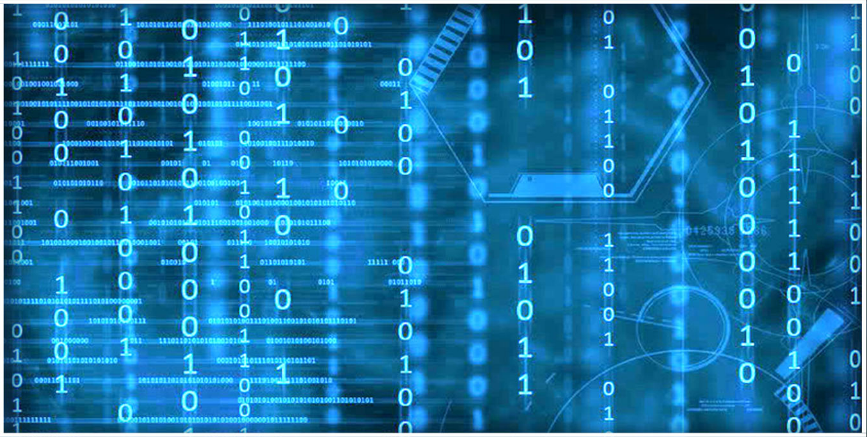
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 10 Issue 9

September 2019



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

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 10 Issue 9 September 2019
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

CONTENTS

Artificial Intelligence Chatbots are New Recruiters

Nishad Nawaz[1], Anjali Mary Gomes[2]
Department of Business Management, College of Business Administration, Kingdom University, Riffa, Bahrain

***Abstract*—The purpose of the paper is to assess the artificial intelligence chatbots influence on recruitment process. The authors explore how chatbots offered service delivery to attract and candidates engagement in the recruitment process. The aim of the study is to identify chatbots impact across the recruitment process. The study is completely based on secondary sources like conceptual papers, peer reviewed articles, websites are used to present the current paper. The paper found that artificial intelligence chatbots are very productive tools in recruitment process and it will be helpful in preparing recruitment strategy for the Industry. Additionally, it focuses more on to resolve complex issues in the process of recruitment. Through the amalgamation of artificial intelligence recruitment process is increasing attention among the researchers still there is opportunity to explore in the field. The paper provided future research avenues in the field of chatbots and recruiters.**

Keywords—Artificial intelligence; chatbots; recruitment process; candidates experiences; employer branding tool; recruitment industry

I. Introduction

In the new phenomenon of information technology and human resource management decades of decades have observed that, the embedded information technology and human resource management in new term as a human resource information system, digital human resource management, automation human resource, enterprise resource planning IOT (Internet of Things), data mining, [1], [2], [3], [4] and freshly artificial intelligence added to the old wine. Again, we can spot the vigor of information technology innovations in business. The latest trouble shooters (solutions) developed by technology to the complex issues of all the various functions of organizations are drawing more attention of the managers from different departments, areas and domains, not excluding the human resource department.

The term artificial intelligence (AI), commonly used for software, machines, system and computers. First time, in the era of industrial revolution, Rossum's Universal Robots (R.U.R) brought into picture by Czech Karel and it is named as ROBOT. But, in the case of artificial intelligence term has been introduced by John McCarthy (1956) appeared in the academic conference and explained the term, as he suggested artificial intelligence (AI) will contribute in future in the following specializations, like applied science, psychology, medical sciences, linguistics, biology, engineering and interdisciplinary programs.

Firstly, we need to know, why HR emerging artificial intelligence (AI), because companies want to extend their business operations to gain profit and new ventures across the globe, it is possible via new technology. Therefore, the organizations aiming to adopt automation process across the functional areas, this will minimize the time and effort of human resources, in other words artificial intelligence (AI) will replace human routine work, enforced them to generate strategies and become craft in the domain.

Artificial intelligence (AI) chatbots developed to make messages to provide assistants to the consumers for 24/7, to answer all queries and acting like FB messenger, webchat, but the competitive environment enthusiastically looking for new added features in artificial intelligence (AI) chatbots to handle all the raised complex problems, therefore artificial intelligence (AI) chatbots are much demanded in chatbot market. Additionally, chatbots present organization to be data driven and pivotal in the success of the business.

The paper is structured as follows, the literature is reviewed, the purpose of the study and then discussion is presented. The paper dismisses with a conclusion and ideas for future research studies.

II. Literature Review

According to [5] recruitment process enhance quality with AI, it will assist employers to select suitable candidate with in a second to ensure whether the candidate is suitable or not. AI has constantly developed over the time to provide deeper insights. This will ensure the organizations not only hire, the right candidate for the organization but also with right skill. [6] main aim of the study is to explain digital technologies influence HR organization. This paper throws light on how digital technologies have reconfigured the HR organization as well as digital technologies transformation and support to the organizational effectiveness, talent strategy execution, succession planning, monitoring operations, transactional case monitoring, recruitment process (robotics automation, artificial intelligence (chatbots), [7], workforce planning, analytics, employee banding management, rewards and engagement, employee relations and effectiveness of organization and finally concluded that digital technology forcing HR process and organization structure into the transformation and [8] through AI can identify right talent leaders, they may easily procure deep insights of training needs, how to cradle time consuming in screening the resumes, unbiased candidate screening and helpful in analyzing personality traits and finally concluded that conduct onboarding via chatbots, this will enhance workforce experience. AI is significantly improving in HR functions, especially in performance management. Traditionally HR managers are evaluating employees performance once in a year, but AI-driven performance management removing unwanted delay in measurement of performance via real time points, face to face communication, chatbots and these technologies providing daily performance, this will prioritize retention across HR [9].

Artificial intelligence understands the human language and not only command but learning from human conversion, transforming like intelligent agent, chatbot is a computer program the conversations through an auditory or methods of textual via natural language processing (NLP), natural language understanding (NLU) and natural language generation (NLG) for interacting humans like ALICE. [10].

An automated mediator stimulate job-seeker to ask questions directly with recruiter about the salary, incentives, leave facilities, FAQ, workforce diversity, complex queries and other related questions [11]. Another piece of work done in this direction chatbots eliminate the routine work in the recruitment process, the RobRecruiters chatbots are automating end to end recruitment process and doing regular process of attendance tracking, goal tracking, reviews of performance, surveys related employees, balance leaves and other activities, enables the HR managers and HR team to move towards success to place organization in top in digitalized era. [12].

(Adams, 2018) chatbots are transforming and assisting in recruitment process to know candidate experience, for effective communication between candidate and recruiter, questions and answers, in identification of qualified candidate and to schedule conversation and finally chatbots taking all requirements from the candidate before his/her entry into the organization. The modern recruiters having more additional responsibilities to have strategies to meet the assigned business goals, keep tracking of competitors, keeping realistic challenges via benchmark, tracking of employees' satisfaction level in all the stages, for this chatbots are useful, because there is a daily conversation database, this will be useful to resolve complex issues Joshi (2019).

The artificial intelligence empowered chatbots to execute human conversation in messaging, the unique way of using words, shorthand, emotions, [13] at the end it will work on the basis of natural language to support conversation process [14], the studies are rare in the field of recruitment process. The authors claim that our understanding of artificial intelligence in recruitment process would benefit more intensive by across all levels of recruitment process. In order to develop better understanding of the recruitment process in human resources organizations in general, at the same time to have effectiveness in different areas of recruitment. Therefore, the authors have proposed the study of artificial intelligence chatbots influence in recruitment process.

III. Methodology

The present study is completely based on literature reviews. The collected literature provided basic understanding of artificial intelligence chatbots and its flow in the recruitment process. To survive this purpose author's selected secondary data such as journals, websites and reports to develop the entire paper.

Additionally, the authors have used library database such as Scopus, ProQuest, EBSCO, Science Direct, Emerald, Elsevier, Taylor and Francis, Springer and Wiley inderscience. The main keywords used for the research include, artificial intelligence, chatbots, recruitment, recruitment process and Internet search engines through Google, Google scholar were utilized to identify and access the relevant working papers, reports, blogs and presentations were used to ensure comprehensive coverage of the literature.

IV. Results

The results identified that there is an increase in the technology development in human resource management, especially in the recruitment process that will have more influence in the future. In past years companies used various technologies for the recruitment such as social media, employee portals, job portals, internal and external networks, social networking, peer referral, emails, mobile messages, cell phone communications are used to attract star performers, best talents for their teams to perform better in the organisation.

In the present transparent digital era, the recruiters, are more experienced because the candidates are more tech-savvy in the mobile environment, and it reinforces the employment brand, an organization employment brand is extremely important, it will attract best performers into their talent pool. Therefore, organizations always want to manage their employment brand to pull candidate towards them.

There are many new implications, solutions and innovative ideas around cognitive technologies, for instance, artificial intelligence, natural language processing, natural language understanding, natural language generation, machine learning, predictive algorithms and robotics process automation, after introduction of Olivia, the chatbots become very popular in the recruitment market and it is providing the guidelines to the candidates, answer their questions. Moreover, facilitating sequence questions to the applicant to answer.

The above technologies adopt smart methods for collecting data of the candidate to make progress in various techniques [15] to identify possible candidate to apply, responding to the unsuccessful candidate, screening the candidate, in sending job offer, and bringing selected candidate into company [16] in other way connecting & collecting all the information of employee into single database and it reveals a new insights for better candidate profile to hire and improve the effectiveness of the recruitment process [17].

Against this reflection, the aim is to study AI chatbots impact in the recruitment process. More precisely, the present study attempts, AI chatbots, significance across the recruitment process function. These are discussed below.

V. Discussion

A. Simplify the First Stage

The present chatbots can accomplish a lot. They can interpret resumes and request elucidations. Interacting with applicants one to one, instant messaging conversations on platforms like Facebook Messenger and text messages, chatbots can get some information about the applicant's experience, answer usual inquiries, and gather a wide range of data and request for a human selection representative to analyze.

Once the application is received screening these applications is an incredibly tedious procedure. Recruiters generally affirm which applicants are appropriate amid the first round of pre-screening calls; while we do know this works, this can be a long procedure which needs revising. Chatbots are an extraordinary arrangement which can make this phase of the evaluation unmistakably increasingly effective! By conveying a text message to every potential applicant that prompts a progression of short, pre-characterized questions, the appropriate applicants can be effectively be sifted through from the unsatisfactory.

Recruiters' can convey several texts in a few minutes and get reply speedily, as compared to the days or even weeks that telephone calls and messages may take.

B. Get the Right Data at the Right Time of the Right Candidate

All organizations dream of updating candidates' database every day, replying to clients faster as well as creating a long term relationship with the candidates. Its not a dream anymore, it is now possible with chatbot. Companies can deploy a chatbot connected to the database to regularly update it. Companies can check their database information which is a key influence. With individual and automated discussions companies can refresh applicants' accessibility, their present position, their mobility or even a new certification. Companies can add value to their database by deploying a chatbot within the preferred audience which can add value to the database.

C. Qualifying Candidates

Chatbots can perform an excellent job of filtering out the good from the bad by asking questions to the applicants related to skills, qualifications, and past experiences which can be otherwise a tedious and time consuming task for the recruiters. It can then effectively rank and qualify a whole group of candidates in terms with the required criteria of the organization.

With all the extensive administrative tasks that come with hiring a candidate, Chatbots can take off a lot of the load by setting up inevitable calls and scheduling meetings keeping in mind both the parties. Apparently, these stages of the process require the human but chatbot ensures all requirements are addressed before humans take over.

D. Get more Qualified Appliers into Job Offers

Companies are aggressively forwarding emails to their database to get more job applicants however the achievement rate isn't as high as anticipated. With a chatbot on messaging apps companies can draw in the applicants database and push them towards the right job at the opportune time. Applicants can apply without any difficulty through the chatbot.

There is no signing in required to go on a website nor have they to go through tedious application process. The job applicants will receive job offers on their messaging apps rather than receiving it through emails. At that point, they can apply for them without changing to another application or site.

E. Increased Numbers of Applications

Due to the impact of social media, facebook recruitment via facebook groups, posts are becoming extremely popular in attracting new candidates. However, the problem is to persuade the applicant to click on the company's career page and submit their application as well. The solution is very simple in the chatbots. By using an automated facebook messenger recruitment chatbot, any potential applicant could be prompted to opt-in for job alerts and provide them with their facebook profile information, as well as showing them job openings, information about the process of application or even videos about working at the company. The possibility of submitting the application by the candidate increases many folds if they are engaged right from the point of initial interest as they've already had contact and established a rapport with the company

The shocking fact is that as per Jobvite Recruiting Funnel 2017 only 8.52 per centage of visitors to a career site literally complete their application. Which means more than 91 per centage of potential candidates just leaves the site without leaving any information for the company, hence even if the company wants to contact them in the future they cannot due to lack of information. This can be a very big problem to companies trying to attract talented and skilled people for their job vacancies. To tackle this issue a chatbot can help by engaging candidates through a messenger, replying to any questions regarding any misgivings they have which might daunt them from submitting any application, it can also give reminder to candidates to complete all the information in case it is insufficient.

F. Question and Answer (FQA)

Before applying the applicants needs to know about the job, the company and various requirements and it is extremely frustrating if a candidate has to look for all these answers on an FAQ page. This can be completely changed with implementation of chatbots. Companies can use chatbots to answer FAQ by recognizing keywords mentioned by the applicant. It's crucial to make the answers understandable and informative which will enable the applicant to acquire knowledge and clear understanding promptly leading to applicant's satisfaction.

It is less intimidating communicating to a chatbot where an applicant can be just themselves and stay calm and composed. Chatbots can respond to the applicants in a user-friendly way and if the answer is not known it will refer the applicant to the right person to create a contented experience. This is the age of data driven decisions hence a chatbots could be linked to a platform which can gather important data. This type of platform will enable the company to tag how many times a particular question has been asked and what candidates want and are looking for. Chatbots are great assistance to recruiters with their prompt replies and instant availability.

G. Responding to the Unsuccessful

According **to** Eyal Grayevsky, Mya "applicants who never get to hear back after submitting their applications from the recruiter are approximately 85 per centage" this results in poor applicant experience and a wrong impression of the company. This can also lead to loosing potential candidates which might be a critical issue, as they will not be motivated to reapply in the same company, who might be better suited for another position in that company in the future.

To remain competitive companies must attract the best talent and skilled professionals. The companies can be successful in this only if they are seen as someone who treats people with respect. With numerous numbers of applications for one job position it is practically not possible for recruiters to personally reject everyone, hence it makes the chatbots the most suitable option. Chatbots can promptly respond to the applicants once a decision is taken. The applicants acknowledge this kind of swiftness and at the same time they don't develop any negative feeling towards the organization as they don't go through the anxiety of waiting time.

H. Screening Candidates Application

Chatbots can initiate a conversation with the applicants once they apply on a company's job site. Chatbots may ask many questions while communicating with them. The questions can range from asking about work experience, previously where they have worked, their areas of interest and so on. When this process is over the chatbots assess the applicant for relevancy of the unfilled position. The decision is taken by the recruitment chatbot based on the conversation that took place, resume details and assessing the job requirement if the candidate is best bit for the job.

I. Assess Candidates' Recruitment Experience

Companies must be aware about how the applicants feel regarding their recruiting experience. Its important to differentiate from the competitors which is possible by getting feedback and this way a strong relationship can be built. To enhance the retention rate of the company its important to ask the applicants about how they feel. The chatbots can assist in checking candidate's feedback after interviews and get deep insights about how companies create a bond with people within the talents pool and useful in recruitment strategy as well.

J. Interview Scheduling of the Candidates

Scheduling the interview with the candidate is another time-consuming task. Intellectual chatbots are capable of accessing the calendar of the recruiters to check if they are available and then schedule the date and time for the relevant candidate. In today's time it's not very much effective to just make calls to the candidate as most of them don't answer to unrecognized phone numbers. Also, it might be bothersome to call the candidate when they are working with their current company or they have to request for a convenient time for both the parties. This whole process could be very time consuming. But for a chatbot this task will not be tedious, and they are great in this type of repetitive tasks.

K. Enhance Candidates On-Boarding

The very first step for a candidate is on-boarding in an organization which is also a long and key process. To smoothen the process the organizations can deploy a chatbot to deliver significant information at the right time to the newly recruited employees.

L. Candidate Experience

It's nothing astonishing that one will need to get a chatbot engaged with this piece of the procedure. With a new job opening numerous numbers of applicants will be pulled in a high volume, it very well may be a major errand for a human to deal with. Proficiently that is, chatbot can intercede with faster reply and speed up the procedure.

To emerge of the commotion, recruiters ought to possibly change their recruitment strategy and make it more applicant driven. The time that applicants take to send resumes and recruiters hit them up ought to be short. Chatbots can get this going. Chatbots can return to the applicants quickly, making the applicants as well as the recruiters contented on that front. With a correct approach companies can make the job search experience where conversation is concerned a more natural process.

In the present day's marketing recruitment the experience of the applicants is becoming extremely important. The applicants' journey must be effortless, straightforward, and inviting and this must be ensured by the talent acquisition leaders. Recruitment chatbots can connect with applicants in a conversational trade as well as answer recruiting FAQs, a boundary that prevents numerous applicants from applying. With recruiting web chat arrangement like career chat, applicants can study the organization and draw in recruiters in live agent modes and computerized modes.

M. Candidate Experience Feedback

Huge number of applicants have poor encounters while presenting their applications and resumes on the web. As a rule, they don't get advised about whether an organization they have connected for has gotten their documents. This vulnerability combined with the distress of sitting tight for input makes a negative impact on an organization's validity.

Through the incorporation of selecting artificial intelligence, many applicants can be obliged promptly and advised with the outcomes of their interview once it's finished. This brings down their nervousness and encourages them proceed onward to discovering openings that are more appropriate for them. Organizations that make HR bots are looking for better approaches to improve their process to fulfill the requirements of clients around the world.

Job pal perceives the requirement for organizations to begin connecting with applicants the minute they apply for work/ job. With that, they have fabricated AI-controlled chatbots to mechanize the correspondence among employees and applicants accordingly accelerating the procuring procedure.

N. Notation Feedback

Structured inquiries that accompany predefined answers will aid in real-time feedback. Instead of using forms the applicants can have the privilege of dynamic interactive conversational interface to share their feedback. In the *point of improvement,* Questions from the employers can be put forward to candidates such as, "where do you think we need to improve" to get an understanding of areas of improvement. The feedback of the applicants could be shared with the recruiters to initiate necessary actions and filter the process to best fit the human capital needs. Using these observations the companies can make sure of providing an engaging experience to candidates and help HRs in eradicating human bias and the possibility for any error.

O. Building Rapport

Correspondence is significant amid the recruitment lifecycle. That is for both the job applicant and recruiter. The informing perspective accessible with chatbots makes it a simple and natural methodology. The chatbots are fueled through guidelines. They're savvy. A chatbot will pose decision inquiries and ensure the discussion is appropriate. Continually recording subtleties prepared for the recruiter to meet the job applicant.

For instance, Mya is an excellent case of a viable chatbot. Planned just for enrollment, it talks and draws in with job applicants through a messaging app. It records every one of the information and answers questions asked by the job applicant. Along these lines, it builds profitability as you can qualify an expansive number of job applicants.

P. Employer Branding Tool

Employer branding helps the company to be outstanding if on the career page of the company have a chatbot. To increase the retention rate of the employees it's very important to engage the right candidates. It's unfortunate and sad that employer branding is often extravagant and time-consuming because it includes building content such as article, blog posts, employee interviews, videos, etc. It continues with marketing all these content to the pool of initial candidates.

Though a lot of organizations are excellent in building employer branding content and posting that content on their career site but unfortunately not many of the candidates are seeing it, but with chatbots it is very much possible to send automated messages to the current subscribers, with the most recent blog posts, videos, etc. all-over multiple messaging platforms.

By giving a personality to the recruiting chatbot some of the companies are taking their branding to the next level. The companies also make sure that the culture and values of the organizations are well aligned to the chatbot and its interactions and promote the employer brand by making sure there is regularity in messaging during the recruitment process.

VI. Conclusion

Indeed, artificial intelligence has it existence in recruitment process and smartly work like human brain in various complex situations. Digital era gains more attention and importance in automating recruitment process when compare to traditional system of recruitment. Artificial intelligence provides smooth process by conducting screening of CV, responding with automated message. Adoption of artificial intelligence has helped in increasing to build momentum, by reducing tedious routine work of recruiters through AI chatbots.

This technology has been successful in simplifying the work and collect related information in candidates experience, building relationship, to answer questions, identification of right candidates, on-boarding, increasing of applications, scheduling interviews and so on will be align with recruiters to have smooth functioning of process to make business success. The study has some limitations. The aim of the study is to know chatbots influence on recruitment process, how it is improving performance of the recruiters. Additionally, it was not studied why recruitment industry is very keen in adopting chatbots in recruitment in the present study.

VII. Further Research Directions

The papers discussed available facets of technology used in recruitment industry for recruitment process. The present study presented fundamental base for future research work in field of AI chatbots and recruitment process. In the future the researchers can take privilege to add to new literature to above discussed topic. The researchers can conduct empirical studies with different perspective. The enthusiastic researchers can take comparative studies before introduction AI chatbots and after in the recruitment process, industry wise, HR designation wise also possible to study. The authors predicting that present study act like "food for brain" for researchers in the field of artificial intelligence.

References

[1] Ernst and Young, "The new age : artificial intelligence for human resource opportunities and functions," USA, 2018.

[2] D. L. Stone, D. L. Deadrick, K. M. Lukaszewski, and R. Johnson, "The influence of technology on the future of human resource management," Hum. Resour. Manag. Rev., vol. 25, no. 2, pp. 216–231, Jun. 2015.

[3] T. Bondarouk, E. Parry, and E. Furtmueller, "Electronic HRM: four decades of research on adoption and consequences," Int. J. Hum. Resour. Manag., vol. 28, no. 1, pp. 98–131, Jan. 2017.

[4] N. Nawaz and A. M. Gomes, "Human resource information system: a review of previous studies," J. Manag. Res., vol. 9, no.3, p.92, Jul. 2017.

[5] N. Nawaz, "A comprehensive literature review of the digital HR research filed," Information Knowl. Manag., vol.7, no.4,pp.15–20, 2017.

[6] J. Gikopoulos, "Alongside, not against: balancing man with machine in the HR function," Strateg. HR Rev., vol.18, no. 2, pp. 56–61, Apr. 2019.

[7] A. DiRomualdo, D. El-Khoury, and F. Girimonte, "HR in the digital age: how digital technology will change HR's organization structure, processes and roles,"Strateg. HR Rev,vol.17,no.5,pp.234–242,Oct. 2018.

[8] N. Nawaz, "Robotic process automation for recruitment process," Int. J. Machanical Eng. Technol., vol. 10, no. 04, pp. 88–91, 2019.

[9] E. He, "Can artificial intelligence make work more human?," Strateg. HR Rev., vol. 17, no. 5, pp. 263–264, Nov. 2018.

[10] B. Buck and J. Morrow, "AI, performance management and engagement: keeping your best their best," Strateg. HR Rev., vol. 17, no. 5, pp. 261–262, Nov. 2018.

[11] P. Nikhila1, G. Jyothi2, K. Mounika3, M. C. Kishor, K. Reddy, and R. Murthy, "Chatbots using artificial intelligence," J. Appl. Sci. Comput., vol. 6, no. 2, pp. 103–115, 2019.

[12] K. Kuksenok and N. Praß, "Transparency in maintenance of recruitment chatbots," in Where is the Human? Bridging the Gap Between AI and HCI Workshop , 2019, pp. 1–4.

[13] B. Sheth, "Chat bots are the new HR managers," Strateg. HR Rev., vol. 17, no. 3, pp. 162–163, Apr. 2018.

[14] A. Følstad and P. B. Brandtzæg, "Chatbots and the new world of HCI," Interactions, vol. 24, no. 4, pp. 38–42, 2017.

[15] J. Hill, W. Randolph Ford, and I. G. Farreras, "Real conversations with artificial intelligence: a comparison between human-human online conversations and human-chatbot conversations," Comput. Human Behav., vol. 49, pp. 245–250, 2015.

[16] E. Parry and V. Battista, "The impact of emerging technologies on work: a review of the evidence and implications for the human resource function," Emerald Open Res., vol. 1, p. 5, Jan. 2019.

[17] P. Cappelli, P. Tambe, and V. Yakubovich, "Artificial intelligence in human resources management: challenges and a path forward," SSRN Electron. J., Nov. 2018.

FPGA Implementation of RISC-based Memory-centric Processor Architecture

Danijela Efnusheva[1]
Computer Science and Engineering Department
Faculty of Electrical Engineering and Information Technologies
Skopje, North Macedonia

***Abstract*—The development of the microprocessor industry in terms of speed, area, and multi-processing has resulted with increased data traffic between the processor and the memory in a classical processor-centric Von Neumann computing system. In order to alleviate the processor-memory bottleneck, in this paper we are proposing a RISC-based memory-centric processor architecture that provides a stronger merge between the processor and the memory, by adjusting the standard memory hierarchy model. Indeed, we are developing a RISC-based processor that integrates the memory into the same chip die, and thus provides direct access to the on-chip memory, without the use of general-purpose registers (GPRs) and cache memory. The proposed RISC-based memory-centric processor is described in VHDL and then implemented in Virtex7 VC709 Field Programmable Gate Array (FPGA) board, by means of Xilinx VIVADO Design Suite. The simulation timing diagrams and FPGA synthesis (implementation) reports are discussed and analyzed in this paper.**

Keywords—FPGA; memory-centric computing; processor in memory; RISC architecture; VHDL

I. Introduction

The growing technological progress over the last several decades has caused dramatic improvements in processor performances, providing speed-up of processor's working frequency, increased number of instructions that can be issued and processed in parallel, [1], [2], multithreading, pre-fetching, etc. According to Moore's law, [3], [4], the integrated circuits production technology has enabled doubling of the number of transistors on a chip every 18 months, which resulted with the creation of multi-core processors over the last decade. This trend of processor technology growth has brought performance improvements on the computer systems, but not for all the types of applications, [5]. The reason for such divergence is due to the bottleneck problem in the communication between the processor and the main memory (which is by default placed out of the processor), caused by the growing disparity of memory and processor speeds, [6]. Therefore, we can say that not long ago, off-chip memory was able to supply the processor with data at an adequate rate. Today, with processor performances increasing at a rate of about 70 percent per year and memory latency improving by just 7 percent per year, it takes a dozens of cycles for data to travel between the processor and the main memory, [7], [8], which is basically placed outside of the processor chip.

The computer systems that are used today are mainly based on the Von Neumann architecture, [9], which is characterized by the strict separation of the processing and memory resources in the computer system. In such processor-centric system the memory is used for storing data and programs, while the processor interprets and executes the program instructions in a sequential manner, repeatedly moving data from the main memory in the processor registers and vice versa, [1]. Assuming that there is no final solution for overcoming the processor-memory bottleneck, modern computer systems implement different types of techniques for "mitigating" the occurrence of this problem, [10], (ex. branch prediction algorithms, speculative and re-order instructions execution, data and instruction pre-fetching, and multithreading, etc.). In fact, the most applied method for approaching data closer to the processor is the use of multi-level cache memory, as faster, but smaller and more expensive data storage than the main memory. Regarding that, the research stated in [11] discusses that the capacity and the area of on-chip cache memory have shown steady growth, as a result of the increased number of on-chip processing cores, which have imposed even greater requirements to the memory system. For example, up to 40% of the chip area in Intel's 65nm processors is occupied by caches, [12], used solely for hiding the memory latency.

Despite the grand popularity of cache memory used in the modern computer systems, we should note that each cache level presents a redundant copy of the main memory data that would not be necessary if the main memory had kept up with the processor speed. According to [13], cache memory causes up to 40% increase of the system's energy consumption, because it adds extra hardware resources and requires the implementation of complex mechanisms, [14], for maintaining memory consistency. Besides that, the misses in cache memory bring unpredictability in the timing of the program, which is not very suitable for real-time systems.

On the other hand, the development of some powerful processor architectures, such as vector, [15], wide superscalar, [16], VLIW (very long instruction word), [17], and EPIC (explicitly parallel instruction computing), [18], did not achieve the expected success, because of their inability to provide fast and high throughput access to the memory system. Considering the difference between the processor and the memory speeds, we believe that the relatively small number of fast GPRs in the processor is the major obstacle for achieving high data throughput. This is mainly expected in the case of executing a program that works with larger data set that needs to be placed into the processor for a short time, but there are not enough free registers. Examples for such applications are:

processing of data flows, calculating vast logical-arithmetical expressions or traversing complex data structures, etc. In such cases, the high speed of access to GPRs doesn't bring many advantages, because the complete set of required data cannot be placed into the register set at the proper time. Therefore, the author of [19] purposes a register-less processor which uses only cache memory (inside and outside of the processor) to communicate with the main memory. Additionally, the authors of [20] and [21] suggest the use of Scratchpad memory as a small software-managed on-chip memory that is separated of the cache memory and can be accessed in a single proc. cycle.

A few decades ago, in the 1990ties, some researches predicted that the memory behavior would be preponderant over the global performances of the computer system. Their proposals suggested the design of "smart memories" that will include processing capabilities. Therefore, several memory-centric approaches of integrating or approaching the memory closer to the processing elements have been introduced, including: computational RAM, [22], Mitsubishi M32R/D, [23], DIVA, [24], Terasys, [25], intelligent RAM, [26] - [28], parallel processing RAM, [29], DataScalar, [30], and an intelligent memory system, known as active pages model, [31]. Within these memory-centric systems, the processor can be realized as some simple RISC or complex superscalar processor and may contain a vector unit, as is the case with the Intelligent RAM.

