. DaSilva, “Spectrum Monitoring for Radar Bands Using Deep Convolutional Neural Networks,” 2017 IEEE Glob. Commun. Conf. GLOBECOM 2017 - Proc., vol. 2018-Janua, pp. 1–6, 2017, doi: 10.1109/GLOCOM.2017.8254105.

[9] L. Gavrilovska et al., “Enabling LTE in TVWS with radio environment maps: From an architecture design towards a system level prototype,” Comput. Commun., vol. 53, pp. 62–72, 2014, doi: 10.1016/j.comcom.2014.07.008.

[10] M. Gail, K. Krickeberg, J. M. Samet, A. Tsiatis, and W. Wong, Statistics for Biology and Health Series Editors. 2010.

[11] G. Biagetti, P. Crippa, L. Falaschetti, G. Tanoni, and C. Turchetti, “A comparative study of machine learning algorithms for physiological signal classification,” Procedia Comput. Sci., vol. 126, pp. 1977–1984, 2018, doi: 10.1016/j.procS.2018.07.255.

[12] S. Khadim, A. Waqar, A. Zeb, I. Khan, and I. Hussain, “Smart Cognitive Cellular Network,” Int. J. Future Gener. Commun. Netw., vol. 10, no. 12, pp. 23–34, 2017.

[13] A. Waqar, S. Khadim, A. Zeb, S. Amir, and I. Khan, “A Survey on Cognitive Radio Network using Artificial Neural Network,” Int. J. Future Gener. Commun. Netw., vol. 10, no. 11, pp. 11–18, 2017.

[14] I. Khan, S. Wasi, A. Waqar, and S. Khadim, “Comparative analysis of ANN techniques for predicting channel frequencies in cognitive radio,” Int. J. Adv. Comput. Sci. Appl., vol. 8, pp. 296–303, 2017.

[15] D. Learning, 技術者が知っておきたい Deep Learning の基礎と 組込みでの利用 ～ 今さら聞いてください Deep Learning ～, vol. 26, no. 7553. 2016.

[16] D. Han et al., “Spectrum sensing for cognitive radio based on convolution neural network,” Proc. - 2017 10th Int. Congr. Image Signal Process. Biomed. Eng. Inform. CISP-BMEI 2017, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/CISP-BMEI.2017.8302117.

[17] M. A. Khan, A. Z. Shaikh, S. Naqvi, S. Khadim, and T. Altaf, “Deep Learning Enabled Spectrum Sensing Radio for Opportunistic Usage,” IJCSNS, vol. 19, no. 11, p. 179, 2019.

[18] S. Zheng, S. Chen, P. Qi, H. Zhou, and X. Yang, “Spectrum Sensing Based on Deep Learning Classification for Cognitive Radios,” pp. 1–7, 2019.

[19] A. Sahai, N. Hoven, and R. Tandra, “Some Fundamental Limits on Cognitive Radio,” Allerton Conf. Control Commun. Comput., pp. 1662–1671, 2004, doi: 10.1.1.123.5645.

[20] C. Liu, J. Wang, X. Liu, and Y. C. Liang, “Deep CM-CNN for Spectrum Sensing in Cognitive Radio,” IEEE J. Sel. Areas Commun., vol. 37, no. 10, pp. 2306–2321, 2019, doi: 10.1109/JSAC.2019.2933892.

[21] W. Rawat and Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural computing. MIT Press Journals, 2017.

[22] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT press, 2016.

[23] A. Botalb, M. Moinuddin, U. M. Al-Saggaf, and S. S. Ali, “Contrasting Convolutional Neural Network (CNN) with Multi-Layer Perceptron (MLP) for Big Data Analysis,” in 2018 International Conference on Intelligent and Advanced System (ICIAS), 2018, pp. 1–5.

[24] J. Bernal et al., “Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review,” Artif. Intell. Med., vol. 95, pp. 64–81, 2019.

[25] N. J. Tustison, B. B. Avants, and J. C. Gee, “Learning image-based spatial transformations via convolutional neural networks: a review,” Magn. Reson. Imaging, 2019.

[26] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, “Medical image analysis using convolutional neural networks: a review,” J. Med. Syst., vol. 42, no. 11, p. 226, 2018.

[27] T. J. Brinker et al., “Skin cancer classification using convolutional neural networks: systematic review,” J. Med. Internet Res., vol. 20, no. 10, p. e11936, 2018.

[28] M. T. McCann, K. H. Jin, and M. Unser, “Convolutional neural networks for inverse problems in imaging: A review,” IEEE Signal Process. Mag., vol. 34, no. 6, pp. 85–95, 2017.