The aim of this paper is to develop a novel RISC-based memory-centric processor architecture, which suggests an integration of processor and memory on the same chip die and proposes removal of general-purpose registers and cache memory (inside and outside of the processor) from the standard memory hierarchy. Contrary to the other memory/logic merged chips, which mostly use the standard memory hierarchy model for data access, the proposed RISC-based memory-centric processor provides direct access to the data into its on-chip memory (without the use of explicit LOAD and STORE instructions) and includes specialized control unit that performs 4-stage pipelining of instructions, allowing every (arithmetical, logical, branch and control) instruction to be completed in a single tact cycle. If this logic is manufactured as an ASIC (application-specific integrated circuit) it cannot be reused for further extensions, so in this paper we are investigating the possibilities to utilize a reconfigurable hardware platform - Virtex7 VC709 FPGA board, [32]. In that process, we are developing a VHDL model of the proposed RISC-based memory-centric processor, and then we are simulating the functionalities of the proposed processor and analyzing the characteristics and the complexity of its FPGA implementation, by means of Xilinx VIVADO Design Suite. In fact, FPGA technology is very suitable for the purposes of this research since it represents a good compromise between performance, price, and re-programmability, [33].

The rest of this paper is organized as follows: Section II gives an overview of different techniques and methods used to alleviate the processor-memory bottleneck and also discusses several memory-centric approaches of computing. Section III presents the proposed RISC-based memory-centric processor, describing its basic architectural characteristics, including instruction set, addressing modes, pipelining support, data forwarding, access to on-chip memory, etc. Section IV presents simulations and synthesis results from the FPGA implementation of the proposed RISC-based memory-centric processor. The paper ends with a conclusion, stated in section V.

II. Current State

The extraordinary increase of microprocessor speed has caused significant demands to the memory system, requiring an immediate response to the CPU (central processing unit) requests. Considering that the memory price, capacity, and speed are in direct opposition, an ideal memory system cannot be implemented in practice, [2]. Therefore, today's modern computer systems are characterized with hierarchical memory, organized in several levels, each of them having smaller, faster and more expensive memory, compared to the previous level.

The hierarchical approach of memory organization is based on the principle of temporal and spatial locality, [1], [2], and the rule "smaller is faster" which states that smaller pieces of memory are usually faster and hence more expensive than the larger ones. According to that, cache memories have lower access time, but on the other hand they bring indeterminism in the timing of the program, as a result of the misses that can occur during the memory accesses (read or write). This is also confirmed with equations 1 and 2, which give the expressions for computing average memory access time and program execution time, accordingly. The relation between these equations is expressed with the CPI (cycles per instruction) parameter, which value depends on the average memory access time. Therefore, if many misses to intermediate memory levels occur, the program's execution time will increase, resulting in many wasted processor cycles.

Average memory access time =

= Hit time + Miss rate * Miss penalty (1)

Execution time = Instructions number *CPI*Clock period (2)

According to the previous assumptions, we can say that multi-level cache memories can cause reduction of the memory access time, but at the cost of additional hardware complexity, increased power consumption, unpredictable program's timing and extra redundancy in the system. Other techniques for memory latency reduction include a combination of large cache memories with some form of branch predictive speculation, or out-of-order execution, [14]. These methods also increase the chip area and cause extra complexity on both the hardware and software level. Even other more complex approaches of computing like vector, wide superscalar, VLIW and EPIC suffer from low utilization of resources, implementation complexity, and immature compiler technology, [15] - [18]. When it comes to processor architectures, we can say that the integration of multiple cores or processors on a single chip die brings even greater demands to the memory system, increasing the number of slow off-chip memory accesses, [8].

In order to tolerate the memory latency and allow the processor to execute other tasks while a memory request is being served, a separate group of memory latency tolerance techniques was introduced. Some of the most popular methods in this group are multithreading, [2], instruction and data pre-fetching, [1] and non-blocking caches, [34]. In general, the usage of these methods contributes to the "reduction" of the memory latency, but on the other hand it increases the memory traffic, leading to a higher instruction and data rate. As a result of the limited bandwidth on the memory interface, additional latency can be generated.

Besides the previously discussed memory latency reduction and tolerance methods, there are several proposals, which present some modifications into the classic multi-level memory hierarchy and provide nonstandard faster access to the main memory. For example, the author of [19] proposes a register-less processor that performs all the operations directly with the cache memory, organized in several layers (on-chip and off-chip), excluding the explicit use of GPRs. Additionally, the authors of [21] suggest the use of Scratchpad memory as a small high-speed on-chip memory that maps into the processors address space at a predefined memory address range. Opposite to the cache memory, the Scratchpad memory is allocated under software control and is characterized with deterministic behavior, allowing single-cycle access time. This small on-chip memory is mostly used for storing in-between results and frequently accessed data, so it requires developing of complex compiler methods for effective data allocation.

Contrary to the standard model of processor-centric computing (Von Neumann model), [9], some researchers have proposed alternative approaches of memory-centric computing, which suggests integrating or placing the memory near to the processor. These proposals are known as computational RAM, intelligent RAM, processing in memory chips, intelligent memory systems, [22] - [31], etc. These merged memory/logic chips implement on-chip memory which allows high internal bandwidth, low latency, and high power efficiency, eliminating the need for expensive, high-speed inter-chip interconnects, [35]. This makes them suitable to perform computations which require high data throughput and stride memory accesses, such as FFT, multimedia processing, network processing, etc., [28].

The integrated on-chip memory in the merged memory/logic chips is usually implemented as SRAM or embedded DRAM, which is mostly accessed through the processor's cache memory. Although the processing in/near memory brings latency and bandwidth improvement, still the system has to perform unnecessary copying and movement of data between the on-chip memory, caches, and GPRs. Besides that, the processing speed, the on-chip memory size, and the chip cost are limited due to the used implementation technology and the production process. Moreover, it is even a greater challenge to develop suitable compiler support for the system, which will recognize the program parallelism and will enable effective utilization of the internal memory bandwidth.

Having in mind that modern processors are lately dealing with both technical and physical limitations, while the memory capacity is constantly increasing, it seems that now is the right time to reinvestigate the idea of placing the processor in or near to the memory in order to overcome their speed difference, [36] - [38]. A promising approach that targets this problem is presented by the Hewlett Packard international information technology company that suggests novel computer architecture, called the Machine, [39], which utilizes non-volatile memory as a true DRAM replacement. A more detailed study about other proposals for overcoming the processor-memory bottleneck is presented in our previous research, given in [40].

Considering the adjustments of the standard memory hierarchy model, presented in some of the previously discussed approaches (ex. PERL, Scratchpad, Machine), we can say that the extension or revision of their work can be a good starting point for further research. In that process, we can first perceive that the relatively small number of fast GPRs in the highest level of the memory hierarchy is the major obstacle for achieving high data throughput. After that, we can consider that the cache memory is a limiting factor in real-time computing, and is also a redundant memory resource, which adds extra hardware complexity and power consumption into the system. Therefore, our research will continue into the direction of developing a novel RISC-based memory-centric processor similar to PERL, which will provide direct access to the memory that is integrated into the processor chip, without the use of GPRs and cache memory. The proposed replacement of the two highest memory hierarchy levels with an on-chip memory is intended to provide: exclusion of unnecessary data copying and individual or block data transfer into the GPRs and cache memory, a decrease of the capacity of redundant memory resources, simplification of the accesses to memory and removal of complex memory management mechanisms.

III. Design of RISC-based Memory-centric Processor Architecture

As a referencing point for designing the proposed RISC-based memory-centric processor, we make use of a RISC architecture implementation (MIPS), which is widely applied in the embedded industry and additionally is well documented and presented in the leading world's literature in the field of processor architectures. The selected MIPS implementation of a single-cycle pipelined RISC architecture, presented by D. A. Patterson and J. L. Hennessy in [1], is also used as a basis in the PERL processor architecture design. In general, MIPS processor is characterized with: fix-length instructions, simple addressing modes, memory accesses with explicit load and store instructions, hardwired control unit, large GPR set and pipeline operation in five stages (fetch, decode, execute, memory access and write back), as shown in Fig. 1.

According to Fig. 1, a MIPS processor includes: Program counter - PC, Instruction Register - IR, pipeline registers, 32 general-purpose registers, separated instruction and data cache memory, 32-bit arithmetical and logical unit, control unit (marked with blue), and other selecting and control logic (multiplexers, decoders, adders, extenders etc.). Therefore, MIPS operates only on operands found in its local GPRs, requiring frequent data transfer between the memory and the processor's registers, via load and store instructions. In order to provide easier access and manipulation of memory data, this paper proposes a modification and extension of the original pipelined RISC architecture and creation of a novel MEMRISC (Memory Access Reduced Instruction Set Computing) pipelined processor architecture, shown in Fig. 2.

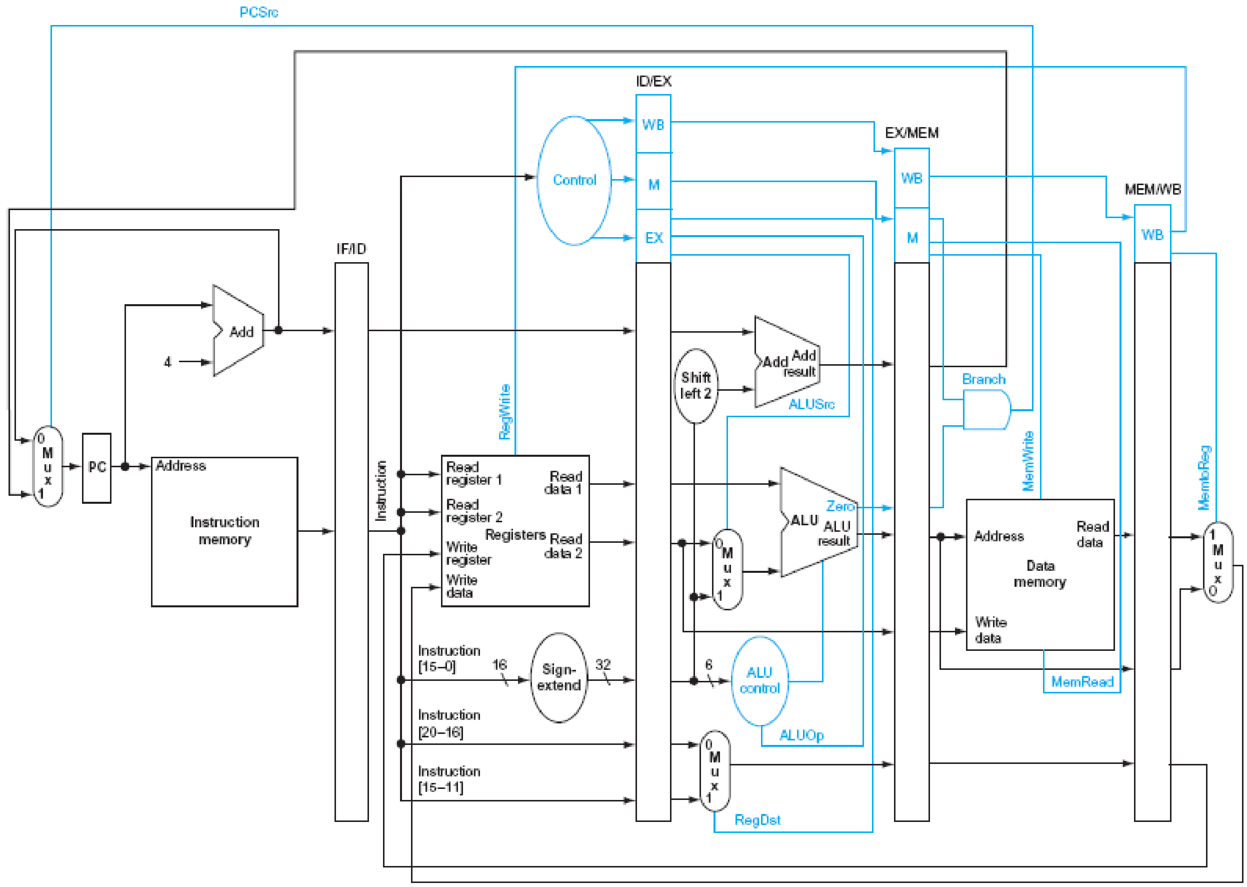


Fig. 1. Pipelined RISC-based MIPS Processor, [1].

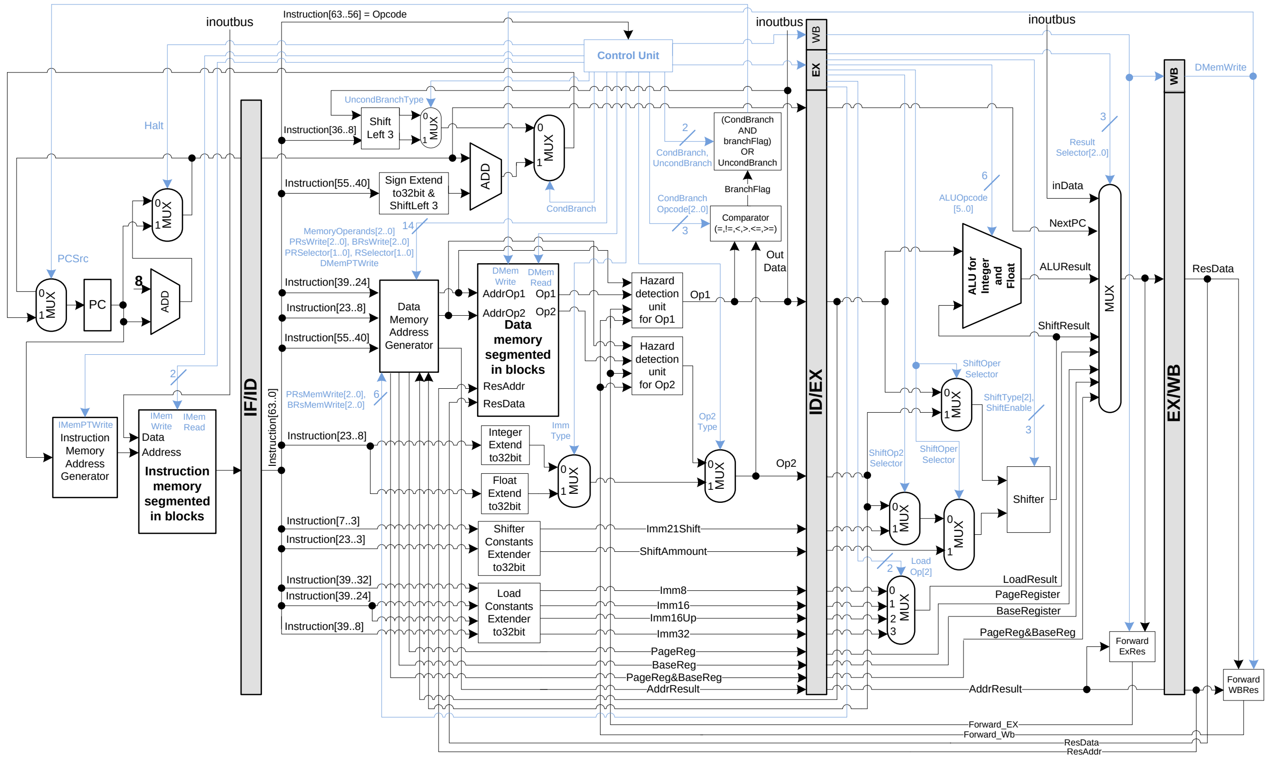


Fig. 2. Pipelined MIMOPS Processor with MEMRISC Architecture.

As shown in Fig. 2, the proposed processor with MEMRISC architecture uses separated on-chip data and program memory, instead of GPRs and on-chip cache memory. This means that the given processor executes all the operations on values found in its on-chip memory, avoiding the unnecessary and redundant data copying and movements, which are performed in the MIPS processor, during (load/store) data transfers. Therefore if the RISC-based MIPS processor is able to execute a million instructions per second, then the proposed processor with MEMRISC architecture would be able to execute a million instructions on memory operands per second, which is the reason why it is called MIMOPS processor in continuation.

The proposed MIMOPS processor excludes the GPRs and the cache memory from the memory hierarchy and thus allows direct and simultaneous access to two sources and one result operand, specified in the instruction. These operands are selected by a specialized memory address generator unit that is used to perform the translation of the input virtual addresses into physical addresses of the paged on-chip memory. Once the operands are read from the on-chip data memory, the operation is executed and then the result is written back to the on-chip data memory. In fact, the MIMOPS processor operates in a 4-stage pipeline (instruction fetch, instruction decode, execute and write back), excluding the MEM phase, and allowing every (arithmetical, logical, branch or control) MIMOPS instruction to be completed in a single tact cycle. The instructions that are supported by the proposed MIMOPS processor are MIPS alike, but the way of their interpretation and execution is slightly different.

Unlike the MIPS processor that is given in Fig. 1, the MIMOPS processor operates directly with the on-chip memory and thus simplifies the access to the operands, the execution of the instructions (pipelining without MEM phase) and the instruction set (removes explicit LOAD/STORE instructions). This way of operation of the MIMOPS processor is managed by a specialized control unit (marked with blue on Fig. 2), which provides support for several addressing modes (ex. direct, immediate, base, PC-direct, and PC-relative addressing). Generally, the memory operands are addressed directly, while the translation of the virtual addresses to physical addresses is performed via specialized hardware support for virtual memory that is implemented inside the MIMOPS processor. This refers to segmentation of the on-chip memory and filling it with virtual pages, and implementation of page translation tables and page replacement mechanisms (ex. FIFO).

The proposed MIMOPS processor implements separated on-chip instruction and data memories that are segmented into M equal-sized physical blocks (for virtual pages). Each of these local memories is organized as an array of N contiguous byte-sized elements, whereas each element has a unique physical address. To provide support for address translation and simultaneous access to the separated instruction and data on-chip memories, the proposed MIMOPS processor implements two dedicated hardware units, called instruction and data memory address generators. These units translate virtual addresses on the fly, performing a look-up in inverted page tables, [14], stored inside the processor's fetch and decode hardware logic, whose contents are managed by the operating system. According to the implemented approach of pipelining, MIMOPS can simultaneously access to a single instruction of an on-chip instruction memory block, and to three operands of up to three on-chip data memory blocks (some operands might be in the same block), as shown in Fig. 3.

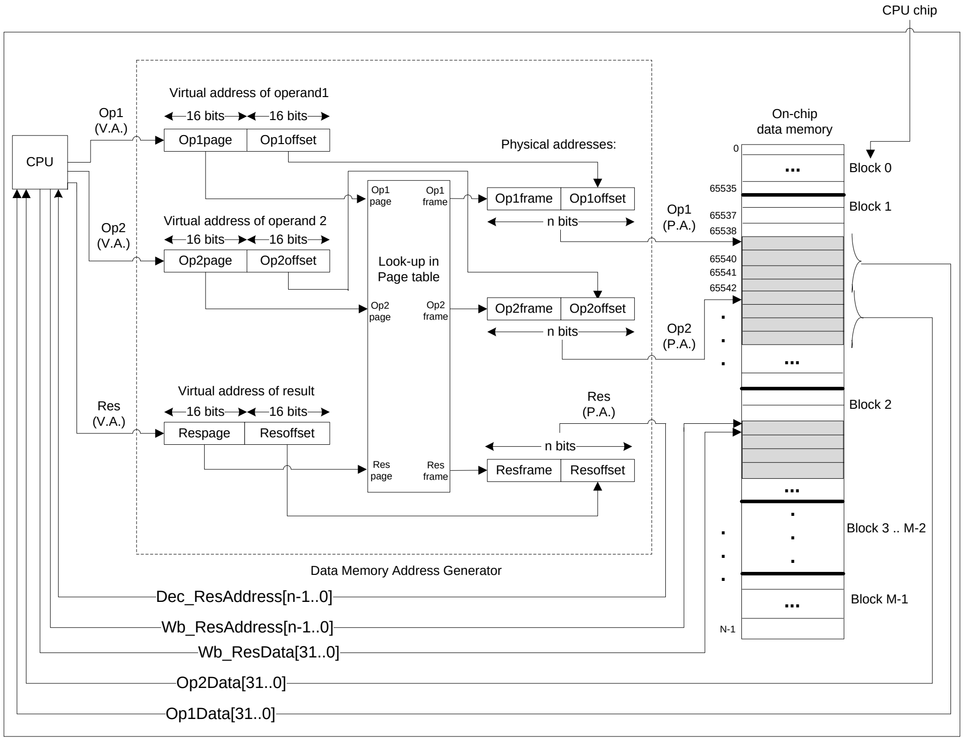


Fig. 3. Virtual Addressing of on-Chip Data Memory.

Fig. 3 shows how the CPU accesses to the on-chip data memory, during the instruction decode pipeline stage. Once the CPU decodes the instruction, it passes three virtual memory addresses (for operand1, operand2, and result) to the data memory address generator unit. This unit performs a look-up in a page table in order to find the appropriate frame numbers for the input page numbers, and thus to generate the effective physical addresses of the two input operands and the result operand. After that, the physical address of operand1 and operand2 are passed to the data memory over the memory bus, while the physical address of the result operand, (Dec_ResAddress), is sent to the next CPU pipeline stage, to be further used during the write-back stage.

According to Fig. 3, the CPU can simultaneously perform two reads and a single write to the on-chip data memory. This is achieved in such a way that the processor fetches two 4-byte (32-bit) data operands, starting at the generated physical addresses of operand1 and operand2, and in parallel stores the received 32-bit result data (Wb_ResData), starting at the physical address (Wb_ResAddress) of the result operand, which is actually passed from the write-back pipeline stage. Similarly to the result data and address forwarding (Wb_ResData, Wb_ResAddress), the fetched operands (operand1 and operand2) are sent to the next CPU pipeline stage, to be further used as input operands for computing some ALU operation in the execute stage.

When it comes to pipelines, it can be noticed that both MIPS and MIMOPS processors provide overlapping of the execution of the instructions, by implementing pipeline registers for every inter-phase (ex. instruction fetch/instruction decode). Besides these similarities, the MIMOPS processor differs from the MIPS processor in many ways, since it allows: reducing of the pipeline stages number by one, finishing the execution of conditional and unconditional branches in the decode pipeline stage and support of data forwarding for overcoming data hazards during parallel instructions execution. Additionally, the MIMOPS processor implements a separate shifter logic that is purposed to generate a second flexible operand for the arithmetical-logical unit (ALU). This is achieved by shifting the second operand by a specific constant value before it is being used by the ALU (this is similar to the ARM - Advanced RISC Machine architecture, [39]). Therefore, the ALU of the MIMOPS processor is able to perform operations over two integer or floating-point input numbers, where the second operand might be previously shifted.

Basically, the instruction set architecture of the proposed MIMOPS processor is RISC-like and includes three types of instructions (M-type, I-type and J-type), organized in four different groups. M-type instructions operate with memory operands placed in the on-chip data memory (similar to registers in R-type MIPS instructions), while I-type and J-type instructions operate with immediate values, whereas J-type instructions are used for unconditional branching. Depending on the function of the instructions, they can belong to arithmetical-logical, shifting, branching or control group. The arithmetical-logical group of instructions includes addition with overflow detection, subtraction, multiplication, integer division (div), modulo division (mod) and AND, OR, XOR and NOT logical bit-wise operations. The shifting group of instructions consists of left and right logical and arithmetical shifts and rotations. The branching group includes instructions for conditional and unconditional change of the program flow. The last group is the auxiliary group, consisting of instructions for program termination and system halt, SET instructions that update the base address units, load instructions for storing 8-, 16- or 32-bit immediate values in the on-chip data memory and IN/OUT instructions for communication with external devices.

The execution of MIMOPS instructions is managed by the control signals generated by the control unit that is specifically defined for the MEMRISC architecture. This unit provides support for several addressing modes, including base, direct, immediate, PC-direct and PC-relative. In addition to the control unit, the MIMOPS processor also includes: arithmetical - logical unit that can operate with integers and floating-point numbers, units for pipelining support, hazard detection unit for overcoming data hazards during pipeline execution of instructions, units that provide hardware support for virtual memory (memory segmentation in blocks, page tables etc), mechanisms for exception handling (ex. incorrect result), I/O (in-/output) control, and additional control and selection logic.

The proposed MIMOPS processor with MEMRISC architecture is expected to save many timing and hardware resources since it removes the complex cache memory management mechanisms and eliminates the use of explicit load and store instructions. Indeed, the MIMOPS processor excludes the many redundant copies of data that occur in GPRs and caches of processors which operate with standard memory hierarchy. This way of operation is very suitable for applications that perform many arithmetical-logical operations over some data set that is accessed with a high degree of locality. Examples of such type of applications are those that perform computations with matrices, such as matrix multiplication programs.

In order to present the performance gains (in terms of speed) of the novel MEMOPS processor with MEMRISC architecture, a comparative analysis between three similar processors is made. This refers to a MIMOPS processor, a register-less PERL processor, and a RISC-based MIPS processor. It is considered that the proposed MIMOPS processor includes on-chip memory with a capacity equal to the amount of cache memory into the MIPS and PERL processors (128KB L1 and 2M L2 cache). The actual analysis measures the execution time of a 32x32 matrix-multiplication program for each of the given processors. The program simulation is done with a MIMOPS instruction-set simulator, explained in [41], a MARS simulator for MIPS, [42] and a special instruction-set simulator for PERL, given in [19].

The results of the analysis are shown in Fig. 4 and Fig. 5, where Fig. 4 shows the execution time of the test program run on each of the three processors (PERL, MIPS, MIMOPS), while Fig. 5 illustrates the improvement that is achieved by MIMOPS. Referring to these results, it can be noticed that PERL provides an improvement of 8.82% in comparison to MIPS, but on the other hand the MIMOPS processor outperforms both of them, achieving 1.33 times (25%) better results than MIPS and 1.21 times (17.7%) better results than PERL. This analysis is made just to show and emphasize the performance potential of the proposed MIMOPS processor.

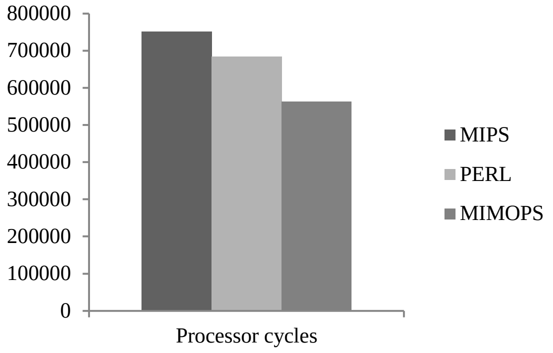


Fig. 4. Execution Time of 32x32 Matrices Multiplication on Three different Processors: MIPS, PERL and MIMOPS.

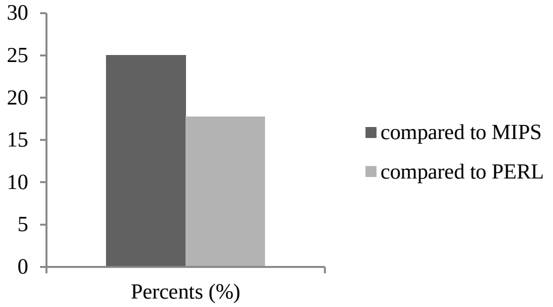


Fig. 5. Percentage Speedup of Execution Time of 32x32 Matrices Multiplication on MIMOPS Processor.

IV. FPGA IMPLEMENTATION OF THE PROPOSED RISC-BASED MEMORY-CENTRIC PROCESSOR ARCHITECTURE

The proposed MIMOPS processor is described in VHDL, by means of Xilinx VIVADO Design Suite. This software environment enables hardware designers to synthesize (compile) their HDL codes, perform timing analysis, examine RTL diagrams, simulate a design's reaction to different stimuli, and configure (program) a target FPGA device. In fact, all these functionalities are achieved by several different tools, including: Vivado regular synthesis and XST (High-Level Synthesis) compiler, Vivado implementation tool (translate, map, place, and route), Vivado Intellectual Property integrator, Vivado Simulator, Vivado serial I/O and logic analyzer for debugging, XDC (Xilinx Design Constraints) tool for timing constraints and entry, Vivado programming (Xilinx impact) tool etc. In general, Vivado is a design environment for FPGA products from Xilinx and is tightly-coupled to the architecture of such chips. Therefore, we use the Vivado tools suite in order to perform FPGA implementation of the proposed MIMOPS processor on Virtex7 VC709 Xilinx evaluation platform, [32].

The VHDL model of the proposed MIMOPS processor is organized in four modules (fetch, decode, execute, write-back) that form the processor's pipelined data-path and an additional module that provides communication with I/O devices. This is also presented in Fig. 6, where a block diagram (schematic) of the VHDL model of MIMOPS processor, generated in Vivado Design Suite, is given.

The fetch module is purposed to read an instruction from the on-chip instruction memory and to generate the next PC value that will be used for instruction fetch in the next tact cycle. This module includes three separate components: instruction memory, instruction page table and IF/ID pipeline register; that are accessed during the instruction fetching phase.

The decode module is purposed to decode the instruction that is sent from the fetch module and to read the instruction operands that are placed inside the on-chip data memory. Besides that, this module also executes shift and sign-extension operations for immediately-addressed operands, comparisons for conditional branching, data hazards detection, and produces control signals with the control unit. This module includes several separate components: data memory, data page table, ID/EX pipeline register, comparator, control unit, and a few multiplexers and extenders; that are accessed during the instruction decoding phase.

The execute module is purposed to execute shifting and arithmetical-logical operations and to select the result value (result from ALU, result from the shifter, etc.) that should be written back to the on-chip data memory. In addition to that, this module also performs forwarding of the result value and address to the decode module in order to prevent the occurrence of data hazards. This module includes several separate components: ALU for integer and real numbers, shifter, EX/WB pipeline register, result selector multiplexer, and several other multiplexers; that are accessed during the instruction executing phase.

The write-back module is purposed to write the result value to the on-chip data memory and to provide forwarding of the result to the decode module in order to prevent the occurrence of data hazards. This module acts as an interface to the decode module, which actually executes the operations of writing and resolving data conflicts.

The I/O communication module is purposed to transfer data between an I/O device and the MIMOPS processor (instruction or data on-chip memory) with IN or OUT instructions. Accordingly, this module uses an in/out data bus to receive data from an I/O device to its on-chip memory (when IN instruction is executed) or to send data to an I/O device from its on-chip memory (when OUT instruction is executed).

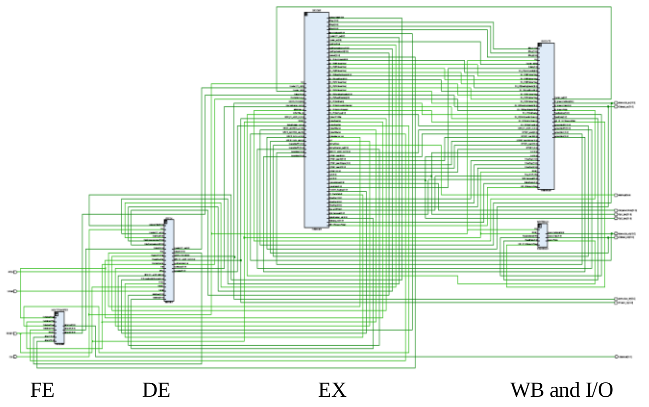
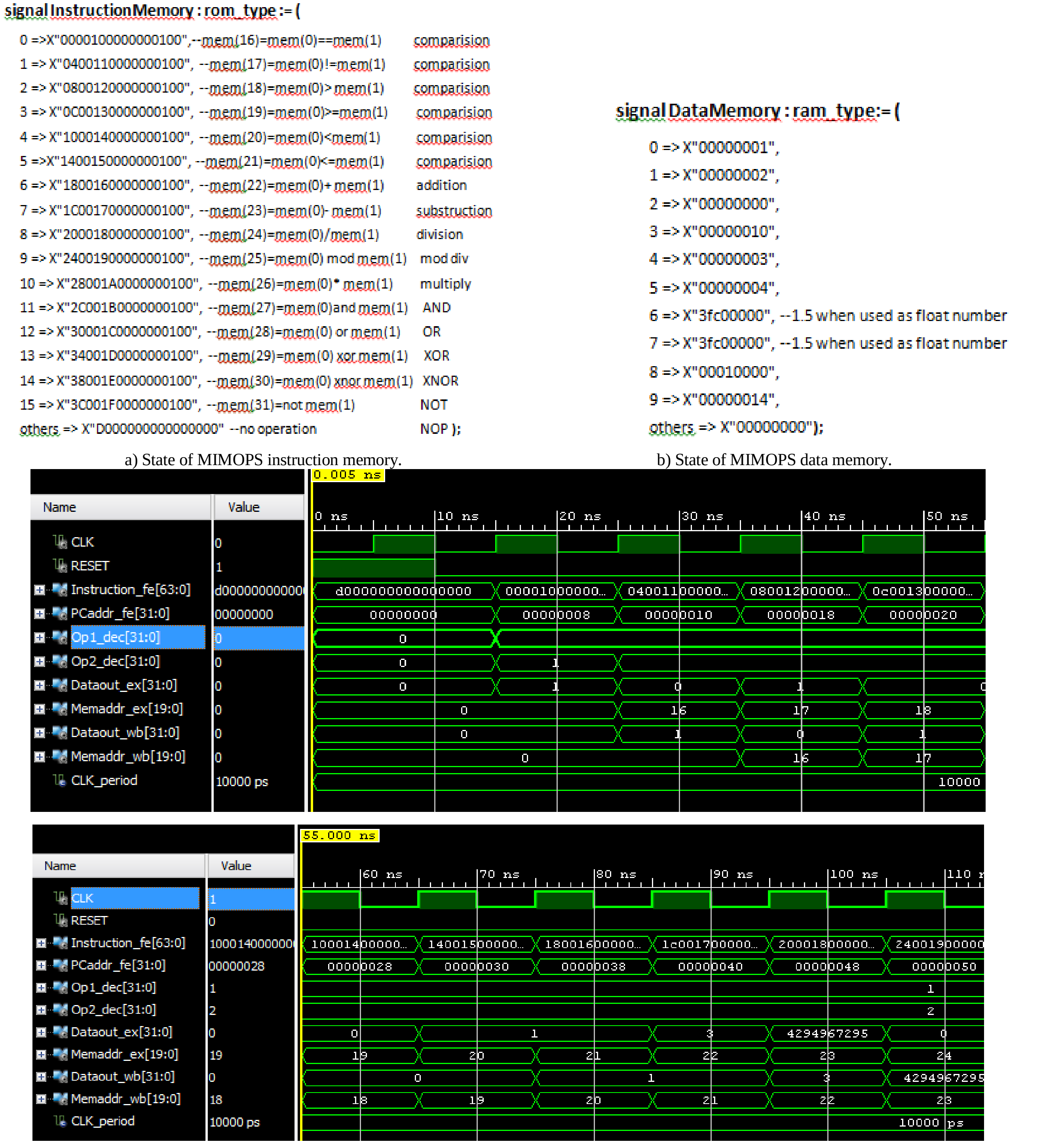


Fig. 6. Block diagram of the VHDL model of MIMOPS.

Each of the given VHDL modules is represented with a block diagram (schematic) that is generated by the Vivado Design Suite. In addition to that, the Vivado Simulator is used to verify the operation of these VHDL modules with separate test-bench programs, written for that purpose. Finally, the complete MIMOPS processor is simulated, and its overall functionality is verified. Therefore, a test-bench is written to analyze the processor's behavior during the execution of a test program that is placed in the processor's instruction memory, (given in Fig. 7(a)). Additionally, it is considered that the processor's data memory is already filled with data, as shown in Fig. 7(b). The results of the test-bench simulation are presented in Fig. 7(c).



```
signal InstructionMemory : rom_type := (

 0 =>X"0000100000000100",--mem(16)=mem(0)==mem(1)       comparision
 1 => X"0400110000000100", --mem(17)=mem(0)!=mem(1)     comparision
 2 => X"0800120000000100", --mem(18)=mem(0)> mem(1)     comparision
 3 => X"0C00130000000100", --mem(19)=mem(0)>=mem(1)     comparision
 4 => X"1000140000000100", --mem(20)=mem(0)<mem(1)      comparision
 5 =>X"1400150000000100", --mem(21)=mem(0)<=mem(1)      comparision
 6 => X"1800160000000100", --mem(22)=mem(0)+ mem(1)     addition
 7 => X"1C00170000000100", --mem(23)=mem(0)- mem(1)     substruction
 8 => X"2000180000000100", --mem(24)=mem(0)/mem(1)      division
 9 => X"2400190000000100", --mem(25)=mem(0) mod mem(1)  mod div
 10 => X"28001A0000000100", --mem(26)=mem(0)* mem(1)    multiply
 11 => X"2C001B0000000100", --mem(27)=mem(0)and mem(1)  AND
 12 => X"30001C0000000100", --mem(28)=mem(0) or mem(1)  OR
 13 => X"34001D0000000100", --mem(29)=mem(0) xor mem(1) XOR
 14 => X"38001E0000000100", --mem(30)=mem(0) xnor mem(1) XNOR
 15 => X"3C001F0000000100", --mem(31)=not mem(1)        NOT
 others => X"D000000000000000" --no operation           NOP );
```

a) State of MIMOPS instruction memory.

```
signal DataMemory : ram_type:= (

  0 => X"00000001",
  1 => X"00000002",
  2 => X"00000000",
  3 => X"00000010",
  4 => X"00000003",
  5 => X"00000004",
  6 => X"3fc00000", --1.5 when used as float number
  7 => X"3fc00000", --1.5 when used as float number
  8 => X"00010000",
  9 => X"00000014",
  others => X"00000000");
```

b) State of MIMOPS data memory.

Name	Value
CLK	0
RESET	1
Instruction_fe[63:0]	d00000000000
PCaddr_fe[31:0]	00000000
Op1_dec[31:0]	0
Op2_dec[31:0]	0
Dataout_ex[31:0]	0
Memaddr_ex[19:0]	0
Dataout_wb[31:0]	0
Memaddr_wb[19:0]	0
CLK_period	10000 ps

Name	Value
CLK	1
RESET	0
Instruction_fe[63:0]	100014000000
PCaddr_fe[31:0]	00000028
Op1_dec[31:0]	1
Op2_dec[31:0]	2
Dataout_ex[31:0]	0
Memaddr_ex[19:0]	19
Dataout_wb[31:0]	0
Memaddr_wb[19:0]	18
CLK_period	10000 ps

c) Results of MIMOPS VHDL Model Simulation.

Fig. 7. Simulation of VHDL model of MIMOPS Processor.

Once the VHDL model of the MIMOPS processor is simulated and verified, the next step is to perform synthesis and implementation of the particular processor in Vivado Design Suite. These activities are performed automatically with the synthesis and implementation tools, which are previously set to target the processor's FPGA realization on Virtex7 VC709 evaluation board, shown in Fig. 8. In general, the VC709 evaluation board provides a hardware environment for developing and evaluating designs targeting Virtex7 XC7VX690T-2FFG1761C FPGA, [32]. This board allows features common to many embedded processing systems, such as DDR3 memories, an 8-lane PCI Express interface, general-purpose I/O, and a UART interface. Other features can be added by using mezzanine cards attached to the VITA-57 FPGA mezzanine connector (FMC) provided on the board.

In the synthesis stage, the VHDL model of the MIMOPS processor is converted to a "netlist", which is composed of generic circuit components interconnected with connections. After the synthesis, the Vivado implementation tool is used to perform: translate, map, place, and route sub-steps. This way, the MIMOPS processor is translated and mapped to Xilinx Virtex7 XC7VX690T FPGA components and after that these components are physically placed and connected together (routed) on the appropriate FPGA board. Fig. 9 presents the state of the Virtex7 VC709 FPGA device after the synthesis and implementation of the MIMOPS processor.

Once the processor's implementation is finished, more detailed reports about the hardware characteristics of the designed MIMOPS processor are generated. According to the resource utilization report, shown in Fig. 10 it can be noticed that the proposed MIMOPS processor can be implemented in Virtex7 VC709 evaluation platform, by utilizing less than 1% of the slice registers and 36% of the slice LUT resources. This result is expected since the MIMOPS processor integrates the memory inside the chip and it implements complex mechanisms that provide hardware support for virtual memory (includes memory address generators with on-chip page tables and performs management of memory blocks etc). In addition to that, the MIMOPS processor includes a more complex control unit that provides support for direct access to memory operands. Besides the control unit, additional complexity is introduced with the implementation of data hazard detection unit, comparison logic purposed for conditional branching in the decode phase, ALU unit that is extended to operate with floating-point numbers and shifter unit that provides support for second source flexible operand. All these hardware units are implemented with the aim to improve the computing performances of the MIMOPS processor, which is actually achieved, but the chip complexity is increased.

In order to program the Virtex7 VC709 FPGA, a constraint file has to be prepared. This file is used to assign the VHDL code signals of the MIMOPS processor to the device pins found on the Virtex7 VC709 evaluation board. For example, the reset signal is assigned to the on-board CPU reset push button switch, which allows the user to manually reset the processor. Similarly, the CLK signal is assigned to the 200 MHz system clock of the FPGA board that is active on a positive edge. In addition to that, the last 8 bits of the ResultData signal that is forwarded from executing to the write-back stage are assigned to the 8 user LEDs of the FPGA board. More details about the Virtex7 VC709 board I/O pin assignments are given in Table 1.

After the FPGA programming, the user can analyze the execution of some program that is already loaded inside the processor's on-chip instruction memory, just by observing the changes of the LEDs state. It is considered that the given program operates with numbers that are in the range of [0-255]. Considering that a MIMOPS processor that works with 200 MHz system clock executes the program very fast, an additional component is defined in order to scale the input 200 MHz clock signal to 1 Hz clock signal (with a period of 1 s). This way, the state of the LEDs changes slowly, so the user can easily monitor the test program's execution.

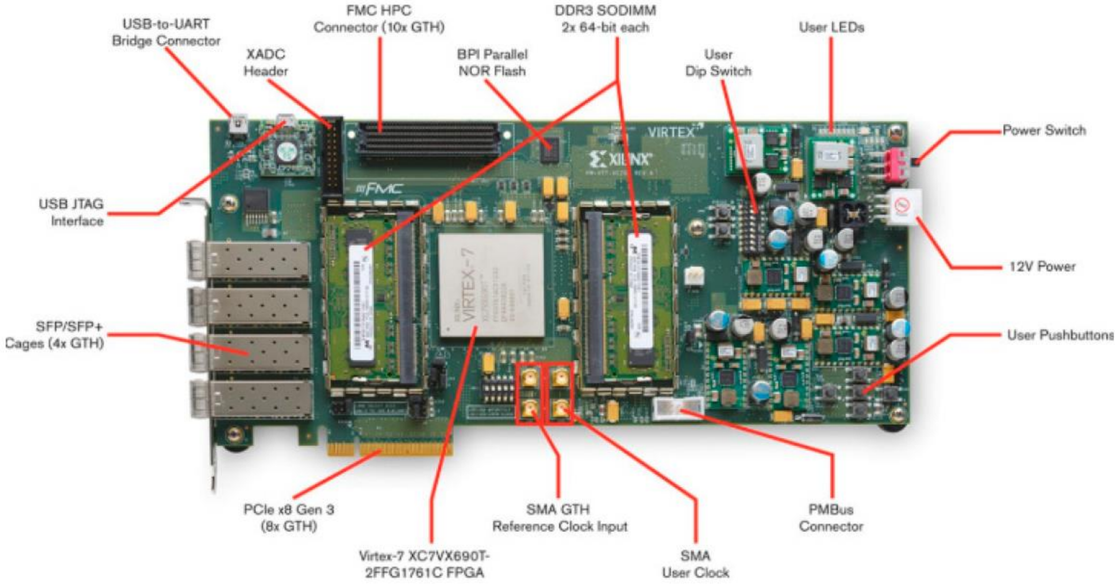
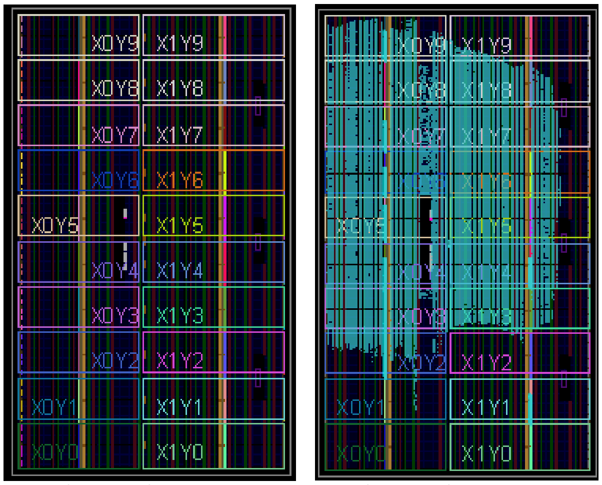
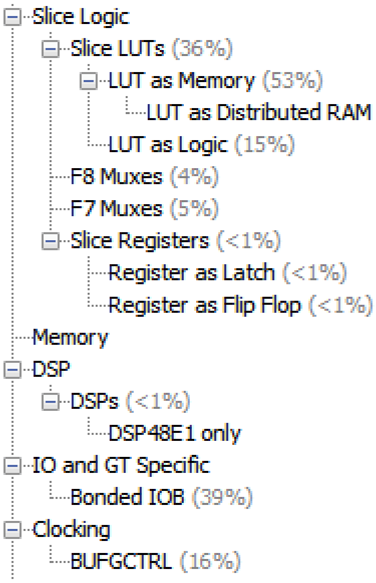


Fig. 8. Virtex7 VC709 evaluation board, [32].



a) State of FPGA after synthsis. b) State of FPGA after implementation.

Fig. 9. FPGA Implementation of MIMOPS Processor on Virtex7 VC709 Evaluation Board.

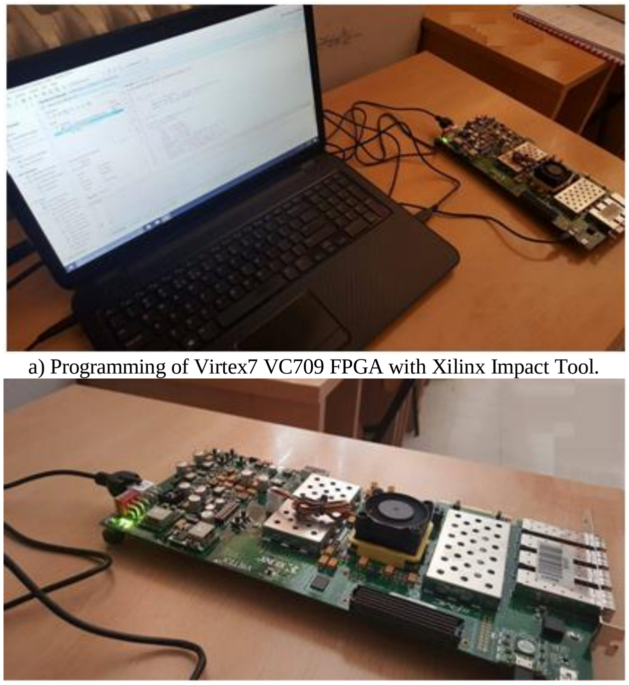


- Slice Logic
  - Slice LUTs (36%)
    - LUT as Memory (53%)
      - LUT as Distributed RAM
    - LUT as Logic (15%)
  - F8 Muxes (4%)
  - F7 Muxes (5%)
  - Slice Registers (<1%)
    - Register as Latch (<1%)
    - Register as Flip Flop (<1%)
- Memory
- DSP
  - DSPs (<1%)
    - DSP48E1 only
- IO and GT Specific
  - Bonded IOB (39%)
- Clocking
  - BUFGCTRL (16%)

Fig. 10. FPGA utilization(%) of MIMOPS Processor.

TABLE. I. VIRTEX7 VC709 BOARD I/O PINS ASSIGNMENT TO MIMOPS PROCESSOR'S SIGNALS

Signal	*Pin*	*Pin Type*	*Pin Function*
CLK	H19	IN	200 MHz system clock active on positive edge
Reset	AV40	IN	Reset push button (PB) switch
ResultData[7]	AU39	OUT	User LED 7
ResultData[6]	AP42	OUT	User LED 6
ResultData[5]	AP41	OUT	User LED 5
ResultData[4]	AR35	OUT	User LED 4
ResultData[3]	AT37	OUT	User LED 3
ResultData[2]	AR37	OUT	User LED 2
ResultData[1]	AN39	OUT	User LED 1
ResultData[0]	AM39	OUT	User LED 0



a) Programming of Virtex7 VC709 FPGA with Xilinx Impact Tool.

b) Simulation of MIMOPS Processor in Real Hardware.

Fig. 11. Hardware Prototype of MIMOPS Processor in Virtex7 VC709 FPGA Board.

Finally, the processor is completely ready to program onto the FPGA device, by means of the Xilinx impact tool. In that process, a bit stream file is generated and used to program the target FPGA device by JTAG cable. The created prototype of the proposed MIMOPS processor in real hardware i.e. Virtex7 VC709 XC7VX690T FPGA board is shown in Fig. 11. After that the test program that is shown in Fig. 7.a is simulated and executed in real FPGA, and the results are verified, according to the FPGA LEDs state and Fig. 7.c simulation diagrams.

V. CONCLUSION

This paper proposes a memory-centric processor core that is based on a standard MIPS implementation of RISC architecture, which is further improved to operate with separated on-chip data and program memory, by excluding the use of GPRs and cache (in and out of the processor chip). The memory-centric approach of processing provides 4-stage pipelining with direct access to the on-chip memory (without MEM phase), fast and simple access to the on-chip memory (without explicit load/store instructions), avoidance of copy operations and transfers of redundant data and blocks into GPRs and cache memory, decrease of capacity of redundant on-chip memory resources, high internal memory bandwidth, and removal of complex cache memory management mechanisms. Actually, it is shown that a MIMOPS processor achieves 25/17.7% better results than a MIPS/PERL processor when executing a 32x32 matrix-multiplication program.

The main focus of this paper is the FPGA implementation of the proposed MIMOPS processor. This includes designing of VHDL hardware model of the proposed processor and experimenting with Xilinx VIVADO Design Suite software environment, which provides support for Virtex7 VC709 FPGA evaluation board. In that process, the hardware model of the proposed RISC-based memory-centric processor is first simulated, by means of Xilinx VIVADO Design Suite Simulator tool. The simulation is performed with test bench programs that generate timing diagrams, which are further used for analyzing the behavior of the hardware model of the proposed processor and its components. The VIVADO synthesis and implementation tools are next employed in creating an RTL model of the proposed processor and implementing the synthesized processor in Virtex7 VC709 FPGA board. The reports that are generated from these tools present that the MIMOPS processor utilizes less than 1% of the slice registers and 36% of the slice LUT resources. The I/O mapping of the MIMOPS processor interfaces with the Virtex7 VC709 FPGA board pins and the programming of the given FPGA Virtex7 VC709 FPGA board are performed at the final stage. The created hardware prototype is used for simulating and analyzing of the proposed MIMOPS processor in real hardware, by means of Virtex7 VC709 FPGA component. This approach makes use of FPGA re-programmability, which has proven to be an ideal solution for achieving reasonable speed at a low price.

The proposed MIMOPS processor provides many advantages, especially in terms of processing speed, but on the other hand it imposes additional requirements to the system's hardware and software, which cause limitations in its application area. Accordingly, the proposed MIMOPS processor implements specific ISA and several special-purpose hardware components that provide direct operation with the on-chip memory. Therefore, it is obvious that a specific software support for the proposed MIMOPS processor should be developed in the future. The primer requirement would be designing of a dedicated compiler that would be able to translate high-level language programs to MIMOPS assembler, (that significantly differs from the assembler of other RISC-based processors) while keeping the standard programming model. Afterward, the next research activities would include developing of a dedicated operating system with process scheduler, which would be able to manage the MIMOPS on-chip memory and to coordinate the complete virtual address space, while multiple processes are being executed. Furthermore, assuming the recent innovation in processing in memory architecture and technology it may become desirable to build a scalable multi-processor MIMOPS-based system in very near future.

References

[1] D. A. Patterson, J. L. Hennessy, Computer Organization and Design: The hardware/software Interface, 5th ed., Elsevier, 2014.

[2] J. L. Hennessy, D. A. Patterson, Computer Architecture: A Quantitative Approach, 5th ed., Morgan Kaufmann Publishers, 2011.

[3] “Moore's law is dead - long live Moore's law,” in IEEE Spectrum Magazine, April 2015.

[4] J. Hruska, “Forget Moore’s law: hot and slow DRAM is a major roadblock to exascale and beyond,” in Extreme Tech Magazine, 2014.

[5] W. A. Wulf, S. A. McKee, “Hitting the memory wall: implications of the obvious,” in ACM SIGARCH Computer Architecture News, Vol. 23, Issue 1, March 1995.

[6] Y. Yan, R. Brightwell, X. Sun, “Principles of memory-centric programming for high performance computing,” in Proc. of Workshop on Memory Centric Programming for HPC, USA, 2017.

[7] D. Patterson, “Latency lags bandwidth,” in Communications of the ACM, Vol. 47, No. 10, 2004, pp 71-75.

[8] D. Jakimovska, A. Tentov, G. Jakimovski, S. Gjorgjievska, M.Malenko, “Modern processor architectures overview,” in Proc. of XVIII International Scientific Conference on Information, Communication and Energy Systems and Technologies, Bulgaria, 2012, pp. 239-242.

[9] R. Eigenmann, D. J. Lilja, “Von Neumann computers,” in Wiley Encyclopedia of Electrical and Electronics Engineering, Volume 23, 1998, pp. 387-400.

[10] A. Bakshi, J. Gaudiot, W. Lin, M. Makhija, V. K. Prasanna, W. Ro, C. Shin, “Memory latency: to tolerate or to reduce?,” in Proc. of 12th Symposium on Computer Architecture and High Performance Computing, 2000.

[11] S. Borkar, A. A. Chien, “The future of microprocessors,” in Communications of the ACM, Vol. 54 No. 5, May 2011, pp 67-77.

[12] Intel Corporation, “New microarchitecture for 4th gen. Intel core processor platforms,” Product Brief, 2013.

[13] W. Bao, S. Tavarageri, F. Ozguner, P. Sadayappan, “PWCET: power-aware worst case execution time analysis,” in Proc. of 43rd International Conference on Parallel Processing Workshops, 2014.

[14] P. Machanick, “Approaches to addressing the memory wall,” Technical Report, School of IT and Electrical Engineering, University of Queensland Brisbane, Australia, 2002.

[15] C. Kozyrakis, D. Patterson, “Vector vs. superscalar and VLIW architectures for embedded multimedia benchmarks,” in Proc. of the 35th International Symposium on Microarchitecture, Instabul, Turkey, November 2002.

[16] J. Silc, B. Robic, T. Ungerer, Processor architecture: From Dataflow to Superscalar and Beyond, Springer, 1999.

[17] N. FitzRoy-Dale, “The VLIW and EPIC processor architectures,” Master Thesis, New South Wales University, July 2005.

[18] M. Smotherman, “Understanding EPIC architectures and implementations,” in Proc. of ACM Southeast Conference, 2002.

[19] P. Suresh, “PERL - a register-less processor,” PhD Thesis, Department of Computer Science & Engineering, Indian Institute of Technology, Kanpur, 2004.

[20] P. R. Panda, N. D. Dutt, A. Nicolu, “On-chip vs. off-chip memory: the data partitioning problem in embedded processor-based systems,” ACM Transactions on Design Automation of Electronic Systems, 2000.

[21] V. Venkataramani, M. Choon Chan, T. Mitra, “Scratchpad-memory management for multi-threaded applications on many-core architectures,” ACM Transactions on Embedded Computing Systems, Vol. 18, Issue 1, 2019.

[22] C. Cojocaru, “Computational RAM: implementation and bit-parallel architecture,” Master Thesis, Carletorn University, Ottawa, 1995.

[23] H. Tsubota, T. Kobayashi, “The M32R/D, a 32b RISC microprocessor with 16Mb embedded DRAM,” Technical Report, 1996.

[24] J. Draper, J. T. Barrett, J. Sondeen, S. Mediratta, C. W. Kang, I. Kim, G. Daglikoca, “A prototype processing-in-memory (PIM) chip for the data-intensive architecture (DIVA) system,” Journal of VLSI Signal Processing Systems, Vol. 40, Issue 1, 2005, pp. 73-84.

[25] M. Gokhale, B. Holmes, K. Jobst, “Processing in memory: the Terasys massively parallel PIM array,” IEEE Computer Journal, 1995.

[26] K. Keeton, R. Arpaci-Dusseau, and D.A. Patterson, "IRAM and SmartSIMM: overcoming the I/O bus bottleneck", in Proc. of the 24th Annual International Symposium on Computer Architecture, June 1997.

[27] C. E. Kozyrakis, S. Perissakis, D. Patterson, T. Andreson, K. Asanovic, N. Cardwell, R. Fromm, J. Golbus, B. Gribstad, K. Keeton, R. Thomas, N. Treuhaft, K. Yelick, “Scalable processors in the billion-transistor era: IRAM,” IEEE Computer Journal, Vol. 30, Issue 9, pp 75-78, 1997.

[28] J. Gebis, S. Williams, D. Patterson, C. Kozyrakis, “VIRAM1: a media-oriented vector processor with embedded DRAM,” 41st Design Automation Student Design Contest, San Diego, CA, 2004.

[29] K. Murakami, S. Shirakawa, H. Miyajima, “Parallel processing RAM chip with 256 Mb DRAM and quad processors,” in Proc. of Solid-State Circuits Conference, 1997.

[30] S. Kaxiras, D. Burger, J. R. Goodman, “DataScalar: a memory-centric approach to computing,” Journal of Systems Architecture, 1999.

[31] M. Oskin, F. T Chong, T. Sherwood, "Active pages a computation model for intelligent memory," in Proc. of the 25th Annual International Symposium on Computer architecture, 1998, pp. 192-203.

[32] Xilinx, "VC709 evaluation board for the Virtex-7 FPGA," User Guide, 2019.

[33] J. M. P. Cardoso, M. Hubner, Reconfigurable Computing: From FPGAs to Hardware/Software Codesign, Springer-Verlag New York, 2011.

[34] S. Li, K. Chen, J. B. Brockman, N. P. Joupp, "Performance impacts of non-blocking caches in out-of-order processors," Technical Paper, 2011.

[35] S. Ghose, K. Hsieh, A. Boroumand, R. Ausavarungnirun, O. Mutlu, "The processing-in-memory paradigm: mechanisms to enable adoption," in book: Beyond-CMOS Technologies for Next Generation Computer Design, 2019.

[36] G. Singh, L. Chelini, S. Corda, A. Javed Awan, S. Stuijk, R. Jordans, H. Corporaal, A. Boonstra, "A review of near-memory computing architectures," in Proc. of the 21st Euromicro Conference on Digital System Design, 2018.

[37] E. Azarkhish, D. Rossi, I. Loi, L. Benini, "Design and evaluation of a processing-in-memory architecture for the smart memory cube," in Proc. of the 29th International Conference Architecture of Computing Systems, Germany, 2016.

[38] E. Vermij, L. Fiorin, R. Jongerius, C. Hagleitner, J. Van Lunteren, K. Bertels, "An architecture for integrated near-data processors," ACM Transactions on Architecture and Code Optimization, Vol. 14, Issue 3, 2017.

[39] Hewlett Packard Labs, "The machine: the future of technology," Technical Paper, 2016.

[40] D. Efnusheva, A. Cholakoska, A. Tentov, "A survey of different approaches for overcoming the processor-memory bottleneck," International Journal of Computer Science & Information Technology, Vol. 9, No. 2, April 2017.

[41] G. Dokoski, D. Efnusheva, A. Tentov, M. Kalendar, "Software for explicitly parallel memory-centric processor architecture," in Proc. of Third International Conference on Applied Innovations in IT, 2015.

[42] K. Vollmar, P. Sanderson, "MARS: an education-oriented MIPS assembly language simulator," in Proc. of the 37th SIGCSE Tech. Symposium on Computer Science Education, 2007.

Authentication and Authorization Design in Honeybee Computing

Nur Husna Azizul[1], Abdullah Mohd Zin[2], Ravie Chandren Muniyandi[3], Zarina Shukur[4]
Center for Software Technology and Management (Softam)
Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

***Abstract*—Honeybee computing is a concept based on advanced ubiquitous computing technology to support Smart City Smart Village (SCSV) initiatives. Advanced ubiquitous computing is a computing environment that contains many devices. There are two types of communication within Honeybee computing: client server and peer-to-peer. One of the authorization techniques is the OAuth technique, where a user can access an application without creating an account and can be accessed from multiple devices. OAuth is suitable to control the limited access of resources to the server. The server use REST API as web service to publish data from resources. However since Honeybee computing also supports peer-to-peer communication, security problem can still be an issue. In this paper, we want to propose the design of a secure data transmission for Honeybee computing by adopting the authorization process of OAuth 2.0 and Elliptic Curve Diffie-Hellman (ECDH) with HMAC-Sha. This article will also discuss the communication flow after adopting OAuth 2.0 and ECDH to the computing environment.**

Keywords—HMAC-Sha; REST API; peer-to-peer; web service; honeybee computing

I. INTRODUCTION

Honeybee computing is a concept based on advanced ubiquitous computing technology to support Smart City Smart Village (SCSV) initiatives1. It is supported by a middleware together with a number of tools such as semantic knowledge tool and predictive analytics for information management. The sources of information in Honeybee Computing are from the web, public and private cloud, and user devices. Since there is a multiple sources of data, it is important that all transactions are secured.

In the development of a software, the effort to secure the software is important, for example a framework cyber security [1] strategy framework is to protect government data, foreign investment and citizens. With many types of attacks, the importance of security is not only to look at securing the data but also to ensure users authenticity [2], especially if the interaction involves third party users. For example, a design for a virtual private network [3] for collaboration specialist users where the authentication becomes the main part of the design and authentication mechanism for an ad-hoc network [4]. One of the popular security problems within a network is the man-in-the-middle (MITM) attack [5][6]. The problem of a MITM attack is more critical in applications that use the single sign on (SSO) method. The Facebook platform that is based on cloud computing is open to multiple types of MITM attacks [5][7].

Authorization and authentication [8][9] are security issues that must be considered during the development of an application. There are multiple cloud service providers with client authentication method, for example Amazon Web services that use HMAC-Sha1, HMAC-Sha256, or X.509 certificate, Azure uses SAML 2.0 or Auth 2.0, Azure Storage uses HMAC-Sha256, and Google App Engine uses OAuth 2.0, shared secret or certificate. HTTP authentication [6][10][11] provides basic and digest access authentication.

Honeybee computing needed authorization authentication that support the architecture, since the Honeybee computing support peer-to-peer and client server, secure communication during data transfer is important to protect the resource. Client server use authorization and authentication that involved storing of key in server side, while peer to peer security mechanism usually involved with encryption and decryption. There is no security method for secure communication with both client server and peer to peer communication. This paper discusses the authorization and authentication process in Honeybee computing.

The rest of this paper is organized as follows: The existing work of the attack, client server and peer to peer method to secure the communication is presented in Section 2. In Section 3, we present the overview of Honeybee Computing, before discussing the findings in Section 4. Section 5 presents the communication flow. Finally, Section 6 concludes the paper and presents the future work.

II. RELATED WORK

A. MITM Attack

Generally, there are three types of MITM attack, namely Address Resolution Protocol (ARP) Cache Poisoning, Domain Name System (DNS) Spoofing and Session Hijacking [12]. The attacks that use ARP spoofing [13] refers to a technique that enables an attacker to pretend to be one of the users in a communication between two users. The DNS spoofing principle [13] is where the victim’s HTTP traffic is intercepted. The program analyzes incoming HTTPS links and replaces them with unprotected HTTP links or homographic ally related secure links. Session hijacking is the hijacking of a valid computer session to the browser. The aim of MITM [12] is to compromise the confidentiality, integrity and availability of messages. Based on the three effects, the scenario that would be caused by MITM would be as follows:

[1] GSIAC Smart City. “ Smart City-Smart Village”. http://gsiac.org/index.cfm?&menuid=36#sthash.74DYwYR0.dpuf [28 January 2015].

Confidentiality: The confidentiality of a message can be compromised where the message can be seen by a man-in-the-middle by interrupting the communication in the middle without both victims realizing. Conti et al. presents the ARP spoofing for the man-in-the-middle to stay in the communication in silence as shown in Fig. 1. The victim, Bob would send a message to Alice without realizing that Eve is actually the one that sent a message to both victims. This would cause a confidentiality risk to the conversations between both victims even though they might believe that their conversation is safe.

Integrity: Message integrity can be compromised by the man-in-the-middle attacking the communication and modifying the message. This is one of the possible effects caused by session hijacking; for example, the attackers can interrupt the information and then modify the message to get session access so that the attacker can access the resources. These types of attacks usually occur during the authentication process and cause a threat to users since the layer of security is not secure. Session hijacking [12] is very dangerous for Internet banking especially since it contains sensitive data. During MITM attacks, traffic is usually interrupted, and a spoofed certification is given to the client to mimic the server.

Availability: There are a lot of prevention methods to ensure the security of the communication during a man-in-the-middle attack. Since the attack starts from the early stage, the victims might expose the authentication procedure to the attackers. One of the methods is to identify if the message is compromised by the attacker during authentication. The issue with a man-in-the-middle attack is that message modification shows the interruption. That is why there is a lot of cryptography methods that involve key sharing to ensure the message is not compromised.

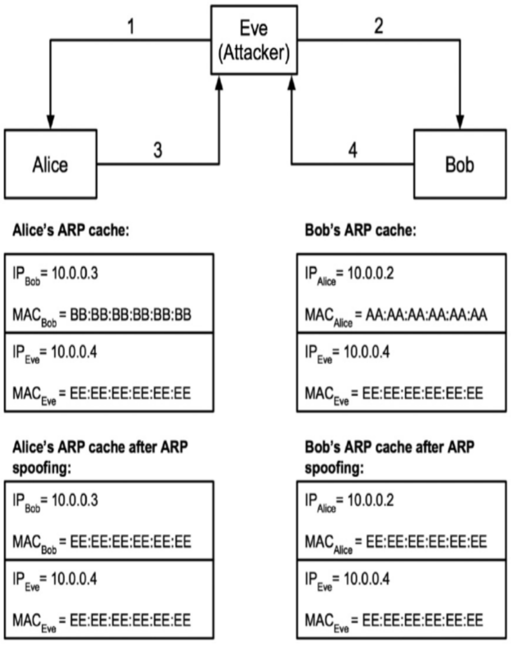

1
Eve
(Attacker)
2
Alice
3
4
Bob
Alice's ARP cache:
IP_{Bob}= 10.0.0.3
MAC_{Bob} = BB:BB:BB:BB:BB:BB
IP_{Eve}= 10.0.0.4
MAC_{Eve} = EE:EE:EE:EE:EE:EE
Bob's ARP cache:
IP_{Alice}= 10.0.0.2
MAC_{Alice} = AA:AA:AA:AA:AA:AA
IP_{Eve}= 10.0.0.4
MAC_{Eve} = EE:EE:EE:EE:EE:EE
Alice's ARP cache after ARP spoofing:
IP_{Bob}= 10.0.0.3
MAC_{Bob} = EE:EE:EE:EE:EE:EE
IP_{Eve}= 10.0.0.4
MAC_{Eve} = EE:EE:EE:EE:EE:EE
Bob's ARP cache after ARP spoofing:
IP_{Alice}= 10.0.0.2
MAC_{Alice} = EE:EE:EE:EE:EE:EE
IP_{Eve}= 10.0.0.4
MAC_{Eve} = EE:EE:EE:EE:EE:EE

Fig. 1. Authorization Process in OAuth 2.0 [6].

B. OAuth

OAuth is suitable to control the limited access of resources of the server.

For the server side, OAuth 2.0 is an authorization framework that is suitable for an environment that involves the use of multiple devices. OAuth 2.0 is an evolution from OAuth where it uses REST API as the development language. This specification is being developed within the Internet Engineering Task Force (IETF). The recent technology of OAuth 2.0 provides new security for users to enable third party applications to access resources from third party providers. These resources can be obtained using REST API[2]. For example, a person who already has a Facebook account can log in to the Spotify application through his Facebook account. Spotify does not have to know the username and password of the Facebook account.

The authorization process can be done using token authentication. Richardson and Ruby [6] present two types of authorization: using web interface or without web interfaces. Fig. 2 shows authorization with a web interface. In this example, the application needs permission from the user before Google provides a token to enable the application to use the Google calendar data.

For the client side, two types of platforms are considered, namely the Android apps and web application. For the Android apps to use third party libraries, an SDK [14] is normally provided. For example, an Android SDK is provided by Google for the programmer to develop apps without the need to register. This is different from using third party libraries online. For example, in a social network such as Facebook, Facebook SDK provides libraries for the Android platform. To access this SDK, a programmer has to create an account in Facebook and register as a programmer. The apps must be registered to Facebook and a secret key is provided for the apps to run online. The interaction is almost the same with web application. Facebook provides a system to enable programmers to register the application profile so that Facebook would recognize the list of applications that have access to the Facebook server.

Fig. 3 shows the sequence diagram of the authorization process. In this figure, a user refers to a person who has an account in the resource server, a user agent is the web application or Android app developed by the programmer. The client is provided by the middleware developer. Client act as interface. The user agent sends a request to the client, and the client will redirect the request to the URL of the server.

C. Infrastructure-Less Communication

In an infrastructure-less communication, such as in a peer-to-peer or ad hoc network, there is no server to manage registered devices. The authorization process is done using a hello protocol. HELLO beacon messages per interval time in order to periodically update link information [15]. To follow hello beacon, the protocol would affect the network performances. Thus, to authorize users in infrastructure-less devices, a key provided by the programmer is sent by device A; if device B as the receiver returns the right string then device B is authorized.

[2] U. Friedrichsen ," OAuth 2.0: A standard is coming of age. codecentric AG". http://www.slideshare.net/ufried/oauth-20-18356495.

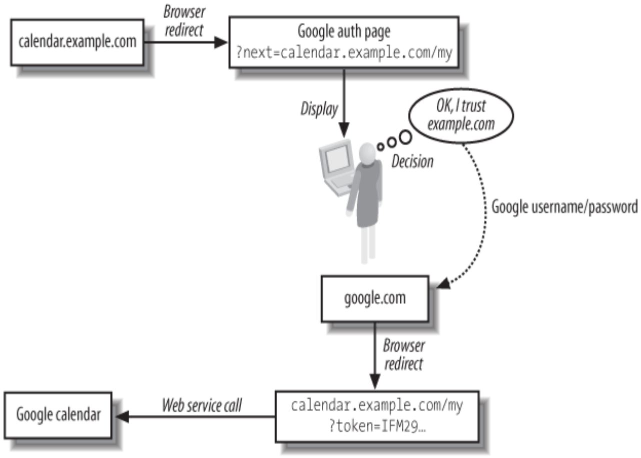


Fig. 2. Token to Access Functionality in Google Calendar [6].

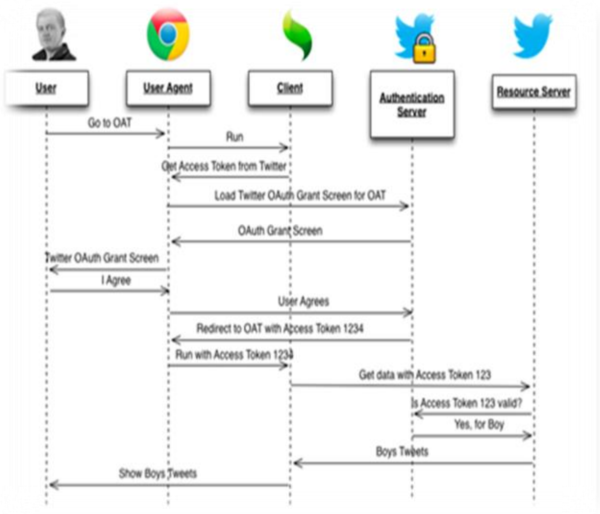

User
User Agent
Client
Authentication Server
Resource Server
Go to OAT
Run
Get Access Token from Twitter
Load Twitter OAuth Grant Screen for OAT
OAuth Grant Screen
Twitter OAuth Grant Screen
I Agree
User Agrees
Redirect to OAT with Access Token 1234
Run with Access Token 1234
Get data with Access Token 123
Is Access Token 123 valid?
Yes, for Boy
Boys Tweets
Show Boys Tweets

Fig. 3. Shows Authorization that Involves Third Party and user [14].

One of the solutions for security in peer-to-peer is provided by the Android API[3] is known as Elliptic Curve Diffie-Hellman [16] that sends a key without sharing the actual key. In this approach, the message is encrypted and decrypted. Cryptographic[3] hash functions are important security primitives especially for data integrity and authentication. HMAC-SHA1 is one of the cryptographic hash functions used to ensure data integrity in computing communication.

III. Honeybee Computing Overview

Honeybee computing is a concept based on advanced ubiquitous computing to support Smart City Smart Village[1]. Honeybee computing contains several components such as semantic knowledge tool and predictive analytics for information management [17]. The sources of information in Honeybee computing are the web, public and private cloud, and devices.

In order to enable applications to be developed in a Honeybee computing environment, a middleware is needed. The architecture of the middleware to support the applications development in a Honeybee computing environment is described in [17]. The middleware architecture is shown in Fig. 4. There are five main components in the middleware, namely, Service Manager, Communication Manager, Security Manager, Semantic Manager, and News Manager. Network Management provides the connection with the user devices, while Resource Manager supports the management of the available resources. Honeybee computing supports two types of network: infrastructure or client-server based network and infrastructure-less or peer-to-peer network. In an infrastructure based network, users can communicate through a server. Communication in peer-to-peer network does not involve an intermediate server, but rather users communicate with each other directly. Most of the users access computing network through wireless communication either through Wi-Fi, ZigBee or 3G/4G technology. In a client-server based network, communication between the server and client is done through web services. There are two methods in providing web services: Representational State Transfer (REST) and Simple Object Access Protocol (SOAP). REST is a more popular choice for middleware as it is simpler to use.

Currently there are two types of applications: normal applications operating on a PC and apps operating on a mobile device. Development of apps within the Honeybee middleware is supported by a software development kit (SDK). Since a user may have a number of devices, the same application or app may be installed in some or all of these devices. Every application or app uses many different types of services. In order to control the type of services that can be accessed by an app, there is a need for permission control. There are two common types of permission provided by an application or app:

1) *Reading:* The application or app can obtain user's data stored in the honeybee server.

2) *Delete:* The application or app can delete user's data stored in the server.

Interaction between users, devices, applications and apps and permission is shown in Fig. 4.

The Honeybee Security manager is responsible in preventing malicious attacks to the server. The protection is done through an authorization key. Other components of the middleware are only accessible if the request contains the approved authorization key.

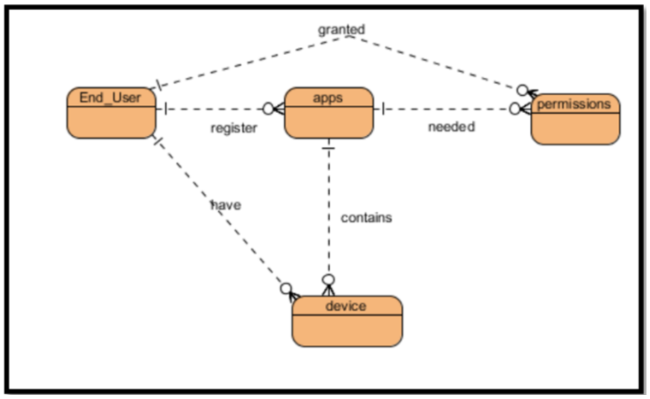


Fig. 4. Shows Interaction of the Security Package.

[3] N. Elenkov,"ECDH on Android sample app". Github. https://github.com/nelenkov/ecdh-kx.

Fig. 5 shows that there are two ways to access the services provided by the Honeybee Security manager. A Honeybee application on a PC can access the Security Web API directly. A Honeybee app on a mobile device will access the Security Web API through an SDK provided by org.honeybee.security.

There are two types of services provided by the Honeybee manager:

1) Authentication and authorization services to access services provided by other parts of the middleware; and

2) Encryption and decryption services for communication between apps on different devices.

Authentication is needed to access the Honeybee middleware. A server as a security manager provides service to ensure whether the app is authorized to access sources. For user sign up, login, and logout, a link to the Honeybee main web page is provided. After an authentication, a user has a session ID that is stored in cookies. The rules are as follows:

1) Programmer needs to register the apps to obtain information such as apps id and apps secret.

2) End_user is granted the type of permissions for the apps to access the source.

3) Each device provides a MAC address to access the server, and every token is provided for each device at each request to the server.

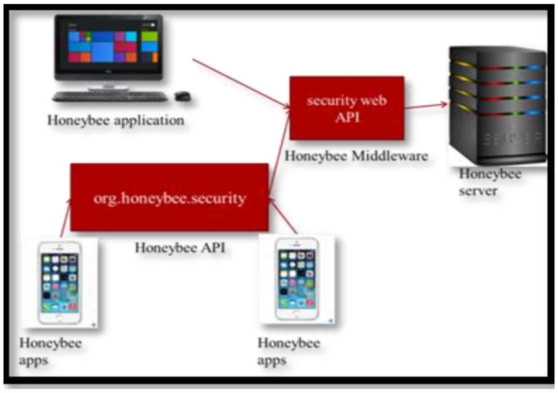


Fig. 5. Secret Generated by HMAC-SHA1 Algorithm for Programmer to Develop Application.

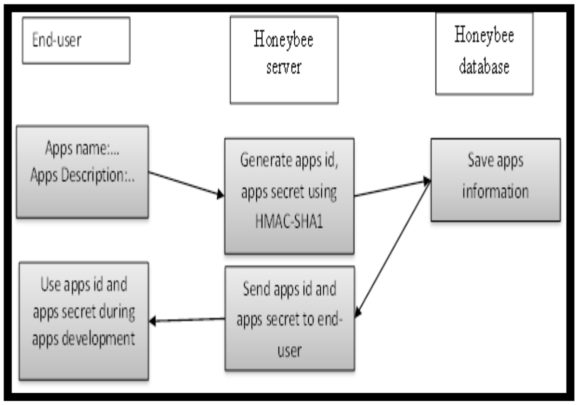

End-user
Honeybee server
Honeybee database
Apps name:...
Apps Description:...
Generate apps id, apps secret using HMAC-SHA1
Save apps information
Use apps id and apps secret during apps development
Send apps id and apps secret to end-user

Fig. 6. Sequence Diagram in Security Manager.

In this system, a programmer is a user, thus user sign up is needed and then the user must register as a programmer to access the developer dashboard.

For an end-user to develop an app, registration is needed where the end-user needs to agree to the programmer agreement. Apps registration is needed for the Honeybee apps to access the Honeybee API. An app ID and app secret is provided to the end-user for development purposes; the sequence is shown in Fig. 6.

IV. Authorization in Honeybee Computing

Honeybee computing authorization follows the OAuth workflow. The authorization is processed using multiple predefined URLs, called endpoints. There are 4 endpoints:

1) Request URI (this endpoint passes the request token).

2) Access URI (exchanges request token for an access token).

3) Authorize URI (confirms that the access token is valid).

4) Refresh token (refresh access token if previous is invalid).

Fig. 7 shows the sequence diagram between user, programmer, client, and server. The programmer develops the Honeybee app or web application and interacts with the Honeybee client before redirecting to the endpoints in the server. This process is adapted from the OAuth2.0 security.

Each number in the diagram is explained as follows:

- Request URI: The first endpoint is the request URI. This request URI is provided by the packages in Honeybee computing. The request URI passes the app secret to check if the user has granted permission to the Honeybee app to access the user account.
- Access URI: In the second endpoint, after the user is granted permission, an access token is provided. This access token can be used for requests to the resource server.
- Authorize URI: In this third endpoint, the access token is checked whether it is valid or not; if the token is valid then data is returned from the resource server.
- Refresh token URI: The fourth endpoint is used to refresh expired tokens; this is used when the user opens the app and the token saved is expired. This endpoint is not shown in the diagram. From Fig. 7, users must grant permission to enable third party applications to the access server.

The Honeybee computing Security manager is responsible for request management. This is to ensure that all requests are authorized. This is because every resource in Honeybee computing needs permission from the user and tokens to get authorized. The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

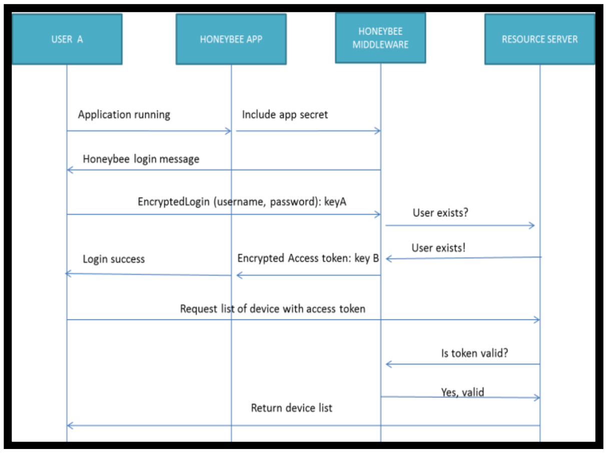


Fig. 7. Communication Flow when Device uses WAP Connection.

V. Communication in Honeybee Computing

There are a number of communication processes involved in Honeybee computing: user registration, application registration, accessing server for resources, finding nearby devices, communicate with other devices, peer-to-peer communication and building an ad hoc network. The descriptions of these processes are as follows:

User registration: Before a user can use the application, the user must have an account with the Honeybee system. This registration is different from the application registration; user must grant different types of permissions. This procedure is to protect multiple access to user data. In every user device there are a lot of applications, however only applications that are granted by the user can access the Honeybee server.

Application registration: Since the Honeybee system is a community system, the third party application would need to be authenticated by the Honeybee server. Thus, the app developer or programmer must register the application to the system. Programmers have to input the application package and then the app ID and app secret are provided for that particular application as a signature to access the server.

Accessing server for resources: The sequence of activities to assess a Honeybee resource is shown in Fig. 8. When a user is successfully logged in to the system, the middleware will provide a token and the application can request the resource by using the provided token. Fig. 8 shows a device communication when connected to WAP.

Finding Nearby Devices: To find a nearby device, the user must use Wi-Fi to connect to the Honeybee server. User must log in to the Honeybee system to get information needed from the server. Honeybee SDK then lists all nearby users who are online using WAP.

Communicate with other devices: The steps needed for communicating with other mobile devices using Wi-Fi are as follows:

1) Device A sends an encrypted message and key to device B. Device B then replies with an encrypted message and key.

2) Device A then decrypts the message and checks the following:

a) Decrypted message follows the format of the message.

b) MAC address inside the message exists in the file downloaded from the Honeybee server.

Device B would follow the same procedure to identify if the key is modified. During this step, both devices can proceed using the key or stop the session and start with a new session to get a new key.

3) If both devices agree to proceed, device A sends an encrypted message to device B without the key sending new keys.

Peer-to-peer communication: Peer-to-peer communication uses a concept similar as WAP since both use Wi-Fi to communicate; at the same time peer-to-peer is also an ad hoc network. To explain the communication, the steps of the procedure are as follows:

1) Device A scans for nearby devices using the Android library and identifies which device to send a message. Device A then encrypts the message and sends with a key to the selected device, which is device B. Device B then replies with the encrypted message and key by following the format needed.

2) Device A and device B then decrypt the message given and identify the following conditions:

a) Decrypted message follows the format to send message; and

b) One of the attributes is following the file downloaded from the Honeybee server earlier.

If any of the conditions are not followed, the communication is suspected as not secure. Both devices could proceed or stop the current session. If both devices agree to continue, the message can be sent by sending an encrypted message without the need to send the key.

Building an ad hoc network: An ad hoc network involves a number of devices connected together to form a network. It does not involve an intermediate server to send messages from one device to another, thus the communication is real time. By using the same authentication method as WAP, the app then can communicate. After logging in to the Honeybee system, the system would provide the information of registered devices to the Honeybee system.

VI. Conclusion

In man-in-the-middle attacks, there are multiple methods to prevent these attacks. There are several methods to prevent the effect of the attacks such as focus on the authentication method and using a different technique of cryptography for messages. There are also methods that focus on the digital signature to prevent any attacks. Honeybee computing security on the other hand, adapts the cryptography method and also validates the application at the server. To get the list of registered devices, the device communicates with the server before connecting to the device. These two communication need to identify if a man-in-the-middle attack is possible during data transfer. To prevent the man-in-the-middle attack, three methods of prevention are covered which are identity identification, resource protection and communication secure.

To ensure the authorization, the OAuth 2.0 method is used. This type of authorization using key sharing involves user agreement before proceeding. With this method, not all data is accessible to the user. This method is also located at the server the location of the resources. Other than that, every authorization and authentication process involves more than one secret key, thus the process to ensure the security of the communication is complex to get easily attacked by intruders. Since the man-in-the-middle is capable of impersonating the victim, one of the approaches in Honeybee computing is identity identification. To ensure if the sender is not an attacker, the server provides a list of devices that helps the user to identify if the sender is actually a registered user or an attacker. Since the list of registered devices is downloaded from the server, the attacker would need to attack the communication with the server to modify the list. By using the OAuth 2.0 method that is secure for the environment, the attacker would have difficulties to interrupt the communication.

One of the causes of man-in-the-middle is message availability where the attacker changes the message without victims realizing. For this case, we use cryptographic and also adopt the key sharing method. To prevent the key from being attacked by intruders, we use the ECDH key for key sharing. This method is one of the methods that are used for man-in-the-middle attacks to prevent any modification to the message. If the attacker is capable of accessing the message, the communication is still secure since the attacker would face other difficulties in decrypting the message. The communication design in Honeybee computing is mainly based on the entities inside the system. Two types of users which are application user and programmer show that Honeybee computing is a community system. Since there is involvement of third party applications in accessing the resources, the security involvement from different aspects is very important. With multiple devices that use multiple applications to access the resources, the OAuth 2.0 technique is used. Since Honeybee computing architecture involves the infrastructure-less communication, the ECDH method is adopted in the Honeybee computing security. These two methods prevent man-in-the-middle attacks to the system. The effects of man-in-the-middle attacks would cause a lot of problems.

As conclusion, it is important for any software that involved with communication transmission to have a secure data transmission. The after effect of MITM would cause many troubles in the resources To ensure the proposed authorization and authentication help to secure communication in honeybee computing, the future research is needed, which is to test whether the communication is secure from MITM attack, a simulation is needed where MITM attack to the application developed in the honeybee computing. With the functionality provided in honeybee computing such as send message, request data from middleware, and register user info, all the functionality must be tested to ensure MITM attack would not happen.

The next task is to validate the proposed security mechanism. The validation process will be done using security testing based on the test cases [18].

ACKNOWLEDGMENT

The authors would like to thank Mr. Muhammad Zharif for his valuable suggestions on the improvement of the paper. This research is one of the Malaysian Government funded projects under the Ministry of Higher Education (MOHE) Fundamental Research Grant (FRGS) with reference number FRGS/1/2015/ICT04/UKM/02/3.

REFERENCES

[1] K. Salamzada,Z. Shukur and M. Abu Bakar , “A Framework for Cybersecurity Strategy for Developing Countries: Case Study of Afghanistan”. Asia-Pacific Journal of Information Technology and Multimedia,4(1): 1 – 10.

[2] Sidra Ijaz, Munam Ali Shah, Abid Khan and Mansoor Ahmed, “Smart Cities: A Survey on Security Concerns” International Journal of Advanced Computer Science and Applications(IJACSA), 7(2), 2016.

[3] A. Kargar Raeespour A and AM Patel, “Design and Evaluation of a Virtual Private Network Architecture for Collaborating Specialist Users”. Asia-Pacific Journal of Information Technology, 5 (1):15 – 30.

[4] MA Catur Bhakti and A. Abdullah, ”EAP Authentication Mechanism for Ad Hoc Wireless LAN”, Journal of Information Technology and Multimedia, 5(2008): 13-40.

[5] V. Rastogi and A. Agrawal, “All your Google and Facebook logins are belong to us: A case for single sign-off”. 2015 Eighth International Conference on Contemporary Computing (IC3) (2015), Noida, 20th–22nd August, India.

[6] L. Richardson and S Ruby, “RESTful web service”. 1st ed. O’Reilly Media.

[7] Mohammed Nasser Al-Mhiqani, Rabiah Ahmad, Warusia Yassin, Aslinda Hassan, Zaheera Zainal Abidin, Nabeel Salih Ali and Karrar Hameed Abdulkareem, “Cyber-Security Incidents: A Review Cases in Cyber-Physical Systems” International Journal of Advanced Computer Science and Applications(IJACSA), 9(1), 2018.

[8] T. Ziebermayr and S. Probst, “Web Service Authorization Framework”. Proc. ICWS, 614-621.

[9] H. Lu , “Keeping Your API Keys in a Safe”. Proc. CLOUD, 962-965.

[10] J. Franks, P. Hallam-Baker, J. Hostetler,S. Lawrence,P. Leach,A. Luotonen and L. Stewart, “RFC 2617: HTTP Authentication: Basic and Digest Access Authentication”. IETF. https://tools.ietf.org/html/rfc2617#section-2

[11] Muhammad Kazim and Shao Ying Zhu, “A survey on top security threats in cloud computing” International Journal of Advanced Computer Science and Applications(IJACSA), 6(3), 2015.

[12] M. Conti, N.Dragoni, and V. Lesyk, “A Survey on Man in the Middle Attack”. IEEE Communications Surveys and Tutorials, 18(3):2027-2051.

[13] AA. Maksutov,IA. Cherepanov and MS. Alekseev, “Detection and prevention of DNS spoofing attacks”. 2017 Siberian Symposium on Data Science and Engineering (SSDSE).

[14] T. Ziebermayr and S. Probst, “Web Service Authorization Framework”. Proc. ICWS, 614-621.

[15] E. Khan, M. El-Kharashi, F. Gebali and M. Abd-El-Barr,”Design space exploration of a reconfigurable HMAC-hash unit”. Journal of Research and Practice in Information Technology. 40(2):109.

[16] E. Khan,M. El-Kharashi,F. Gebali and M. Abd-El-Barr ,” Design space exploration of a reconfigurable HMAC-hash unit”. Journal of Research and Practice in Information Technology. 40(2):109.

[17] NH. Azizul, A. Mohd Zin and E. Sundararajan,”The Design and Implementation of Middleware for Application Development within Honeybee Computing Environment”. IJASEIT, (6)6:937-943.

[18] A. Lunkeit and I. Schieferdecker “Model-Based Security Testing - Deriving Test Models from Artefacts of Security Engineering”. 2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 244-251.

Smartphone Image based Agricultural Product Quality and Harvest Amount Prediction Method

Fertilizer Control through Quality Prediction by using Smartphone Images of Agricultural Products

Kohei Arai[1]
Faculty of Science and Engineering
Saga University, Saga City, Japan

Osamu Shigetomi[2], Yuko Miura[3], Satoshi Yatsuda[4]
Saga Prefectural Agricultural Research Institute
Saga Prefecture, Saga City, Japan

***Abstract*—A method for agricultural product quality and harvest amount prediction by using smartphone camera image is proposed. It is desired to predict agricultural product quality and harvest amount as soon as possible after the sowing. In order for that, satellite imagery data, UAV camera based images, ground based camera images are used and tried These methods do cost significantly and these do not work so well due to some reasons, in particular, most of farmers cannot use these properly. The proposed method uses just smartphone camera acquired images. Therefore, it is totally easy to use. If the results of prediction of product quality and harvest amount are not satisfied, then farmers have to add some additional fertilizer at the appropriate time. The experimental results with soy plantations show some possibility of the proposed method.**

Keywords—Smartphone camera image; agricultural product quality and harvest prediction; fertilizer control; soy plantation

I. Introduction

Isoflavones are most specific feature of soy. Therefore, soy farmers intend to find species which is isoflavone rich species and also intend to control water supply, fertilizer etc. for improving soy quality and harvest amount. Because isoflavones are ubiquitous in the germ part. Therefore, it is not so easy to estimate the isoflavone content in the planted soy.

Nitrogen content estimation of rice crop based on Near Infrared (NIR) reflectance using artificial neural network (ANN) is conducted [1]. Rice crop field monitoring system with radio controlled helicopter based near infrared cameras through nitrogen content estimation and its distribution monitoring is proposed [2]. Rice crop quality evaluation method through regressive analysis between nitrogen content and near infrared reflectance of rice leaves measured from near field radio controlled helicopter is also proposed and well reported [3]. Furthermore, a relation between rice crop quality (protein content) and fertilizer amount as well as rice stump density derived from helicopter data is well reported [4]. Then estimation of rice crop quality and harvest amount from helicopter mounted NIR camera data and remote sensing satellite data is proposed [5].

Effect of stump density, fertilizer on rice crop quality and harvest amount in 2015 investigated with drone mounted NIR camera data is also confirmed [6]. Method for NIR reflectance estimation with visible camera data based on regression for NDVI estimation and its application for insect damage detection of rice paddy fields is discussed [6]. Artificial Intelligence: AI baaed fertilizer control for improvement of rice quality and harvest amount is proposed and well validated the proposed method with rice paddy field data. [7]. Method for NIR reflectance estimation with visible camera data based on regression for NDVI estimation and its application for insect damage detection of rice paddy fields is proposed [8]. Also, artificial intelligence baaed fertilizer control for improvement of rice quality and harvest amount is proposed [9].

It is not always possible to acquire the remote sensing satellite data due to the limitation of revisit cycle of the satellite orbit. Neither, it is not possible to acquire UAV camera images due to the limitation of wind speed, weather condition, and so on. On the other hand, ground based cameras have problems on sun illumination condition difference, shading and shadowing, etc. Moreover, these methods do cost very much.

Meanwhile, the proposed requires only smartphone camera derived images. Therefore, it is easy to use and does not require much cost. Only thing the farmers have to do is just acquire images of the soy plantations (example of agricultural products) and send these to the image collection center. After the farmers send their acquired images, the image collection center analyzed data, and predict their product quality and harvest amount and send the predicted result with some appropriate instructions for fertilizer control (fertilizer amount and timing) to the farmers.

In the following section, the proposed prediction method and system will described followed by experimental set-up together with experimental results. After that, concluding remarks and some discussions will be described.

II. Proposed Method

A. System Configuration

Fig. 1 shows the proposed system configuration and the procedure. Following is the detailed procedure:

1) Farmer sends smartphone images to the image collection center.

- Register the system
- Report conditions

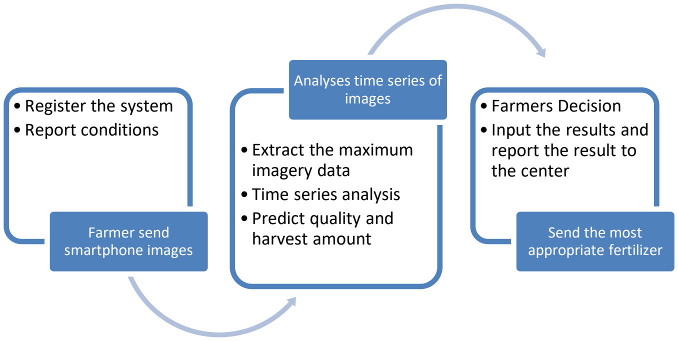

• Register the system
• Report conditions
Farmer send smartphone images
Analyses time series of images
• Extract the maximum imagery data
• Time series analysis
• Predict quality and harvest amount
• Farmers Decision
• Input the results and report the result to the center
Send the most appropriate fertilizer

Fig. 1. Proposed System Configuration and the Procedure.

2) Analyses time series of the acquired smartphone images

- Extract the maximum imagery data
- Time series analysis
- Predict quality and harvest amount

3) Send the most appropriate fertilizer control

- Farmers Decision
- Input the results and report the result to the center

B. *Method for Prediction of Harvested Soy Products and Harvest Amount*

In this regard, the prediction methods of product quality and harvest amount are key issues here. These methods are based on the experimental results which are described in the next section. These are basically based on a correlation analysis, and a linear regressive analysis. In the year 2018, some experiments were conducted at the soy farm areas which are situated at the Saga prefectural research institute. The experimental data encourages us to use the correlations between the maximum image pixel value and isoflavone content and protein content in the harvested soy beans as well as harvest amount. Therefore, the results of the regressive analysis are reliable.

Essentially, the proposed methods for soy bean quality evaluation and harvest amount prediction are based on regressive analysis by using smartphone acquired camera imagery data.

III. Experiments

A. *Experiment Procedure*

Sowing time, fertilizer time and flowing time of the two types of soy (Sakukei 207: New specie, and the traditional Fukuyutaka) are shown in Table I. Smartphone camera data are acquired on the following dates:

July 26, August 17, 22, 27, September 11, 25 in 2018

TABLE. I. The Time for Sowing, Fertilizer, Flowing of the Two Species of Soy

	Sakukei207	Sakukei207	Sakukei207	Fukuyutaka
Sowing	June 4	June 27	July 12	July 12
Original fertilizer	July 24	Aug.14	Aug.24	Aug.27
Second	Aug.6	Aug.28	Sep.6	Sep.10
Third	Aug.20	Sep.11	Sep.21	Sep24
Flowing	July 23	Aug.12	Aug.22	Aug.25

In conjunction with camera data acquisition, spectral reflectance measurements and sample collection is conducted. The collected samples are used for the truth data of isoflavone, phospholipid, nitrogen content, water content of the harvested soy beans. The method for the chemical truth variables are as follows:

1) High performance liquid chromatography: The total amount of daidzin, glycitin, genistin and their respective aglycones, acetyls and malonyls, acetyls and malonyls were calculated as daidzin, glycitin or genistin and corrected with molecular weight.

2) Phospholipids as stearo, oleo and lecithin: It was converted by a factor of 25.4 from phosphorus determined by the colorimetric method.

B. *Intensive Study Area*

Intensive study area is situated at the Saga Prefectural Agricultural Research Institute as shown in Fig. 2. There are four lines of moth. From the right, Sakukei207 (6/4), Sakukei207 (6/27), Sakukei207 (7/12) and Fukuyutaka (7/12) are aligned in the soy farm area of test site.

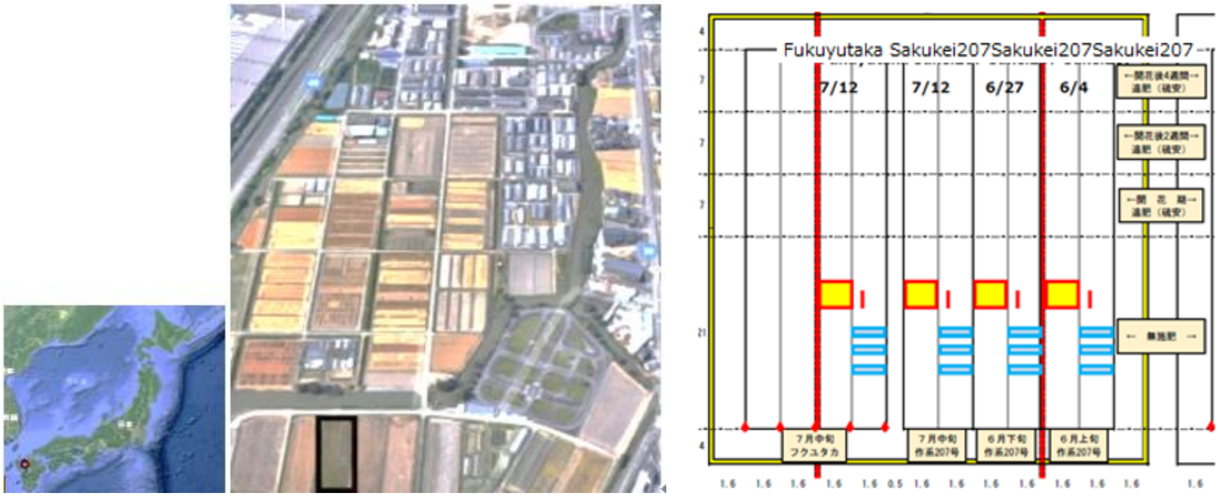
There are two types of cameras, visible camera and NIR filter attached camera. As shown in the previous paper, it is possible to replace NIR filtered camera with visible camera. Therefore, visible cameras are fine to collect the photos of the soy plantations for estimation of product quality and harvest amount.

C. *Acquired Smartphone Camera Images*

Examples of the acquired smartphone camera images are shown in Fig. 3. Each figure includes green color of histogram as well as mean, standard deviation, minimum and maximum pixel values of the rectangle areas. These statistics are shown in Table II.

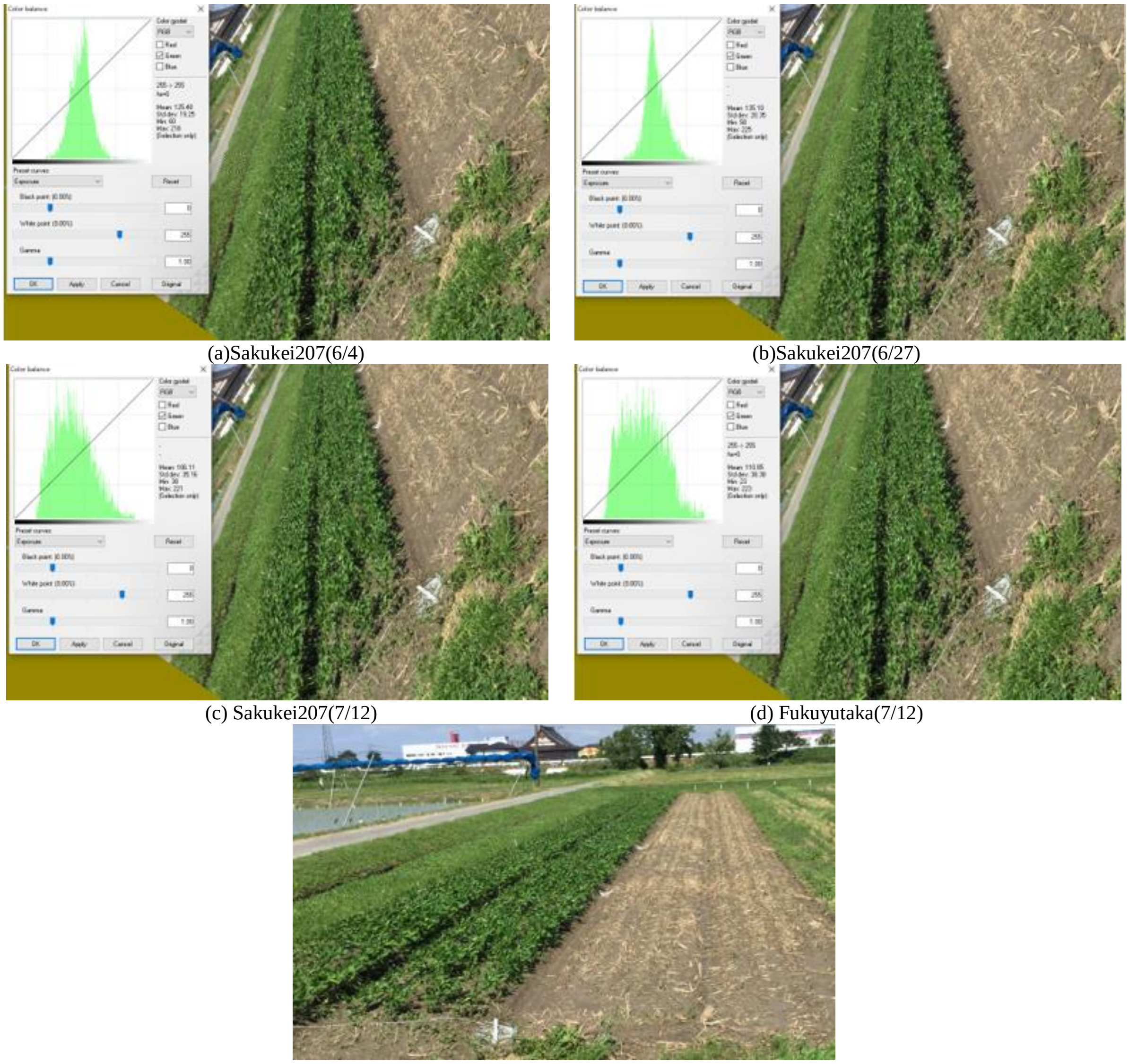
TABLE. II. Mean, Standard Deviation, Minimum and Maximum Pixel Values of the Rectangle Areas

	June 4 S	June 27 S	July 12 S	July 12 F
Mean	125.48	135.1	106.11	110.85
Std.	19.25	20.35	35.16	38.38
Min.	60	58	38	22
Max.	218	225	230	223



(a) Test Site on Map (Test Site is Situated at the Red Circle) (b) SPARI (c) Test Site on 3D Aerial Photo Image of Google Map.

Fig. 2. Test Site.



(a)Sakukei207(6/4)

(b)Sakukei207(6/27)

(c) Sakukei207(7/12)

(d) Fukuyutaka(7/12)

(e) Original Image

Fig. 3. Example of the Acquired Soy Plantation Images with Smartphone.

Where, “S” stands for Sakukei207 while “F” stands for Fukuyutaka, respectively.

D. Acquired Spectral Reflectance of Soy Plantations

Meanwhile, spectral reflectance of the soy plantations is measured. Therefore, NIR image (at around 800nm) of image can be estimated with visible colored smartphone camera images. Example of the measured spectral reflectance is shown in Fig. 4.

If just reflectance at 800nm of the soy plantation horizontally is taken into account, then Fig. 5 of horizontal profile of the measured reflectance of soy plantation can be shown.

These reflectance are measured on September 25 2018.

E. Measured Truth Data

The measured truth data are as follows:

Soy isoflavone, Phospholipid, Nitrogen content, water content and protein content in the harvested soy beans and harvest amount. Firstly, the following correlation analysis is conducted with the measured reflectance of the soy plantation and protein content, water content, as well as harvest amount.

F. Correlation Analysis and Linear Regression Between Smartphone Camera Derived Reflectance at 800nm and the Truth Data

The relation between the measured reflectance of soy plantation at 800 nm and water content, protein content of the harvested soy beans and harvest amount is shown in Fig. 6. From the figure, it is found that there is not so small correlation between the measured reflectance and protein content and also harvest amount. Therefore, it is possible to predict these two parameters (harvest amount and protein content) from reflectance measurement with not only spectral-radiometer, but also smartphone camera. Even if the spectral coverage of the smartphone camera rages from blue to red, it is still possible to estimate the reflectance at 800nm with visible smartphone camera if a calibration between visible camera and spectral-radiometer is conducted. On the other hand, there is no such correlation between the measured reflectance at 800nm and water content in the harvested soy beans.

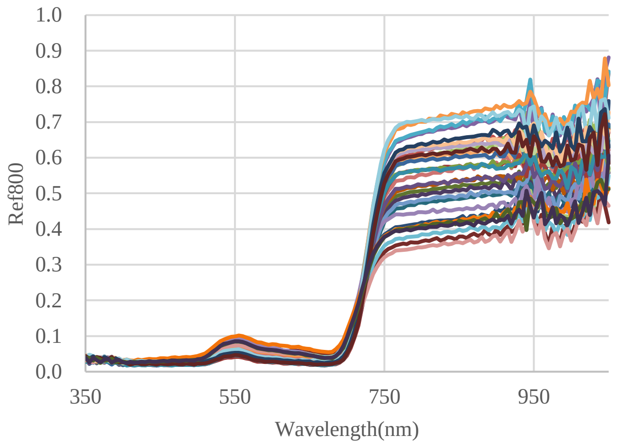


Fig. 4. Example of the Measured Spectral Reflectance of the Soy Plantations.

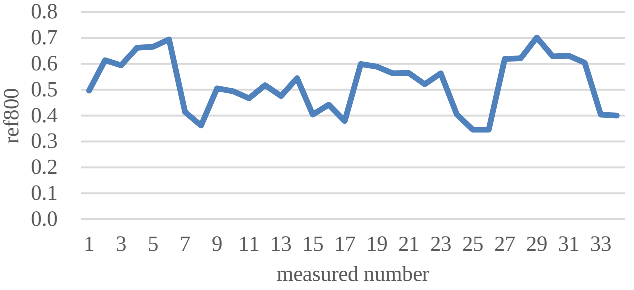


Fig. 5. Horizontal Profile of the Measured Reflectance of Soy Plantation.

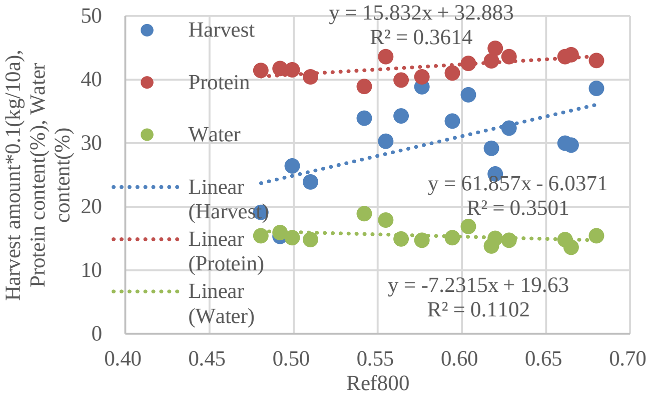


Fig. 6. The Relation between the Measured Reflectance of Soy Plantation at 800nm and Water Content, Protein Content of the Harvested Soy Beans and Harvest Amount.

The regressive equations and R square values for protein content in soy beans, harvest amount and water content in soy beans are as follows:

$$p = 15.832x + 32.883$$

$$R^2 = 0.3614 \tag{1}$$

$$h = 61.857x - 6.0371$$

$$R^2 = 0.3501 \tag{2}$$

$$w = -7.2315x + 19.63$$

$$R^2 = 0.1102 \tag{3}$$

Therefore, it can be said that protein content in soy beans and harvest amount can be predicted with the measured reflectance of the soy leaves at 800 nm in some sense (around 0.6 of correlation coefficient between both). Also, it is not possible to estimate water content in soy beans with the measured reflectance of the soy leaves at 800 nm.

It is also found that there is strong correlation between maximum pixel value and soy isoflavone and Phospholipids nevertheless the correlation between reflectance at 800nm and soy isoflavone and Phospholipids is week. The correlation between soy isoflavone is much stronger (0.821) than that between Phospholipids (0.309) as shown in Table III.

Because the pixel values in the smartphone camera image are variated due to the angle of the soy plant leaves and camera looking angle are so different, the maximum pixel value is much more appropriate for the correlation analysis.

TABLE. III. CORRELATION BETWEEN MAXIMUM PIXEL VALUE OF THE SMARTPHONE CAMERA IMAGE AND SOY IS OF LAVONE AND PHOSPHOLIPIDS

	Soy isoflavone (g/100g)	Phospholipid (g/100g)	y=(x-m)^2/s^2	1/y	Max.	R$_{800}$
Sakukei207 (June 4)with	0.34	1.6	0.104	9.592	218	0.6
Sakukei207 (June 27)	0.34	1.64	0.104	9.592	225	0.48
Sakukei207 (July 12)	0.43	1.63	3.072	0.326	230	0.58
Fukuyutaka (July 12)	0.32	1.61	0.615	1.626	223	0.66
Sakukei207 (June 4)without	0.34	1.64	0.104	9.592	218	0.68
Correlation R800	0.313	-0.162	0.454	-0.543		
Correlation max.pixel	0.732	0.309	0.821	-0.710		

As the result, it is found that it is possible to predict soy isoflavone content, protein content and harvest amount with the acquired smartphone camera image in the early stage of the soy plantation. In this case the smartphone camera image which is taken on July 26 2018 is used. It means that soy isoflavone content, protein content as well as harvest amount can be predicted 13 days after the sowing. Therefore, fertilizer control can be done properly and appropriately.

On the other hand, Table IV shows the result from the correlation analysis between the maximum pixel value and nitrogen content in the moth, seed and stem as well as the measured reflectance of soy plantation at 800nm and nitrogen content in the moth, seed and stem. As shown in Table IV, there is strong correlation between the maximum pixel value and nitrogen content in the moth as well as the maximum pixel value and nitrogen content in the seed.

However, it is found that the correlation between the maximum pixel value and nitrogen content in the stem is very week. Furthermore, the correlations with the maximum pixel values are much stronger than the measured reflectance at 800nm. Therefore, it is concluded that correlation analysis would be better to be conducted with the maximum pixel value of the acquired smartphone camera images rather than the measured reflectance at 800nm.

Also, it is found that it is possible to predict nitrogen content in soy plant moth and seed by taking smartphone camera image of the soy plantation in concern.

TABLE. IV. RESULT FROM THE CORRELATION ANALYSIS BETWEEN THE MAXIMUM PIXEL VALUE AND NITROGEN CONTENT IN THE MOTH, SEED AND STEM AS WELL AS THE MEASURED REFLECTANCE OF SOY PLANTATION AT 800NM AND NITROGEN CONTENT IN THE MOTH, SEED AND STEM

	Moth	Seed	Stem	Max.	R800
Sakukei207(June 4)with	1.19	6.92	0.72	218	0.6
Sakukei207(June 27)	0.97	6.44	0.73	225	0.48
Sakukei207(July 12)	0.93	6.63	0.69	230	0.58
Fukuyutaka(July 12)	0.76	6.8	0.51	223	0.66
Sakukei207(June 4)without	1.25	6.59	0.8	218	0.68
Correlation between R800	0.02357	0.396693	-0.18483		
Correlation between max.pixel	-0.67314	-0.41736	-0.28206		

IV. CONCLUSION

A method for agricultural product quality and harvest amount prediction by using smartphone camera image is proposed. It is desired to predict agricultural product quality and harvest amount as soon as possible after the sowing. In order for that, satellite imagery data, UAV camera based images, ground based camera images are used and tried These methods do cost significantly and these do not work so well due to some reasons, in particular, most of farmers cannot use these properly. The proposed method uses just smartphone camera acquired images. Therefore, it is totally easy to use. If the results of prediction of product quality and harvest amount are not satisfied, then farmers have to add some additional fertilizer at the appropriate time. The experimental results with soy plantations show some possibility of the proposed method.

It is also found that there is strong correlation between maximum pixel value and soy isoflavone and Phospholipids nevertheless the correlation between reflectance at 800nm and soy isoflavone and Phospholipids is week. The correlation between soy isoflavone is much stronger (0.821) than that between Phospholipids (0.309).

There is strong correlation between the maximum pixel value and nitrogen content in the moth as well as the maximum pixel value and nitrogen content in the seed. However, it is found that the correlation between the maximum pixel value and nitrogen content in the stem is very week. Furthermore, the correlations with the maximum pixel values are much stronger than the measured reflectance at 800nm. Therefore, it is concluded that correlation analysis would be better to be conducted with the maximum pixel value of the acquired smartphone camera images rather than the measured reflectance at 800nm. Also, it is found that it is possible to predict nitrogen content in soy plant moth and seed by taking smartphone camera image of the soy plantation in concern.

V. 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 authors would like to thank to Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

References

[1] Setia Damawan Afandi, Yeni Herdiyeni, Lilik B. Prasetyo, Wahyudi Hashi, Kohei Arai, Hiroshi Okumura, Nitrogen Content Estimation of rice Crop Basedon Near Infrared (NIR) reflectance Using Artificial Neural Network (ANN), Procedia Environmental Sciences, Elsevier, 33, 63-69, 2016.

[2] Kohei Arai, Osamu Shigetomi, Yuko Miura, Hideaki Munemoto, Rice crop field monitoring system with radio controlled helicopter based near infrared cameras through nitrogen content estimation and its distribution monitoring, International Journal of Advanced Research in Artificial Intelligence, 2, 3, 26-37, 2013.

[3] Kohei Arai, Rice crop quality evaluation method through regressive analysis between nitrogen content and near infrared reflectance of rice leaves measured from near field radio controlled helicopter, International Journal of Advanced Research in Artificial Intelligence, 2, 5, 1-6, 2013.

[4] Kohei Arai, Masanori Sakashita, Osamu Shigetomi, Yuko Miura, Estimation of protein content in rice crop and nitrogen content in rice leaves through regressive analysis with NDVI derived from camera mounted radio-control helicopter, International Journal of Advanced Research in Artificial Intelligence, 3, 3, 7-14, 2014.

[5] Kohei Arai, Masanori Sakashita, Osamu Shigetomi, Yuko Miura, Relation between rice crop quality (protein content) and fertilizer amount as well as rice stump density derived from helicopter data, International Journal of Advanced Research on Artificial Intelligence, 4, 7, 29-34, 2015.

[6] Kohei Arai, Masanori Sakashita, Osamu Shigetomi, Yuko Miura, Estimation of Rice Crop Quality and Harvest Amount from Helicopter Mounted NIR Camera Data and Remote Sensing Satellite Data, International Journal of Advanced Research on Artificial Intelligence, 4, 10, 16-22, 2015.

[7] Kohei Arai, Gondoh, Miura, Shigetomi, Effect of Stump density, Fertilizer on Rice Crop Quality and Harvest Amount in 2015 Investigated with Drone mounted NIR Camera Data, International journal of Engineering Science and research Technology, 2, 2, 1-7, 2016.

[8] Kohei Arai, Kenji Gondoh, Osamu Shigetomi, Yuko Miura, Method for NIR Reflectance Estimation with Visible Camera Data Bsed on Regression for NDVI Estimation and Its Application for Insect Damage Detection of Rice Paddy Fields, International Journal of Advanced Research on Artificial Intelligence, 5, 11, 17-22, 2016.

[9] Kohei Arai, Osamu Shigetomi, Yuko Miura, Artificial Intelligence Baed Fertilizer Control for Improvement of Rice Qualityt and Harvest Amount, International Journal of Advanced Computer Science and Applications: IJACSA, 9, 10, 61-67, 2018.

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.

Implementing a Safe Travelling Technique to Avoid the Collision of Animals and Vehicles in Saudi Arabia

Amr Mohsen Jadi
Collage of Computer Science and Engineering
University of Hail, Hail, Saudi Arabia

Abstract—In this work, a safe travelling technique was proposed and implemented a LoRa based application to avoid the collision of animals with vehicles on the highways of Saudi Arabia. For the last few decades, it has been a great challenge for the authorities to secure the life of animals and human being on the roads due to the sudden passage of animals on the highways. In such situations, drivers are not aware of the animal movement, and serious damage is observed with the life of both humans and animals. A LoRaWAN based architecture with a variety of advantages towards low cost and high accuracy of finding the movement of animals is possible with the proposed method and could deliver good results as well. The accuracy of this method was improved to a maximum extent as compared to the existing system due to the usage of LoRa sensors implanted in the animal's skin to trace with the nodes and base stations easily.

Keywords—LoRa; Sensor-based mobile applications; runtime monitoring; tracking; global positioning system

I. Introduction

The Middle East is facing a serious problem with the presence of camels on the roads/highways and is the reason for many accidents with an average of about 600-700 deaths in a year. Presence of the stray animals is not only the problem of the Middle East but also found to be a major concern in world-wide countries. In daylight somehow the people are still managing to escape, but in night times it is always a challenging risk to address carefully to save the animal-human life's and avoids the major damages. In this context, a lot of research was carried out by different authors to find reasonable solutions. In the early 70's, Al-Ghamdi highlighted the 30 times rise in road accidents in the Kingdom of Saudi Arabia (KSA) [1]. Later in 80's, it was reported by Tamimi et al., that most of the accidents are due to driver errors (around 90%) and they are between the ages of 20 and 40 [2]. Later, Qayed reported a huge number of accidents (around 6,117) in Saudi Arabia from June 1994 to 1995 and among which 2,551 people got injuries and 151 lost their lives [3]. The author in his work explained that due to vehicle collision with other vehicle and fixed objects is mostly seen. However, the collision of vehicles with animals in the night times proved to be a critical issue to be considered carefully on the highways. In a statistical analysis carried out by Ansari et al. revealed that 564,762 people (which is equivalent of 3.5% of the total population in KSA) were died or injured in road accidents [4]. Later, the government of KSA carried out stringent action against the rash driving citizens and even penalized for any kind of death or injuries caused to the animals. However, this was misused by some of the camel owners to earn money by pushing their animals on roads as reported by Ansari et al. [5]. This has become a serious issue for the governments and the commuters to overcome as it is created the problem. Similar problems of a motor vehicle collision with Kangaroo's were reported in Australia in the night times by Abu-Zidan [6]. Surprisingly the rate of accidents in winter due to the kangaroo-vehicle collision was reported to be very low in Australia. The vehicle collisions with the larger animals generally cause severe trauma to occupants due to fatal attacks caused by the wildlife animals [7]. Therefore, to address these kinds of animal-vehicle collision (AVC) or camel-vehicle collision (CVC) problems various countermeasures have been introduced in the past by many researchers with the help of the government in Saudi Arabia [8]. Similarly, Bashir and Abu-Zidan proposed different types of preventing techniques to avoid motor vehicle collision (MVC) with large animals [9]. This includes alarming signs, underpasses or overpasses for animals, frightening reflectors, etc.

A lot of data was collected by Huijser to study different types of reasons for AVC revealed that sex and age of the animals are also playing a vital role in the AVC. Sometimes the animals become more protective and aggressive to save their family or even to entertain themselves [10]. Later, Huijser and Kociolek presented a detailed summary with the highlights related to issues, problems, and experiences of different operations such as false positives, false negatives, maintenance, etc. [11]. It was also noted that most of the accidents found to be with the vehicle speeds of above 88 km/h (55mi/h) and proposed to alert the drivers to reduce speed with a distance of almost 21 m (68 ft). A lot of research has been carried out in this work towards wildlife-vehicle collision (WVC) and provided different types of mitigation measures to prevent accidents due to animals. The author also discussed different types of animal detection systems and wildlife crossing structures that influence towards reducing the number of accidents in his work [12]. Similar work was carried out by Al-Shimemeri and Arabi on Arabian camels and highlighted the types of injuries that occur due to the collision with the Arabian camels with a weight of over 726 kg [13]. Many camels tend to sit and relax on the highways in the night times and are killed by accident or by some of the ruthless drivers as shown in Fig. 1.

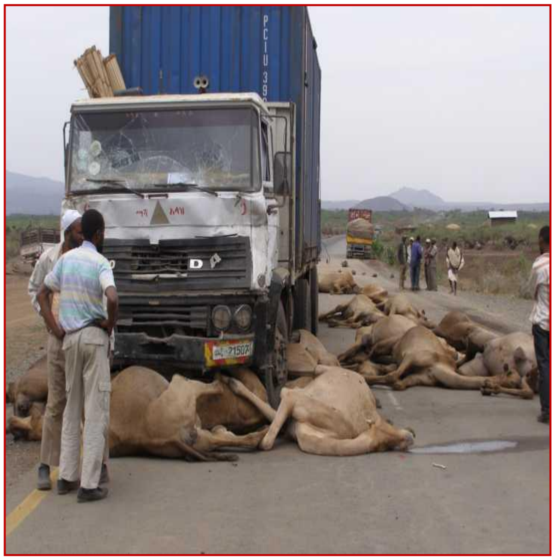


Fig. 1. Shows the Huge loss of Camels due to Accidents with Vehicles on the Highways/Roads.

Global positioning system (GPA) based intelligent Camel-Vehicle Accident Avoidance system (CVAAS) was introduced by Zahrani *et al.* for the practical usage on the roads/highways of Saudi Arabia [14]. This work was simulated by following the European Telecommunications Standard Institute (ETSI) frequency allocation of frequency between 5875 to 5905 MHz [15]. This system helps to detect the position of the camels, their movement and direction on their travelling. Based on these parameters a warning will be given to the drivers on the roads/highways. They classified the roads/highways into zones with respect to the distance of the camel's movement and accordingly the alarm system will be changing to alert the commuters. The authors claim to help the SA government to saving billions of Saudi Riyals. In recent times, the KSA is considered to be a developed nation with respect to the infrastructure, roads, and network availability [16]. In such circumstances, the role of mobile phones along with the GPS may help most of the commuters to avoid the accidents due to the collision of vehicles and camels in SA, since KSA tops the world for having the maximum number of mobile phone users [17]. The usage of phosphorous bands in the necks and bodies of the camels initiated by the Oman officials proved to be a bit an effective approach to some extent [18]. However, the drivers with very high speed tend to face the same challenges with these camels with phosphorous bands as well. Some of the challenges due to the camels are shown in Fig. 2.

Now it is reported statistically that SA is spending 13 billion Saudi Riyals annually to treat road-traffic related accidents according to DeNicola *et al.* [19]. The biggest concern for the SA government is towards the projected growth in road accidents due to animals crossing the roads is going to be more than the present conditions. Recently, it is observed that AVCs have been reduced to some extent by installing the animal crossing warning signboards on the highways and most of the drivers are being trained to get accustomed to the W11-3 and W11-4 signs [20]. Hosseini *et al.* suggested three important areas to be sorted out for the AVC problems: a) camel identification problem, b) irresponsible sailors, and c) unidentified camels by the drivers on the roads/highways. The camel owners can be active members in solving these issues to some extent by taking some of the precautionary steps as suggested in the past by some of the authors by providing phosphorous neck bands, body bands, etc. However, the loose animals (especially loose camels) are going to be a big threat to the commuters and these things needed to be addressed carefully.

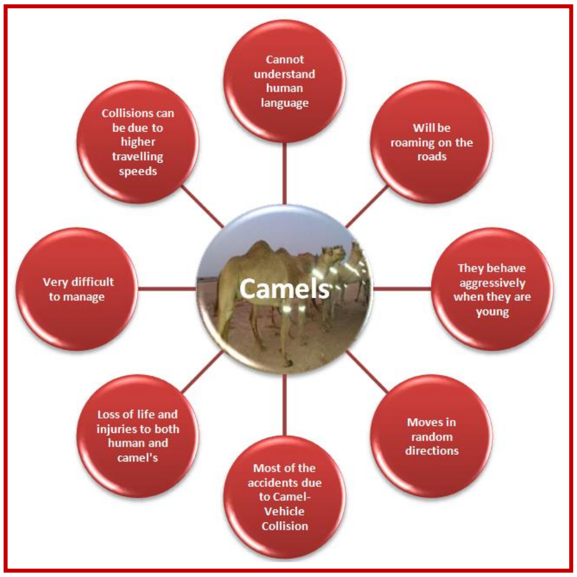

Cannot understand human language
Will be roaming on the roads
Collisions can be due to higher travelling speeds
Camels
Very difficult to manage
They behave aggressively when they are young
Loss of life and injuries to both human and camel's
Moves in random directions
Most of the accidents due to Camel-Vehicle Collision

Fig. 2. Shows the different Challenges due to Camels in the Deserts that Cross the Roads/Highways.

In this paper, Section II explores the existence methods and different types of AVC. It tries to explain different types of methods and technologies used for avoiding the collision. Section III explains the proposed architecture and the internal components involved with their role to obtain the desired results to avoid the collision of animals and vehicles. Section IV deals with the functioning of the architecture and implementation methods are discussed in detail, followed by the last section to conclude the overall work carried out in this paper.

II. Existing Methods to Avoid the Animal Vehicle Collision (AVC)

There are so many ways already introduced to avoid the animal-vehicle collision (AVC) based on a) road-based technologies, b) animal-based technologies, and c) vehicle-based technologies [15]. All these technologies try to minimize the collisions on the roads and aimed to save human and animal lives. The classification of these technologies can be seen in Fig. 3, where some of the latest technologies are introduced as compared with the work of Ragab [15]. Only GPS was introduced in Ragab's work in the category of vehicle-based technology. However, few more technologies are introduced in this paper such as mobile technology based on code division multiple access (CDMA)/orthogonal frequency division multiple access (OFDMA), and sensor-based mobile technology and LoRa (long range) wireless radio frequency (RF) technology.

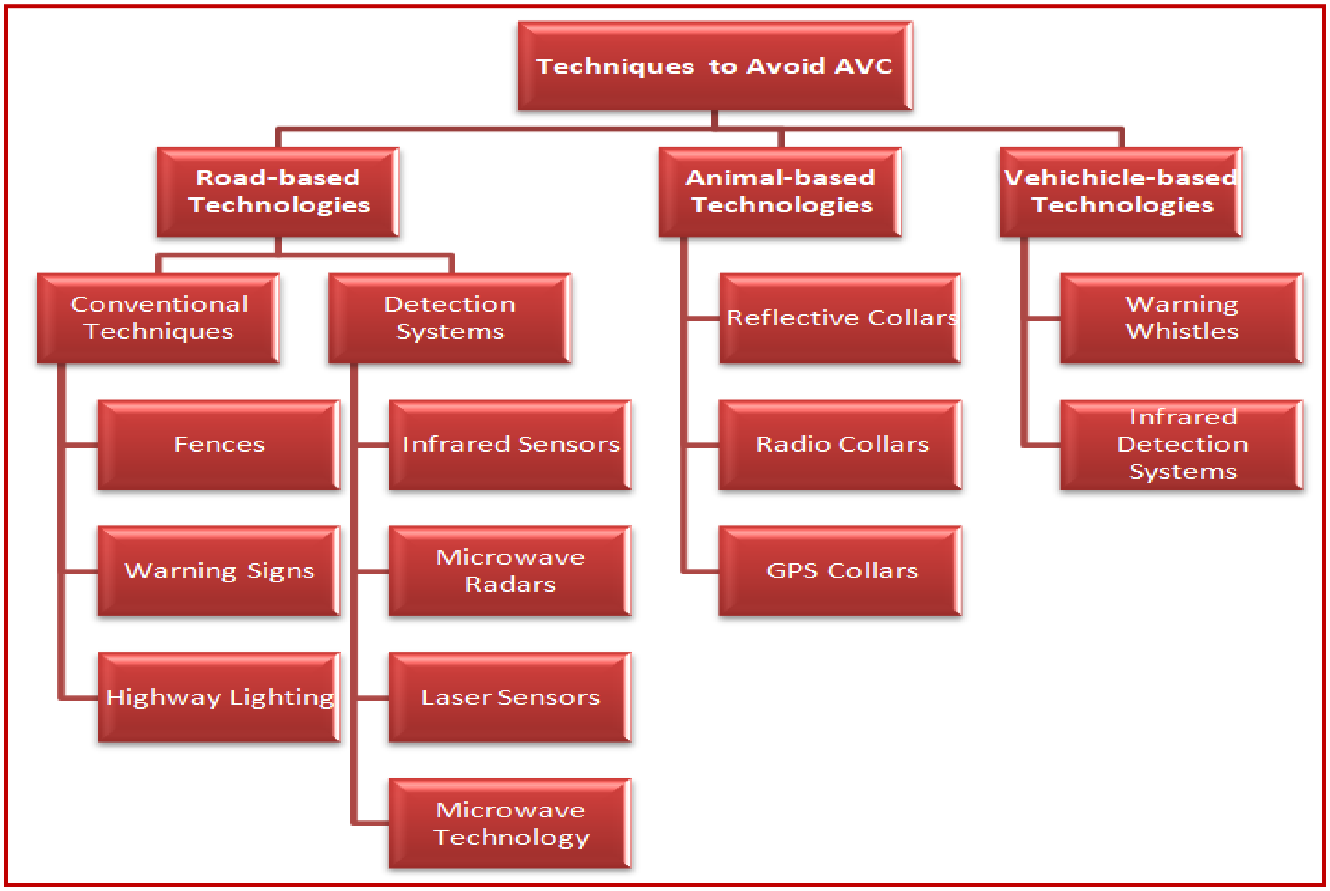


Fig. 3. Shows different Types of Techniques to Avoid AVC on the Roads/Highways.

A. *Road-based Technologies*

The road based techniques are classified into two categories as a) conventional techniques and b) detection systems.

- **Conventional Techniques:** To avoid collision between animals and vehicles, the ***fences*** are used to avoid the passage of animals through the roads/highways. The height of the fence must be at least 2m to avoid any kind of jumping [21]. However, this method needs regular maintenance and checkups to avoid any kind of exploit breaks with the massive groups of animals hitting the fencing arrangements. At some places, ***warning signs*** are installed to inform the commuters about the frequency of animal movement with different types of signals/indicators/direct short message on the signboards. This method helps the commuters to slow the vehicle speed and pass the roads carefully with more attention [15]. Finally, the ***highway lighting*** is used from sunset to sunrise for easy passage of the commuters with clear visibility and proper roadside instructions at various junctions about the level of seriousness with the moving animals on the roads/highways along with the signboards [22].

- **Detection Systems:** There are many methods to detect the animals passing through the roads using sensing devices and technology-based devices/equipment. There are special categories of vehicle detection systems available in the market, which only identifies the vehicles based on the specifications loaded into their database [23]. These systems cannot detect animals. In such cases, the animals are alerted with the help of long-range audio and visual signals from the designated locations on the roads/highways. In this method, the probability of animals ignoring the caution signals may be noticed.

In the other cases, the ***infrared sensors*** are used to identify the animals within a range of 30-100 m radius on the roads will be installed, so that on the animal detection some kind of signals are being activated to alert the drivers about the animal movement on the roads [15]. The possibility of false detection in this system is more due to broken sensors, loss of power due to improper functioning of solar panels, broken lamps, etc. due to heavy winds and fog during the nighttime with heavy cold environments. Similarly, ***microwave radar sensors*** are installed at some locations to identify the movement of larger animals up to 50m at a distance with 60^0 horizontal angles. Symbols of animals with LED displays are turned ON when the identification of animals will take place [15]. The possibility of false detection during the winter is more due to heavy snow and fog at some of the cold places.

Apart from the above two methods, ***laser sensors*** also installed at some locations of Washington, USA, with a pair of lasers installed on both sides of the roads/highways. Whenever a rectangle shaped vehicle passes the alert will not be given but when the shape of the animal (non-rectangle) is identified means an alert will be given to the drivers with the help of broken lights. However, in this method, if a deer stands between the laser beams for a longer time (more than a minute) means the warning signals will be turned OFF. The biggest drawback of this method is to use only for the shortest distance and requires high maintenance cost. Finally, microwave technology using a series of transmitters and receivers animal detection is possible using continuous microwave RF signals [27]. Using these method animals are alerted when a vehicle is detected passing through the road by using the variety of noise or light signals. But this system is not useful for high vehicle density roads because the noise will be ringing continuously with the passage of the vehicles [15].

B. Animal-based Technologies

Different types of collars are introduced in this method to mitigate the AVC by providing a blinking signal system. They are specially classified into two types: a) reflective collars and b) radio collars. These two collars require a minimum infrastructure requirement and are easily available in the market. However, based on the GPS systems also the collars were installed on the animal's neck or at other body parts. These systems can cover a long range of distance and a massive range of animals can be covered using this system [15].

- **Reflective Collars:** There are many methods to detect the animals passing through the roads using sensing devices and technology-based devices/equipment. There are special categories of vehicle detection systems available in the market, which help the drivers to identify the animals from a long distance due to the reflective tape on the neck of animals. But these collars need to be maintained properly by the owners of the animals so that they are visible from distance and avoidance of collision is sometimes not possible if the distance is very high [28].

- **Radio Collars:** The first time they were introduced in the year 1999 in Olympic Peninsula, Washington. The animals were radio-collared and the receivers used to scan the frequencies of different radio collars for the whole day. In the presence of a signal at a particular radio-collar within a distance of 400 m of the roadside, a signal activates the flashing beacons. However, due to the operation of these radio collars is for 24/7, the batteries run out after several years and must be replaced, which is a serious concern using these radio collars [15].

- **GPS Collars:** Movement of a large number of animals can be traced out by this method and can be helpful in collecting a massive amount of the data to analyze the tendency of the animal movement in different scenarios and meteorological conditions.

C. Vehicle-based Technologies

In the early stages, the technology was used in two ways to avoid AVC by using: a) the warning whistles, and b) the in-built infrared detection system.

- **Warning Whistles:** They are independent of any kind of installations, such as the roadside equipment, etc. This concept was introduced in the late 70's [24]. The warning whistles are of air activated type and will be mounted on the front side of the vehicles. These whistles produce ultrasonic frequency waves for any kind of wind rushing towards them. The sounds made by these whistles scare away the animals and the animals are supposed to run away from the roadside. But it is reported that the audio signals will make any effect on the behaviour of the animals [10].

- **Infrared Detection System (IDS):** In this case, the drivers are alerted with the help of infrared sensors connected to the vehicle when the animals are identified within a range of distances. However, the range of these sensors must be within the reach of animals and must be sufficient enough to allow the drivers to apply the brakes within the time [25]. In this method, a monochromatic display will be used to see the hot objects in white colour and cold objects as black in the images. This method helped the drivers to some extent but some people reported to have headaches. Apart from these, there are many maintenance and technical problems while using IDS. Also, false positive and false negative problems as discussed earlier are common to find by using this method [26].

III. Proposed Architecture

At present few countries have got their own GPS satellites includes the US, Russia, European Countries, China and India. Therefore it will be a great deal to adopt the GPS systems in its direct form as a complete solution to resolve the problem in Saudi Arabia due to a huge investment. Therefore in this work, a LoRa based approach is proposed to deal with the collision between animals and vehicles in the proposed system with the help of GPS based Google Maps and sensor-based mobile technology to initiate the alerts.

In the recent times, there are so many advanced technologies evolved into the market that is very much suitable to use for implementing the safe travelling methods for the commuters in Saudi Arabia by avoiding the collision with animals. There are seven technologies in the market that are helpful to track the location. They are Beacons, Wi-Fi, Radio-Frequency Identification (RFID), Near-Field Communication (NFC), Quick Response Codes (QR), LoRa and GPS as listed in Table I with different technological parameters. The range, cost, accuracy, and security of these technologies play a key role to define the type of application where they can be implemented with low risk. There is nothing like a winner or a loser for the tracking game out of all the following tracking technologies [29]. In recent times, the accuracy of GPS proved to be increasing and Bluetooth beacons are well equipped to transfer the larger data. Above and all the usage of mobile phones increased drastically with the increased competition between mobile manufacturers. Most of these technologies proved to be excellent for tracking the people in indoor locations.

TABLE. I. CLASSIFICATION OF DIFFERENT LOCATION TRACKING TECHNOLOGIES WHICH CAN BE USED FOR SAFE TRAVEL USAGE

Technology	**Symbol**	**Range**	**Cost**	**Accuracy**	**Accessibility**	**Security**
Bluetooth Beacon	ᛒ	1 m – 100 m	★★★★★ (2)	★★★★★ (3)	★★★★★ (4)	★★★★★ (3)
Wi-Fi	📶	20 cm – 100 m	★★★★★ (3)	★★★★★ (3)	★★★★★ (3)	★★★★★ (2)
RFID	RFID	1 cm – 100 m	★★★★★ (4)	★★★★★ (4)	★★★★★ (2)	★★★★★ (4)
QR Code	▦	10 cm or less	★★★★★ (2)	★★★★★ (5)	★★★★★ (4.5)	★★★★★ (3)
NFC	NFC	30 cm or less	★★★★★ (1)	★★★★★ (5)	★★★★★ (3)	★★★★★ (4.5)
LORA	LoRa	45 km or more	★★★★★ (1)	★★★★★ (1)	★★★★★ (1)	★★★★★ (1)
GPS	🛰	Unlimited	★★★★★ (1)	★★★★★ (2)	★★★★★ (4.5)	★★★★★ (3)

Even till date, GPS is the most useful tracking technology as compared to others available in the market. The basic advantage of GPS is that they can be installed in smartphones and even on basic phones as well. Due to the unlimited range of GPS, it is always the best method to use along with mobile phones. Whereas in the case of Bluetooth, the upgraded Bluetooth 5 helped most of the mobile manufacturers to increase the range 4 times, speed by 8 times and enhanced the broadcast message capacity, but it cannot be used for any kind of animal tracking due to the range. Similarly, RFID, QR Code, and NFC are a bit more accurate but cannot be used due to the low range [29]. Now there is some scope by using the LORA technology due to its range and flexibility to install far locations easily with low power requirements and long battery durability. A detailed explanation of the proposed architecture based on four major technologies involved in this work is explained. The four technologies include: a) Microchip Implantations, b) GPS Based Technology, c) Sensor-Based Mobile Technology, and d) LoRa Wireless RF Technology.

a) Implantations: In recent times, most of the animals are implanted with microchips under the skin of the animals to track their pets. These microchips are very much capable of working for more than 25 years and these chips can be scanned by using the universal scanners produced by the microchip companies. These implants are used with RFID technology with a broad spectrum of frequencies, interfaces, devices, and protocols [30]. The usage of RFID tagging helped to track the animals successfully in the groups [31]. A tracking device suggested by Movers provides a microchip transmitter encapsulated in biological inert material [32]. The major disadvantage of this method is the range and accuracy of the tracking system when the animals are in masses. In the proposed architecture (as shown in Fig. 4), the animals (in case of SA the camels will be focused) will be implanted with LoRa equipped sensors with possible care using the cylinder like small microchips of bio-safe epoxy resin [30]. These chips are field powered and do not require any kind of power sources.

b) GPS Based Technologies: The GPS systems are used to track the position of a particular entity by installing a device in any kind of vehicle, cell phone, or special devices that are made for a special purpose. These systems use the global navigation satellite systems (GNSS) networks, which incorporate the range of satellites using microwave signals which are transmitted to the GPS devices giving the information of the location [33]. Now in the present problem scenario, the animals can be tagged with some kind of GPS collars in their neck or body to observe the movement of animals remotely and communicated the same with the help of mobile based applications. This system can be very much useful to access the location, speed of the animal movement, time and direction as well. The biggest advantage of this system is that it can cover longer ranges with minimum effort as compared to any other system. In the proposed technique, the GPS based Google Maps platform will be integrated with the newly developed application to get the information of animal movement on the screen of the mobile phones, as we see the traffic information in the Google Maps. For this, a Google ID has been created for developing the Google Map tracker and the refresh of the animal information will be done by using the Meta tag and jarring effects on the final map outputs. In this process, the base station will be getting the information of the GPS data with longitude, latitude, altitude, speed and time of the animal movements. This data will be converted into decrypt form and dumps the same into MySQL database and updates the same on the Google Maps.

c) Sensor-based Mobile Technology: There is an exponential rise in using smart-phones as a sensing device with all kinds of sensors (such as gyroscope, GPS, temperature, etc.) embedded within mobile phones. The introduction of the internet on mobile technologies made this combination more powerful to make it an emerging tool for various applications and as a real-time tracking tool as well [34]. This technology with a full arsenal of sensors makes this to be used for navigation, location-based services, mobility analysis, etc. However, for the current problem, one needs to have a customized device, which is suitable to tag with the animals travelling in extreme weather conditions as well. Some of the sensors existing in the mobile technology may not be very much suitable for extreme heat conditions of Saudi Arabia. Therefore, the selection of the devices and sensors must have a careful look before they are adopted in this problem statement. In the current architecture, the sensors are being used to generate an audio alert for the drivers when the movement of animals found to be communicated by the LoRa Base Station. This audio voice will be loud enough and will be OFF only when the driver sitting in the car press the message acknowledge button.

d) LoRa Wireless RF Technology: This is also known as a LoRaWAN, which is a low power wide area network (LPWAN) grabbing the attention of a large number of people across the globe. This is one of the latest technologies connecting the devices in rural areas and urban areas for up to 30 miles in indoor environments with low energy consumption [35]. The battery life of these devices can be up to 10 years; hence, reducing the battery replacement cost gives an edge for most of the tracking systems. This is one of the GPS-free tracking application, which provides low power benefits as compared to other technologies in the market. This technology can be implemented as the greatest tool for the animal tracking purpose due to the flexibility and adaptability with the devices in motion with low power consumption, reduced cost, minimum infrastructure, battery replacements and low overall operating expenses. It supports mobile technologies, millions of messages and requirements of public network operators to serve huge markets.

The biggest advantage of LoRa lies in its ability to provide efficient, flexible and with the reasonable economic solutions for most of the real-time problems in most of the rural and indoor applications, where most of the cellular and WiFi networks proved to be ineffective. It is a good choice for long range and Internet of Things (IoT) solutions with low power applications. It also enables different types of tracking applications in the absence of GPS and the LoRaWAN protocol helps to create the interoperability among different applications, IoT devices, and telecommunication operators. The architecture of a LoRaWAN consists of four major sections: a) end nodes, b) concentrator/gateway, c) network server and d) Application server as shown in Fig. 5.

- Functioning and Security of LoRa Architecture

Most of the IoT devices using the LoRa consist of a combination of the sensors with LoRa transceivers at the endpoints. These endpoints are connected in a star network and transmit the signals received from the sensors when they detect the movement of the devices consisting of LoRa sensors. The base station receives the information and passes the same through different gateways using standard IP connections. The data use different types of standard protocols to connect with telecom networks. The network servers manage the data based on the functions to eliminate the duplicated packets. The application servers will control the actions of the endpoints and/or collect data from the endpoints.

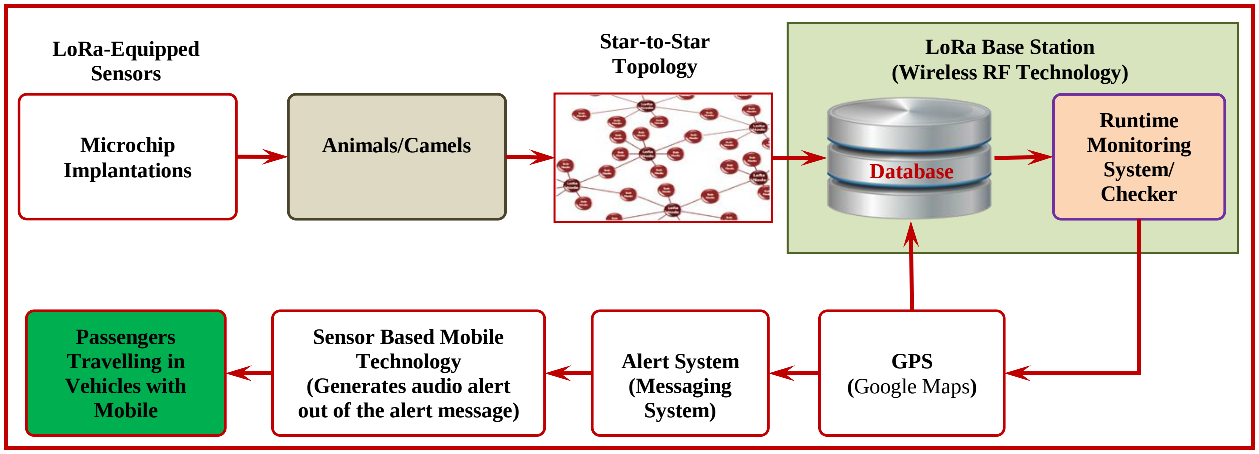

LoRa-Equipped Sensors
Microchip Implantations
Animals/Camels
Star-to-Star Topology
LoRa Base Station (Wireless RF Technology)
Database
Runtime Monitoring System/ Checker
GPS (Google Maps)
Alert System (Messaging System)
Sensor Based Mobile Technology (Generates audio alert out of the alert message)
Passengers Travelling in Vehicles with Mobile

Fig. 4. Shows the Proposed Architecture for Safe Travelling in Saudi Arabia.

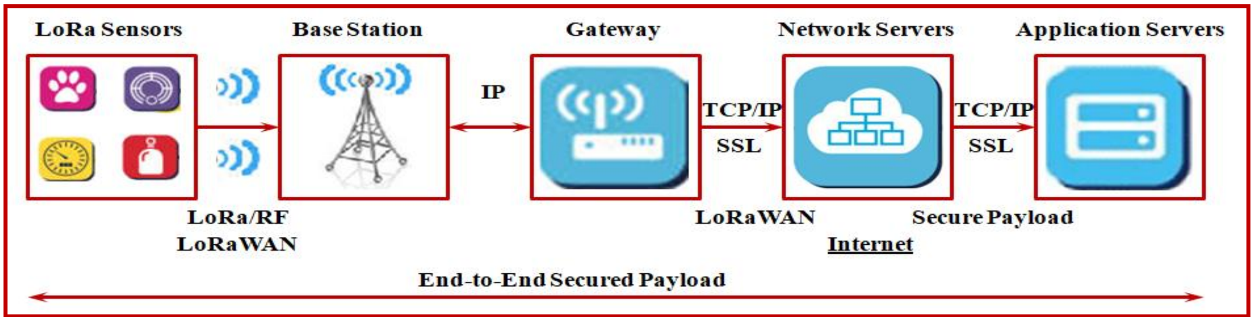


Fig. 5. Shows the LoRaWAN Architecture.

In the LoRa messaging, the messages are transmitted by using the LoRaWAN network by the end devices, which helps to improve the resilience of the network. Multiple base stations in an area may increase the deployment capital but also helps to enhance the performance. Multiple receptions are filtered at LoRa server, which also helps in providing the security checking. The security on LoRa network is provided by unique network key (EU164) at network levels, unique application key (EU164) at the application level and device-specific key (EU128) for the device levels.

IV. Functioning of the Proposed Architecture

The following steps will be implemented to monitor the movement of camels on the highways/roads of Saudi Arabia.

a) Installation of LoRa Base Stations to establish the communication between animals and Nodes (sub-nodes too): In this work, the animals are implanted with LoRa sensors into their skin in the first stage. On the other hand star topology-based networks are created with nodes and sub-nodes of specific identification techniques also. Throughout the area, these nodes and sub-nodes are covered using the LoRaWAN. All these nodes and sub-nodes are connected with certain base stations. Such a combination of nodes, sub-nodes and base stations are repeated throughout the highways at different places as a **Unit**. The nodes will communicate with the base stations when an animal (which is implanted with LoRa sensor) is identified within the different ranges as shown in Fig. 6. The range of a Unit will be between 10-300 m from the highway divided into three zones: **red** zone (0-10 m), **yellow** zone (10-100 m) and **green** zone (100-300 m).

b) Establishing communication between the base stations and GPS based mobile phones: The Units are connected with the GPS systems simultaneously giving specific information about each node to update the latitude, longitude, altitude, etc. on a regular basis. Google maps APK is installed with the present application which displays the sensor data received from the animals at different locations.

The runtime monitoring system will check the database for animal movements from the information received from the base stations (see Fig. 7). The Checker component will try to assess the information obtained from the base stations and identify the animal movements in the different zones.

c) Establishing an alert system using the proposed mobile application based on the distance between the animals and main roads: If the animals are within the green zone (i.e. safe zone) means there will not be any kind of alert. But if the animal movement is in the yellow zone (i.e. the possibility of animals rushing on to the highway is more) means there will be an alert to the drivers on their mobile application with a message and caution sound. However, if the animals are moving in the red zone (i.e. indicating a need for serious attention by the drivers) means there will be a message alert and the sensors of the mobile phone will activate a vibration with ring tone to alert the drivers to be more careful and reduce the speed of the vehicle. The sequence diagram with important components and their interaction with different components at different levels in the proposed architecture are shown in Fig. 8.

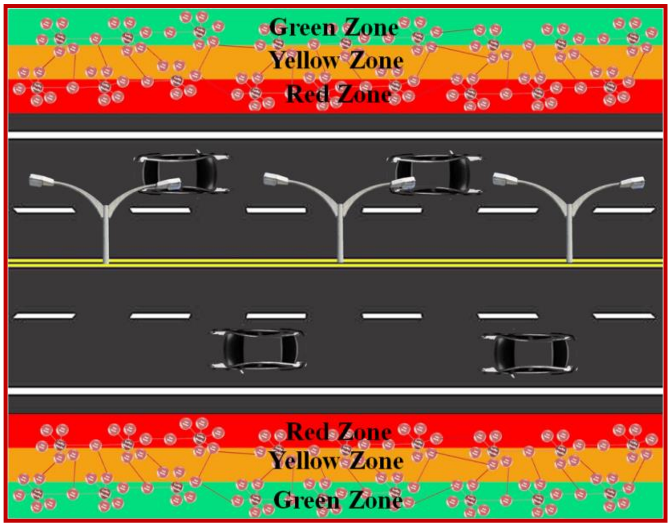


Fig. 6. Defined the different Zones across the Highways for Detecting the Animals with Respect to the Distance between the Road and Animals.

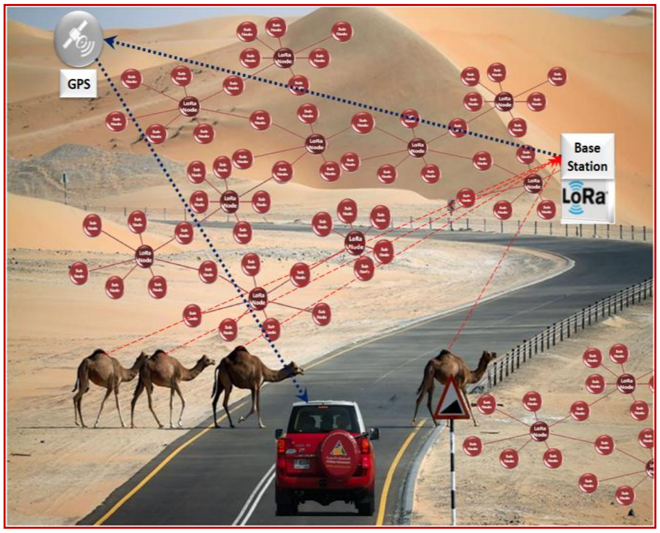


Fig. 7. Shows the Functioning of the Network in the Proposed Architecture.

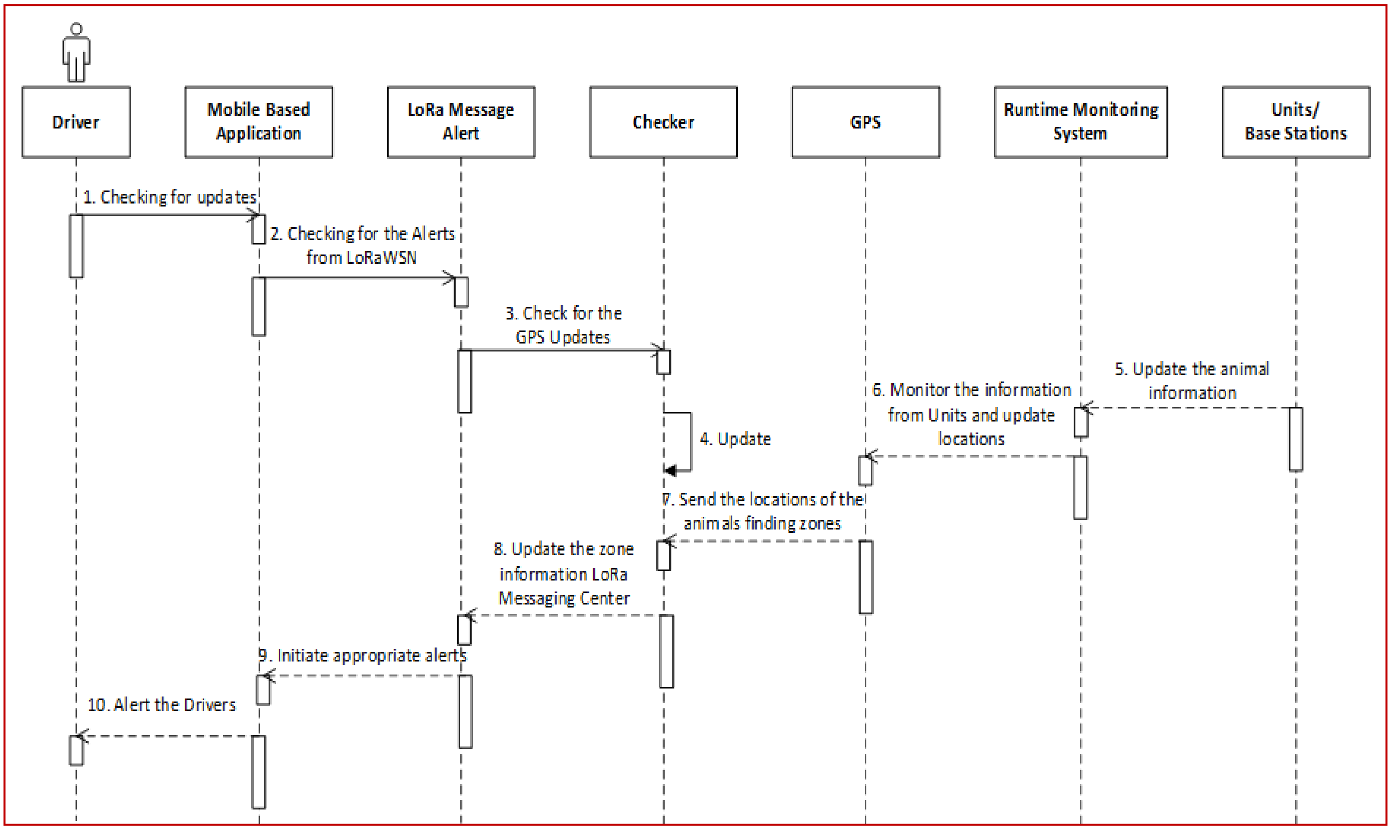

Driver
Mobile Based Application
LoRa Message Alert
Checker
GPS
Runtime Monitoring System
Units/ Base Stations
1. Checking for updates
2. Checking for the Alerts from LoRaWSN
3. Check for the GPS Updates
4. Update
5. Update the animal information
6. Monitor the information from Units and update locations
7. Send the locations of the animals finding zones
8. Update the zone information LoRa Messaging Center
9. Initiate appropriate alerts
10. Alert the Drivers

Fig. 8. Shows different Types of Component Interactions using a Sequence Diagram.

V. Conclusions

In the proposed method, the usage of GPS was limited by using LoRaWAN to establish communication between different nodes and sub-nodes after detecting the animals near the highways/roads. The implanted LoRa sensors are capable of working for more than 10-25 years as they are not dependent on any of the external sources. Google Maps platform was used along with the newly developed APK to monitor the sensor data on the mobile for any kind of animal movement near or on the highways. The information of the animal movement will be transmitted to the nearest base stations and the data will be analyzed for false alarms and false negatives in the process in the runtime monitoring system. The location of the nodes detecting the animals will be regularly updated with the GPS system and then based on the checker component analysis the zone wise information will be alerted in the form of LoRaWAN messaging services. The sensor-based mobile applications will be generating the alerts based on the zones as defined in the above discussions. The efficiency of the proposed method revealed satisfactory results by improving the alert quality and response as well. The overall system implementation is very much cheaper as compared with the GPS based systems for the conditions of Saudi Arabia without its own GPS system.

Acknowledgment

The author would like to thank first the Almighty, Allah for his grace and blessings. The author also thanks his parents, family and teachers for their unconditional support and encouragement throughout the career.

References

[1] S. Al-Ghamdi. "Road accidents in Saudi Arabia: a comparative and analytical study." WIT Transactions on the Built Environment 26 (1970).

[2] T. M. Tamimi, M. Daly, M. A. Bhatty, and A. H. M. Lufti. "Causes and types of road injuries in Asir Province, Saudi Arabia, 19751977: preliminary study." Saudi medical journal 1, no. 5 (1980): 249-256.

[3] M. H. Qayed. "Epidemiology of road traffic accidents in Al-Ahssaa Governorate, Saudi Arabia." (1998).

[4] S. Ansari, F. Akhdar, M. Mandoorah, and K. Moutaery. "Causes and effects of road traffic accidents in Saudi Arabia." Public health 114, no. 1 (2000): 37-39.

[5] S. A. Ansari, M. Mandoorah, M. Abdalrahim, and K. R. Al Moutaery. "Dorsal spine injuries in Saudi Arabia—an unusual cause." Surgical neurology 56, no. 3 (2001): 181-184.

[6] F. M. Abu-Zidan, K. A. Parmar, and S. Rao. "Kangaroo-related motor vehicle collisions." Journal of Trauma and Acute Care Surgery 53, no. 2 (2002): 360-363.

[7] T. P. Pynn, and B. R. Pynn. "Moose and other large animal wildlife vehicle collisions: implications for prevention and emergency care." Journal of Emergency Nursing 30, no. 6 (2004): 542-547.

[8] A. S. Al-Ghamdi, and S. A. AlGadhi. "Warning signs as countermeasures to camel–vehicle collisions in Saudi Arabia." Accident Analysis & Prevention 36, no. 5 (2004): 749-760.

[9] M. O. Bashir, and F. M. Abu-Zidan. "Motor vehicle collisions with large animals." Saudi medical journal 27, no. 8 (2006): 1116-1120.

[10] M. P. Huijser. Animal-vehicle collision data collection. Vol. 370. Transportation Research Board, 2007.

[11] M. P. Huijser, and A. V. Kociolek. "Wildlife-vehicle collision and crossing mitigation measures: a literature review for Blaine County, Idaho." Western Transportation Institute, Montana State University, Bozeman (2008).

[12] M. P. Huijser, K. J. Paul, and L. Louise. Wildlife-Vehicle Collision and Crossing Mitigation Measures: A Literature Review for Parks Canada, Kootenay National Park. No. 4W1929 A. Western Transportation Institute, College of Engineering, Montana State University, 2008.

[13] A. Al Shimemeri, and Y. Arabi. "A review of large animal vehicle accidents with special focus on Arabian camels." Journal of Emergency Medicine, Trauma and Acute Care 2012, no. 1 (2012): 21.

[14] M. S. Zahrani, K. Ragab, and A. U. Haque. "Design of gps-based system to avoid camel-vehicle collisions: A." Asian J Appl Sci 4, no. 4 (2011): 362-377.

[15] K. Ragab. "Simulating camel-vehicle accidents avoidance system." International Journal of Future Generation Communication and Networking 4, no. 4 (2011): 43-56.

[16] H. M. Hassan, L. Dimitriou, M. A. Abdel-Aty, and A. S. Al-Ghamdi. Analysis of Risk Factors Affecting Size and Severity of Traffic Crashes in Riyadh, Saudi Arabia. No. 13-2333. 2013.

[17] R. Al-Awaal. "KSA tops world's mobile phone users." Saudi Gazette (2014).

[18] A. A. Abdo, and A. A. Al-Ojaili. "Assessment of awareness of livestock-vehicle traffic accidents in Dhofar region, Oman." International Journal of, Applied Engineering Research (IJAER) 10, no. 18 (2015): 38955-38959.

[19] E. DeNicola, O. S. Aburizaize, A. Siddique, H. Khwaja, and D. O. Carpenter. "Road traffic injury as a major public health issue in the Kingdom of Saudi Arabia: A Review." Frontiers in public health 4 (2016): 215.

[20] M. Khalilikhah and K. Heaslip. "Improvement of the performance of animal crossing warning signs." Journal of safety research 62 (2017): 1-12.

[21] A. Ward. Lorin. Mule deer behavior in relation to fencing and underpasses on Interstate 80 in Wyoming. No. 859. 1982.

[22] S. M. R. Hosseini, D. Khorasani-Zavareh, and A. Abbasi. "Challenges and strategies for preventing vehicle collisions with camels in South Khorasan Province: a qualitative study." Safety Promotion and Injury Prevention 6, no. 1 (2018): 43-48.

[23] W. Saad, and A. Alsayyari. "Loose Animal-Vehicle Accidents Mitigation: Vision and Challenges." In 2019 International Conference on Innovative Trends in Computer Engineering (ITCE), pp. 359-364. IEEE, 2019.

[24] K. K. Knapp. Deer-vehicle crash countermeasure toolbox: a decision and choice resource. Midwest Regional University Transportation Center, Deer-Vehicle Crash Information Clearinghouse, University of Wisconsin-Madison, 2004.

[25] H. Bender. Deterrence of kangaroos from roadways using ultrasonic frequencies-efficacy of the Shu Roo. University of Melbourne, Department of Zoology, 2001.

[26] M. P. Huijser, P. T. McGowen, and W. Camel. Animal vehicle crash mitigation using advanced technology phase I: review, design, and implementation. No. FHWA-OR-TPF-07-01. Western Transportation Institute, 2006.

[27] K. Finkenzeller. RFID handbook: fundamentals and applications in contactless smart cards, radio frequency identification and near-field communication. John Wiley & Sons, 2010.

[28] M. M. Hurwitz. "Interchangeable attachments for collars, leashes, belts and accessories." U.S. Patent 8,142,053, issued March 27, 2012.

[29] Lighthouse. 2019. 6 Technologies that can be used to track location. [Online] available on URL: <https://blog.lighthouse.io/6-technologies-that-can-be-used-to-track-location/>, [June 21, 2019].

[30] Y. Grauer. A practical guide to microchip implants. [Online] available on URL: <https://arstechnica.com/features/2018/01/a-practical-guide-to-microchip-implants/>, [June 23, 2019].

[31] A. Weissbrod, A. Shapiro, G. Vasserman, L. Edry, M. Dayan, A. Yitzhaky, L. Hertzberg, O. Feinerman, and T. Kimchi. "Automated long-term tracking and social behavioural phenotyping of animal colonies within a semi-natural environment." Nature communications 4 (2013): 2018.

[32] M. G. T. Mowers. "Tracking device for pets." U.S. Patent 5,850,196, issued December 15, 1998.

[33] P. Bertagna. How does a GPS tracking system work? [Online] available at URL: <https://www.eetimes.com/document.asp?doc_id=1278363#>, [June 25, 2019].

[34] S. Plangi, A. Hadachi, A. Lind, and A. Bensrhair. "Real-Time Vehicles Tracking Based on Mobile Multi-Sensor Fusion." IEEE Sensors Journal 18, no. 24 (2018): 10077-10084.

[35] SEMTECH. What is LoRa? [Online] available at URL: <https://www.semtech.com/lora/what-is-lora>, [June 21, 2019].

AUTHOR'S PROFILE

Amr Jadi is an Associate Professor of Software Engineering at Collage of Computer Science and Engineering, University of Hail.

Dr. Jadi received PhD degrees from De Montfort University and Master's Degree from Bradford University, UK. The author is specialized in with an area interest in Software Engineering, Early warning systems, Risk management and Critical Systems. Presently the author is also involved in various development activities within the University of Hail and abroad as a consultant.

A Compact Broadband Antenna for Civil and Military Wireless Communication Applications

Zaheer Ahmed Dayo*[, 1], Qunsheng Cao*[, 2], Yi Wang[3], Saeed Ur Rahman[4], Permanand Soothar[5]
College of Electronic and Information Engineering
Nanjing University of Aeronautics and Astronautics (NUAA), 211100, P.R China[1, 2, 3, 4]
School of Electronic and Optical Engineering, Nanjing University of Science and Technology (NJUST), 210094, P.R China[5]

***Abstract*—This paper presents a compact broadband antenna for civil and military wireless communication applications. Two prototypes of the antenna are designed and simulated. The proposed antenna is etched on low cost substrate material with compact electrical dimensions of 0.207λ×0.127λ×0.0094λmm^3 at 2GHz frequency. The simple microstrip feeding technique and antenna dimensions are involved in the design to attain the proper impedance matching. An optimization of variables is carried out by multiple rigorous simulations. The designed antennas have achieved the broadband impedance bandwidth of 89.3% and 100% at 10dB return loss. The antennas exhibit omni directional radiation pattern at lower resonances and strong surface current distribution across the radiator. The peak realized gain of 5.2dBi at 10.9GHz resonant frequency is realized. Results reveal that the proposed broadband antenna is a better choice for WiMAX, UWB, land, naval and airborne radar applications.**

Keywords*—Compact antenna; broadband; microstrip feeding; civil and military; peak realized gain and impedance bandwidth

I. Introduction

In modern communication systems, the requirement of smart antennas is growing rapidly in the market. These antennas are economical, small in size, light weight with enhanced charactestics [1]. Recently, microstrip patch antennas with different shapes are good choice for different wireless communication systems. The narrower impedance bandwidth (BW) and larger physical and electrical dimensions are major concerns of patch antennas [2]. Therefore, antenna design engineers are working on the enhancement of important parameters of the compact patch antennas such as impedance BW, gain, stable radiation pattern and radiation efficiency.

Nowadays, researchers are paying attention on the design of simple structure antennas with enhanced features. Besides, these antennas can be used in civilian and military platforms. There are different wireless communication applications defined in the electromagnetic spectrum. These applications include WiMAX (3.5GHz-5.8GHz), H-Band (6GHz-8GHz), Ultra-Wideband (UWB) (3.1GHz-10.6GHz) and airborne, land and naval radars (8.5GHz-10.5GHz). These frequency bands are allocated for different wireless communication applications after the approval of Federal Communication Commission (FCC) [3].

Numerous studies about compact broadband antennas for different wireless communication applications have been proposed. The different shapes of radiating patch and modified ground plane were reported in [4] and [5]. Further, researchers have been proposed different techniques i.e. meta-material resonators, different shapes of slots, tuning stubs and proper selection of the feed line [6][7][8][9]. These techniques were used for the improvement in impedance BW. The existing study proposed the compact elliptical patch-based planar monopole antenna by embedding arc-shaped slot for UWB applications [10]. The Daisong Zhang and Yahya Rahmat Samii have been designed the antenna with top cross loop engraved on the compact substrate 0.345λ×0.575λ×0.02λ mm^3 at 3.43 GHz resonant frequency. The proposed antenna has achieved 91% of fractional impedance BW [11]. Another work was presented on the broadband antenna with parastic patch technique. The antenna exhibited the relative BW of 80% with compact dimensions of 0.521λ×0.521λ×0.012λmm^3 particularly at 2.32GHz operating frequency [12]. However, the designed antennas have the complex structures and larger dimensions. Moreover, Arash Valizade *et al.* demonstrated the protrude ground plane structures [13]. Jian-Feng Li *et al.* presented the idea on the isolation of antennas with T-Shaped slits in the antenna design structure [14]. Asghar Mousazadeh *et al.* presented the work on the broadband antenna with inverted L-shaped grounded strips. The antenna has the compact dimensions of 0.601λ×0.601λ×0.008λmm^3 [15]. The defected ground structure concept was utilized in [16] and [17]. Different broadband antennas were suggested in [18][19][20]. However, the employed techniques focused in the literature were complex and excessive variables utilized in the designed antennas might result in computational complexity. In the modern antenna topology, the compact size and adjustment in the designed antenna dimension in terms of variables is required. This adjustment can be achieved by electromagnetic (EM) simulation software which has the capability of the rigorous optimization.

Moreover, the authors have proposed a new palm tree structure wideband antenna. The antenna is capable to cover the 4GHz to 10.4GHz operable frequency range [21]. Kalyan Mondal *et al.* demonstrated the inverted question mark wideband antenna. The proposed antenna exhibited the good gain of 5.5dBi across the frequency span [22]. Further, the authors have presented the antennas which were capable to cover the different wireless communication applications [23].

Recently, the authors have presented the multiband antennas. The different feeding techniques were utilized and achieved the multiband characteristics [24] and [25]. A novel

*Corresponding Author.

miniaturized UWB antenna has been presented. The proposed antenna has the larger physical dimensions [26]. A high gain tapered slot antenna array was reported in [27]. The authors have achieved the high gain and substantial impedance BW with Wilkinson power divider approach.

In this paper we have designed the two models of the compact antennas and analyze their performance. The antenna design topology is very simple. The proposed antenna dimensions are calculated with the standard formulation. Moreover, the parametric study of different variables has been carried out. The performance of antenna parameters at multiple resonances is observed. The simulation results of prototype-I (reference antenna) and prototype-II (proposed antenna) is compared. The broadband fractional impedance BW of 100% at 10dB return loss has been observed. Moreover, substantial impedance BW improvement of 10.7% is analyzed. Finally, the proposed antenna design achieves near monopole like stable radiation pattern, maximum gain and strong current distribution across the surface of the antennas.

The key contributions of this manuscript are explained as:

- The designed models of the antennas exhibit the broadband impedance BW, good gain, strong current distribution and stable radiation pattern across the standard planes.
- The proposed antennas possess the compact physical and electrical dimensions.

The organization of the paper is categorized mainly in five sections. The antenna layout and mathematical strategy covers in Section II. The optimization of the variables and analysis of results such as peak realized gain, return loss, radiation pattern and current distribution of the antenna is explained in Section III. Summary of proposed antenna is stated in Section IV. Future work is elucidated in Section V.

II. Antenna Layout and Mathematical Strategy

The reference and proposed antenna model and their visualization from the front, back and side view perspective are depicted in the Fig. 1(a)-(c). The antenna designs are composed of compact patch; microstrip feed line and partial ground plane (PGP). A simple shape of radiator is engraved on the top surface of thick substrate with compact dimensions 31.7×19×1.4 mm^3. Low cost FR4 Epoxy laminate is used as substrate material with dielectric relative permittivity value ε_r=4.4 and dielectric loss tangent δ=0.02. Moreover, the proposed antenna is feed by 50Ω simple microstrip feeding line. The antenna is composed of three layer sheets, i.e. the first layer sheet is dielectric substrate, the second layer sheet consists of the compact patch and feeding line etched on top of the laminate and the third layer sheet covers the PGP etched on the back side of the laminate. These all elements are made up of copper clad material. The feed line has the dimension of 16.6×2 mm^2 which have great influence to achieve the proper impedance matching. The PGP is taken as optimized value for broader impedance BW. The variables of the radiator are adopted to adjust the return loss. Moreover, L-shape slots engraved on the upper side of PGP are used to realize the improved impedance BW.

Moreover, the prototype-II of proposed antenna is depicted in Fig. 1(b). It is operated between 3.5GHz-10.5GHz centered at 6.7GHz performing the operating BW of 7GHz. It can be observed that PGP, substrate thickness and the dimensions of L-shape slots have resulted in improved impedance BW. The proposed antenna is designed and simulated by the EM solver HFSS version 13.0.

The approximated initial calculated values have been obtained from the equations explained in this section. After multiple experimental simulations the optimized values of the designed antenna are listed in Table I.

The values of dimensions of patch is calculated by the following equations (1) and (2), respectively [28].

$$L_p = \frac{F}{\left\{1+\frac{2h_s}{\pi\varepsilon_r F}\left[\ln\left(\frac{\pi F}{2h_s}\right)+1.7726\right]\right\}^{\frac{1}{2}}}\times 2 \tag{1}$$

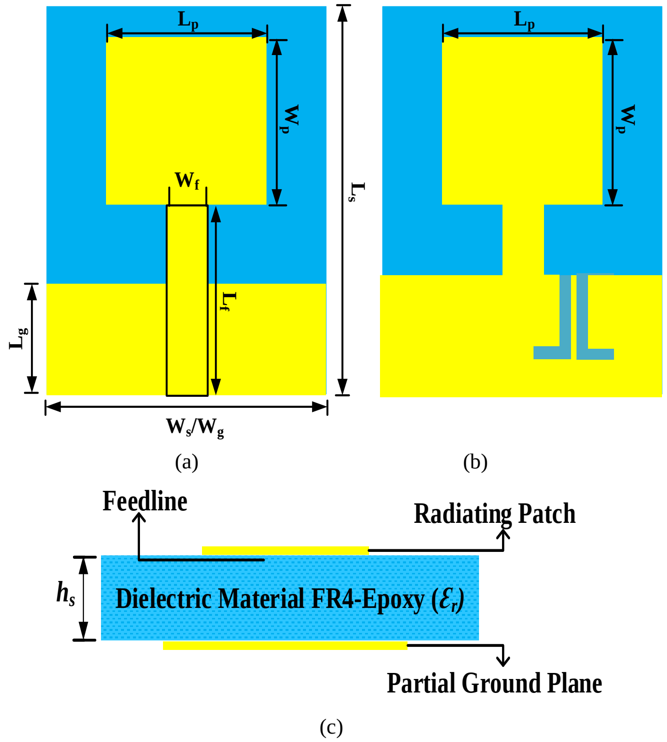


Fig. 1. (a) Top View of Prototype-I (b) Bottom View of Prototype-II (c) Lateral View of the Proposed Antenna.

TABLE. I. Proposed Antenna Defined Variables

Variable name	*Values (mm)*	*Variable name*	*Values (mm)*
L_p	9.5	L_{PGP}	14.25
W_p	9.5	W_{PGP}	19
L_{fl}	16.6	L_s	31.7
W_{fl}	2.0	W_s	19

The value of F can be calculated as follows:

$$F = \frac{8.79\times10^{9}}{f_r\sqrt{\varepsilon_r}} \tag{2}$$

In the above equations (1) and (2) variable h_s represents the thickness of dielectric substrate ε_r is relative permittivity, *F* represents the wavelength of substrate and f_r is resonant frequency.

Moreover, feedline width can be calculated by using the standard numerical equations (3) and (4), respectively [29].

$$\frac{W_{fl}}{h_s} = \frac{8e^{A}}{e^{2A}-2} \tag{3}$$

Where variable A can be calculated as:

$$A = \frac{Z_0}{60}\left(\frac{\varepsilon_r+1}{2}\right)^{1/2} + \frac{\varepsilon_r-1}{\varepsilon_r+1}\left(0.23+\frac{0.11}{\varepsilon_r}\right) \tag{4}$$

Z_0 is the characteristic impedance.

III. Simulated Results and Analysis

In this section impedance matching performance related to variables used in the antenna design is explained. Moreover, the results of return loss *(S_{11})*, peak realized gain *(dBi)*, surface current distribution *(Jsurf)* and radiation pattern are also discussed and analyzed.

A. Parametric Study

This section investigates the impact of the feeding line length *(L_f)* and width of patch *(W_P)*, Effect of PGP length *(L_{PGP})* and width *(W_{PGP})*. These effects realize the matching performance of proposed antenna. Moreover, parametric study of the proposed antenna in terms of variables is accomplished by running the multiple times rigorous simulations. The effects of different values of variables are observed. Finally, the optimized values are chosen to validate the proposed antenna prototype.

1) Variation in feedline(L_f) and Patch (W_p): Microstrip feeding line is key part of the proposed antenna. The antenna radiator can be excited with the feeding line. It is very important to set the proper dimensions of feed line in order to achieve the perfect impedance matching. Fig. 2 shows the different optimetric values of feedline length ranges from 16.2mm to 16.6mm. It is analyzed that the proposed antenna achieves the good matching performance at 16.6mm value.

Moreover, the dimensions of patch also influences over the impedance matching of the proposed antenna. Fig. 3 demonstrates the variation of patch width from 9.1mm-9.5mm. It is analyzed that the optimized values for length and width of patch is achieved at 9.5mm.

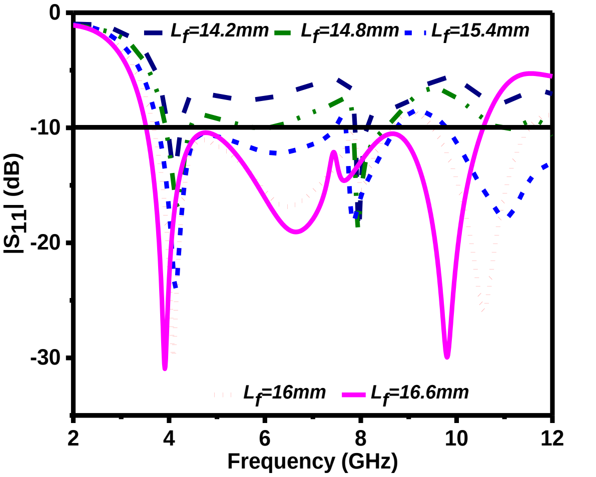


Fig. 2. Impedance Matching Analysis Related to Feed Line Length.

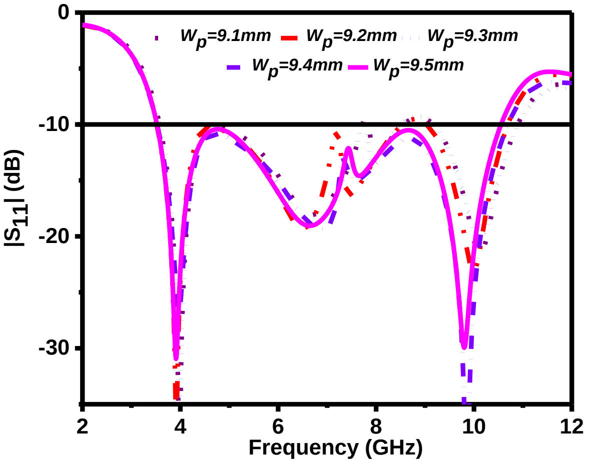


Fig. 3. Impedance Matching Analysis Related to width of Patch.

2) Variation of PGP with respect to length (L_{PGP}) and width (W_{PGP}): Length of the PGP plays a vital role to achieve broadband impedance BW. The PGP dimensions are almost half of the dimensions of the dielectric substrate. Fig. 4(a) demonstrates the optimum matched result of return loss S_{11}<10dB at 14.25mm.

Fig. 4(b) shows the optimized value of the width of PGP at 19mm. It is observed that the change in the dimensions of PGP results in the wide impedance BW and proper impedance matching**.** Finally, the simulation results of radiating patch, feed line and PGP shows the optimized antenna design geometry covers the different wireless communication application.

B. Return Loss (S_{11})

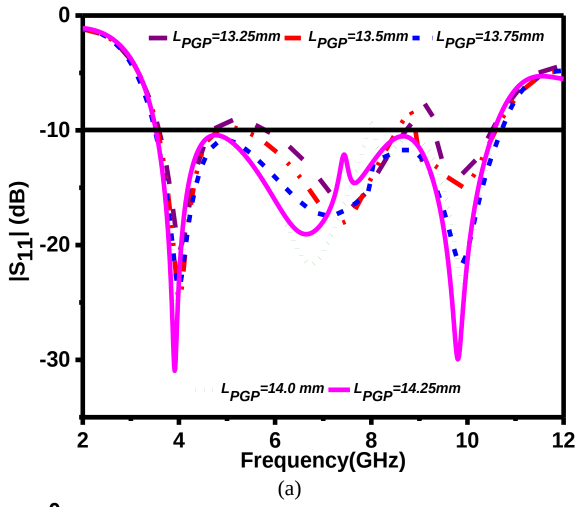
Fig. 5 delineates the return loss of prototype-I (reference antenna) and prototype-II (proposed antenna) across the operable frequency range. It is analyzed that the reference antenna achieves the broadband impedance BW of 6GHz at

10dB return loss ranging from 3.68GHz to 9.75GHz, which constitutes the fractional impedance BW of 89.3%.The reference antenna has the three resonances at 4.2GHz, 6.5GHz and 8GHz respectively. Moreover, proposed antenna achieves the 7GHz broadband BW varies from 3.4GHz to 10.5GHz which corresponds to fractional impedance BW of 100% at 10dB return loss. Furthermore, it is analyzed that proposed antenna has the three resonances at 3.9GHz, 6.6GHz and 9.8GHz, respectively.

From the above analyzed results of the return loss, it is concluded that proposed antenna achieved almost 10.7% improvement in impedance BW as compared to reference antenna prototype.

C. Peak Realized Gain (dBi)

Fig. 6 shows the peak realized gain of reference antenna and the proposed antenna.It can be observed at 11.6GHz frequency reference antenna exhibits peak realized gain of 5dBi. Besides, the multiple resonances such as: at 4.2GHz the antenna exhibits the gain of 4.1dBi, at 6.5GHz the acceptable gain of 3.45dBi and at 8GHz the gain of 3.6dBi is achieved.



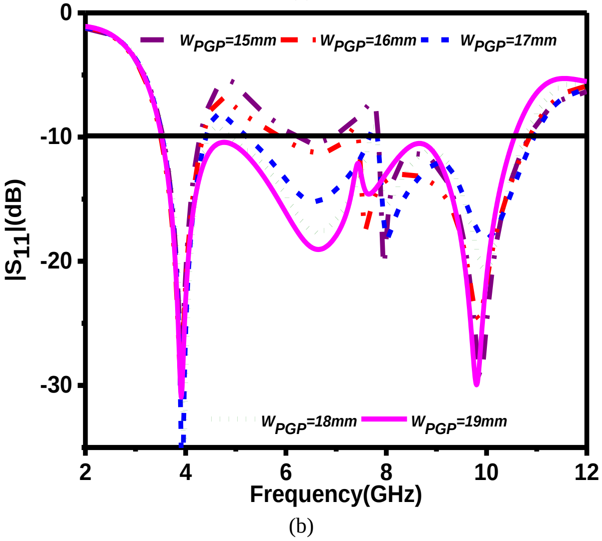


Fig. 4. (a) Impedance Matching Analysis Related to Length of PGP. (b) Impedance Matching Analysis Related to width of PGP.

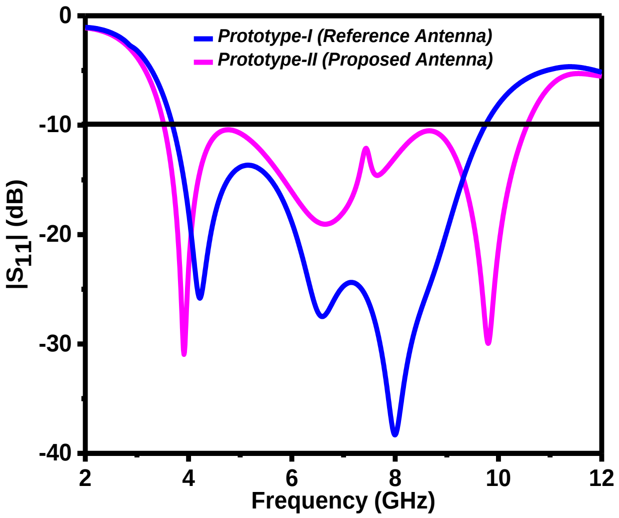


Fig. 5. S_{11} vs. Frequency Plot of Proposed and Reference Antenna Design.

Moreover, it is analyzed that the proposed antenna has achieved the high gain of 5.2dBi at 10.9GHz. However, the multiple resonances such as: at 3.9GHz, the acceptable gain of 3.7dBi, at 6.6GHz, the gain of 3.2dBi and at 9.8GHz the good gain of 4.9dBi is observed. Besides, it can also be seen that from 7.2GHz to 7.6GHz the gain is degraded upto 0.18dBi. The degradation in the gain is observed because of the L-shaped slots etched behind the feedline.

D. Surface Current Distribution (Jsurf)

The surface current distribution validates the effectiveness of proposed antenna. Therefore, it is essential to analyze a flow of current across surface of the antenna. The reference antenna current distribution simulation results are displayed in Fig. 7(a)-(c). The strong flow of current across the radiating patch, ground plane and feedline at 4.2GHz, 6.5GHz and 8GHz resonant frequencies is observed. Moreover, the proposed antenna design has almost the same current flowing effects as the reference antenna regardless of L-shape slots engraved on the conducting PGP as shown in Fig. 7(d)-(f).

E. Radiation Pattern

The 2-D Far-field radiation pattern along standard planes, i.e. E-plane (θ=0°) and H-Plane (θ=90°) of reference & proposed antenna is delineated in the Fig. 8. The reference antenna radiation pattern at multiple resonances is depicted in Fig. 8(a)-(c). It is analyzed that reference antenna possesses an stable omni directional radiation at 4.9GHz frequency, near monopole like radiation pattern at 6.5GHz and 8GHz resonant frequencies. Moreover, the proposed antenna radiation pattern at multiple resonances is illustrated in Fig. 8(d)-(f). It is analyzed at 3.9GHz frequency the proposed antenna achieved the omni directional radiation pattern. However, at 6.6GHz and 9.9GHz the radiation pattern along H-plane is near omni directional. The variation in radiation pattern is observed due to the L-shape slots inserted on the PGP. It is concluded from the above discussion that the reference antenna and proposed antenna has consistent and symmetrical radiation pattern at lower resonances and an acceptable change has been observed at higher resonances.

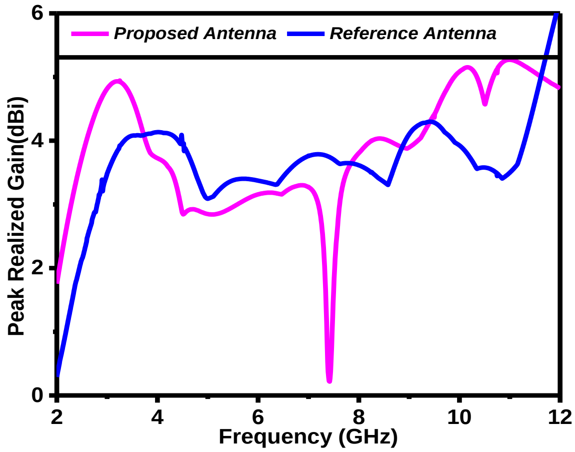
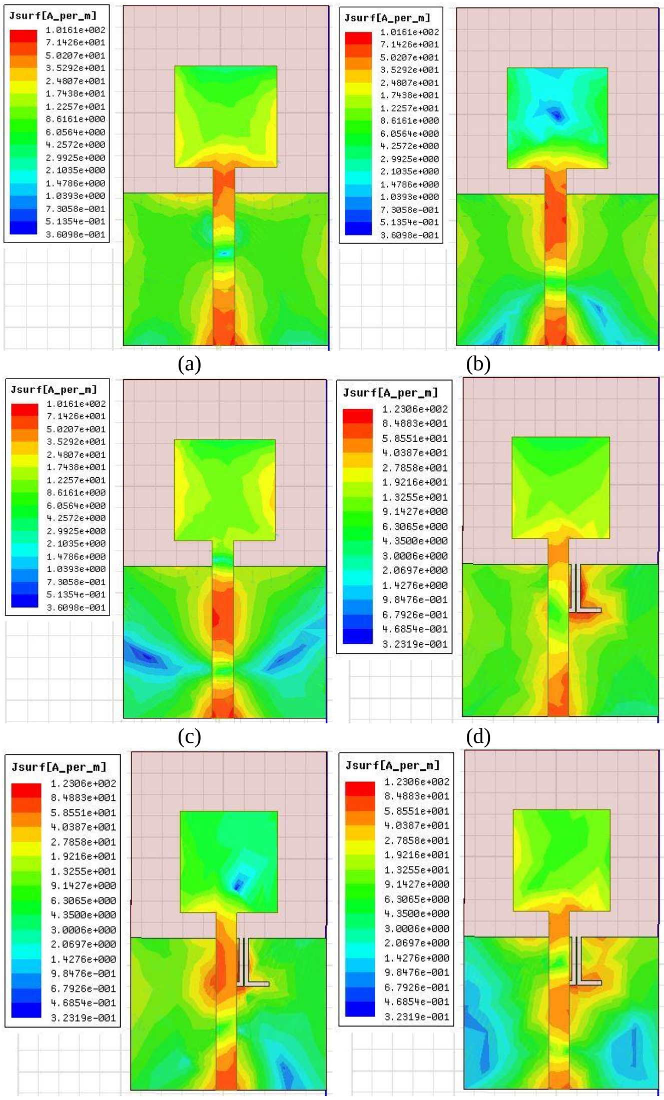


Fig. 6. Peak Realized Gain (dBi) vs. Frequency Range of Reference Antenna and Proposed Antenna.

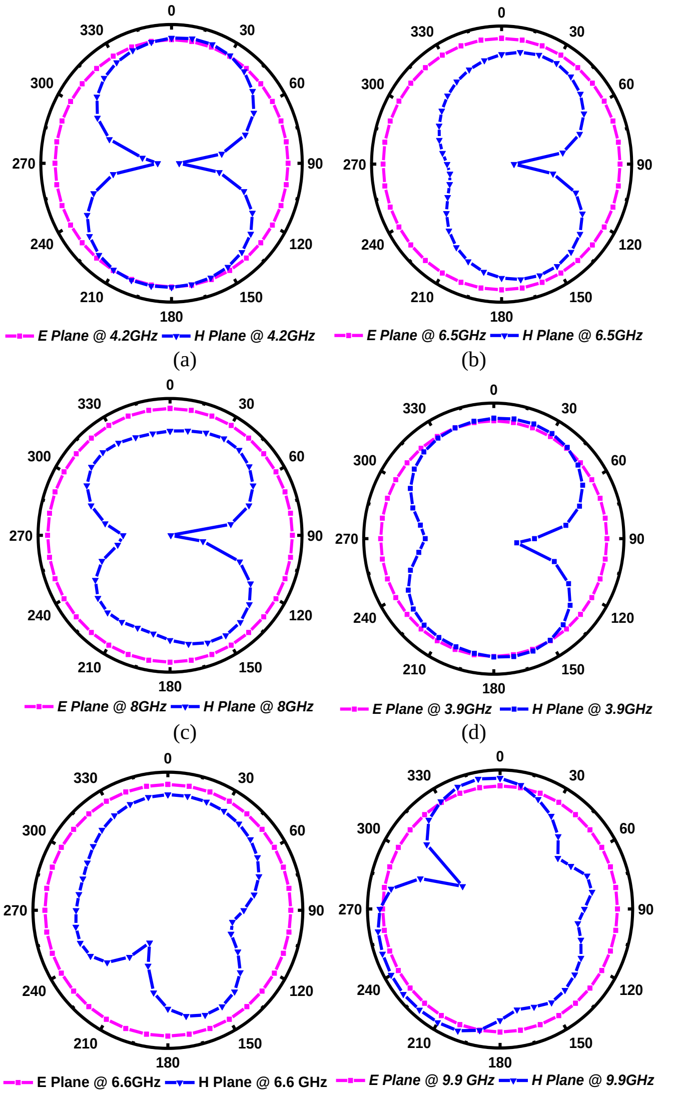


(a) (b)

(c) (d)

(e) (f)

Fig. 7. Surface Current Distribution (Jsurf) of the Reference Antenna and Proposed Antenna : (a) at 4.2GHz Resonance (b) 6.5GHz Resonance (c) 8GHz Resonance (d) at 3.9GHz Resonance (e) at 6.6GHz Resonance (f) at 9.8GHz Resonance.



(a) (b)

(c) (d)

(e) (f)

Fig. 8. Radiation Patterns at Azimuth and Elevation Plane of the Prototypes at Multiple Resonances:(a) Reference Antenna at θ=0° and θ=90°. (b) Reference Antenna at θ=0° and θ=90° (c) Reference Antenna at θ=0° and θ=90° (d) Proposed Antenna at θ=0° and θ=90° (e) Proposed Antenna at θ=0° and θ=90° (f) Proposed Antenna at θ=0° and θ=90°.

IV. Conclusion

In this paper two prototypes of compact broadband monopole antenna have been designed and simulated. The variables used in the designed antenna prototypes were calculated by the standard formulas. Besides, an impedance matching analysis by multiple times rigorous simulations has been carried out. The proposed antennas has been achieved the fractional impedance BW of 89.3% and 100% respectively. The designed antennas have the stable omni-directional radiation pattern at lower resonances. The peak realized gain of 5.2dBi has been observed at 10.9GHz resonant frequency. The strong visualization of current across the surface of antenna is observed at multiple resonances. The proposed antenna designs are well suitable candidate for WiMAX, UWB, land, naval and airborne radar applications.

V. Future Work

The work presented in this paper can be further extended to create the reconfigurable notch band functions at particular frequency band of spectrum. Moreover, a broadband antenna array topology will also be focused by implementing the efficient Wilkinson power divider and their performance will be analyzed in real time environment.

Acknowledgment

Authors Zaheer Ahmed Dayo, Qunsheng Cao, Yi Wang and Saeed Ur Rehman gratefully acknowledge the support from National Natural Science Foundation of China under grant No. 61871219.

References

[1] J. R. Verbiest and G. A. E. Vandenbosch, “A novel small-size printed tapered monopole antenna for UWB WBAN,” IEEE Antennas Wirel. Propag. Lett., vol. 5, no. 1, pp. 377–379, 2006.

[2] A. Dastranj and H. Abiri, “Bandwidth enhancement of printed E-shaped slot antennas fed by CPW and microstrip line,” IEEE Trans. Antennas Propag., vol. 58, no. 4, pp. 1402–1407, 2010.

[3] F. C. Commission, “Revision of Part 15 of the Commission’s Rules Regarding Ultra-Wideband Transmission Systems,” First Rep. Order …, no. FCC02-48, pp. 1–118, 2002.

[4] Y. Sung, “Triple band-notched UWB planar monopole antenna using a modified H-shaped resonator,” IEEE Trans. Antennas Propag., vol. 61, no. 2, pp. 953–957, 2013.

[5] M. Gupta and V. Mathur, “A new printed fractal right angled isosceles triangular monopole antenna for ultra-wideband applications,” Egypt. Informatics J., vol. 18, no. 1, pp. 39–43, 2017.

[6] M. Ojaroudi, N. Ojaroudi, and N. Ghadimi, “Dual band-notched small monopole antenna with novel W-shaped conductor backed-plane and novel T-shaped slot for UWB applications,” IET Microwaves, Antennas Propag., vol. 7, no. 1, pp. 8–14, 2013.

[7] M. Naser-Moghadasi, R. A. Sadeghzadeh, T. Sedghi, T. Aribi, and B. S. Virdee, “UWB CPW-fed fractal patch antenna with band-notched function employing folded T-shaped element,” IEEE Antennas Wirel. Propag. Lett., vol. 12, pp. 504–507, 2013.

[8] M. Koohestani and M. Golpour, “U-shaped microstrip patch antenna with novel parasitic tuning stubs for ultra wideband applications,” IET Microwaves, Antennas Propag., vol. 4, no. 7, p. 938, 2010.

[9] A. T. Mobashsher, M. T. Islam, and N. Misran, “Wideband compact antenna with partially radiating coplanar ground plane,” Appl. Comput. Electromagn. Soc. Newsl., vol. 26, no. 1, pp. 73–81, 2011.

[10] M. C. Tang, T. Shi, and R. W. Ziolkowski, “Planar ultrawideband antennas with improved realized gain performance,” IEEE Trans. Antennas Propag., vol. 64, no. 1, pp. 61–69, 2016.

[11] D. Zhang and Y. Rahmat-Samii, “Top-cross-loop improving the performance of the UWB planar monopole antennas,” Microw. Opt. Technol. Lett., vol. 59, no. 10, pp. 2432–2440, 2017.

[12] Y. Sung, “Bandwidth enhancement of a microstrip line-fed printed wide-slot antenna with a parasitic center patch,” IEEE Trans. Antennas Propag., vol. 60, no. 4, pp. 1712–1716, 2012.

[13] A. Valizade, J. Nourinia, B. Mohammadi, and P. Rezaei, “New design of compact dual band-notch ultra-wideband bandpass filter based on coupled wave canceller inverted T-shaped stubs,” IET Microwaves, Antennas Propag., vol. 9, no. 1, pp. 64–72, 2014.

[14] J. Li, Q. Chu, Z. Li, and X. Xia, “Compact Dual Band-Notched UWB MIMO Antenna With High Isolation,” EEE Trans. Antennas Propag., vol. 61, no. 9, pp. 4759–4766, 2013.

[15] A. Mousazadeh, M. Naser-Moghaddasi, F. Geran, S. Mohammadi, and P. Zibadoost, “Broadband CPW-Fed circularly polarized square slot antenna with arc-shaped and inverted-L grounded strips,” Appl. Comput. Electromagn. Soc. J., vol. 28, no. 4, pp. 314–320, 2013.

[16] A. Katuru and S. Alapati, “Design and analysis of modified circular patch antenna with DGS for UWB applications,” in Lecture Notes in Electrical Engineering, 2018, vol. 434, pp. 537–545.

[17] A. Kamalaveni and M. Ganesh Madhan, “Halve dumbbell shaped DGS tapered ring antenna for dual-band notch characteristics,” Electromagnetics, vol. 38, no. 3, pp. 189–199, 2018.

[18] K. F. Jacob, M. N. Suma, R. K. Raj, M. Joseph, and P. Mohanan, “Planar branched monopole antenna for UWB applications,” Microw. Opt. Technol. Lett., vol. 49, no. 1, pp. 45–47, 2007.

[19] Y. Z. Cai, H. C. Yang, and L. Y. Cai, “Wideband monopole antenna with three band-notched characteristics,” IEEE Antennas Wirel. Propag. Lett., vol. 13, pp. 607–610, 2014.

[20] S. Koziel and A. Bekasiewicz, Multi-Objective Design of Antennas Using Surrogate Models. 2016.

[21] S. K. Palaniswamy, K. Malathi, and A. K. Shrivastav, “Palm tree structured wide band monopole antenna,” Int. J. Microw. Wirel. Technol., vol. 8, no. 7, pp. 1077–1084, 2016.

[22] K. Mondal, A. Shaw, and P. P. Sarkar, “Inverted question mark broadband high gain microstrip patch antenna for ISM band 5.8 GHz/WLAN/WIFI/X-band applications,” Microw. Opt. Technol. Lett., vol. 59, no. 4, pp. 866–869, 2017.

[23] M. L. Meena, M. Kumar, G. Parmar, and R. S. Meena, “Design analysis and modeling of directional UWB antenna with elliptical slotted ground structure for applications in C- & X-bands,” Prog. Electromagn. Res. C, vol. 63, no. April, pp. 193–207, 2016.

[24] P. V. Naidu and A. Malhotra, “A small ACS-fed tri-band antenna employing C and L shaped radiating branches for LTE/WLAN/WiMAX/ITU wireless communication applications,” Analog Integr. Circuits Signal Process., vol. 85, no. 3, pp. 489–496, 2015.

[25] Z. A. Dayo, Q. Cao, P. Soothar, M. M. Lodro, and Y. Li, “A compact coplanar waveguide feed bow-tie slot antenna for WIMAX, C and X band applications,” in 2019 IEEE International Conference on Computational Electromagnetics (ICCEM), 2019, vol. 26, no. 3, pp. 1–3.

[26] L. Guo, M. Min, W. Che, and W. Yang, “A Novel Miniaturized Planar Ultra-Wideband Antenna,” IEEE Access, vol. 7, pp. 2769–2773, 2019.

[27] P. Soothar, H. Wang, B. Muneer, Z. A. Dayo, and B. S. Chowdhry, “A Broadband High Gain Tapered Slot Antenna for Underwater Communication in Microwave Band,” Wirel. Pers. Commun., no. 123456789, 2019.

[28] C. A. Balanis, Antennas Third Edition, vol. 45, no. 3. 2005.

[29] D. M Pozar, Microwave Engineering, 3rd Edition. 2004.

Crowd-Generated Data Mining for Continuous Requirements Elicitation

Ayed Alwadain[1]
Computer Science Department. King Saud University
Riyadh, Saudi Arabia

Mishari Alshargi[2]
Information Systems Department. King Saud University
Master Degree Student

***Abstract*—In software development projects, the process of requirements engineering (RE) is one in which requirements are elicited, analyzed, documented, and managed. Requirements are traditionally collected using manual approaches, including interviews, surveys, and workshops. Employing traditional RE methods to engage a large base of users has always been a challenge, especially when the process involves users beyond the organization's reach. Furthermore, emerging software paradigms, such as mobile computing, social networks, and cloud computing, require better automated or semi-automated approaches for requirements elicitation because of the growth in systems users, the accessibility to crowd-generated data, and the rapid change of users' requirements. This research proposes a methodology to capture and analyze crowd-generated data (e.g., user feedback and comments) to find potential requirements for a software system in use. It semi-automates some requirements-elicitation tasks using data retrieval and natural language processing (NLP) techniques to extract potential requirements. It supports requirements engineers' efforts to gather potential requirements from crowd-generated data on social networks (e.g., Twitter). It is an assistive approach that taps into unused knowledge and experiences emphasizing continuous requirements elicitation during systems use.**

Keywords—***Requirements engineering; RE; crowd data mining; NLP; Twitter; continuous requirements elicitation***

I. INTRODUCTION

Requirements engineering (RE) is the process of collecting, defining, documenting, and maintaining the requirements of a software system [1]. It is fundamental during the software development cycle to obtain users' needs by utilizing effective means of requirements elicitation, analysis, and management [2]. Getting the requirements right is important because mistakes cascade to subsequent development stages. Owing to poor RE practices, deficiencies at this phase cost more later and often result in systems failure [2-4].

Traditionally, elicitation is done at the beginning of software development. Recent approaches have advocated continuous requirement elicitation to capture user feedback and experiences during system's use [5]. Elicitation is needed during system's use to understand new feature requests, issues, and emerging requirements [6]. Requirements elicitation for traditional software systems has been well-studied, but new computing paradigms (e.g., social media, mobile apps, and cloud computing) require different assumptions and approaches [5]. These new computing paradigms enable users to express their feedback and experiences online via social-network sites, forums, and blogs.

Because of changing contexts and user needs, continuous requirements elicitation should be adopted to ensure that requirements stay refreshed and that needs are addressed [7]. Stakeholder needs and technologies change over time, exacerbated by the rise of crowd-generated data. Automated requirements elicitation methods and analysis should be incorporated to enable requirements engineers to acquire and analyze online data efficiently. Automation facilitates access to online crowd-generated data and the use of these data for systems' improvements [7]. Automated or semi-automated requirements elicitation approaches should be able to overcome issues facing existing traditional approaches [8]. This research proposes a methodology that collects crowd-generated data from social networks (e.g., Twitter) and processes the data using natural language processing (NLP) techniques to extract potential emerging requirements for a certain software product.

The rest of the paper is organized as follows. Section 2 presents the literature review while Section 3 details the proposed methodology and its supporting tool. Sections 4 and 5 respectively present the discussion and the conclusion of this study.

II. LITERATURE REVIEW

The success of a system development or an upgrade depends on a well-developed RE process that successfully elicits and manages stakeholder requirements, resulting in a higher level of satisfaction [4, 9]. Requirements elicitation is traditionally the first phase of obtaining requirements. Elicitation is the most important phase because the collection of poor requirements can lead to project failure [10-12]. The involvement of users and customers in the RE process leads to many benefits, such as improved system acceptance, more accurate and complete requirements, and improved project success rates [13]. Many issues lead to poorly collected requirements, such as ambiguous project scopes, poor system understanding, and volatility where the evolved users' needs do not meet the original requirements [14].

Various requirements elicitation approaches have been suggested [15]. Most existing techniques are manual and assume the presence of the stakeholders involved. Employing such techniques can rapidly become expensive and resource-intensive, particularly when dealing with larger stakeholder populations [16-18]. Employing such techniques to engage a large user base has always been a challenge, especially when there are large numbers of software users beyond the organization's reach [7]. Traditional RE approaches ignore opportunities to continuously engage large and heterogeneous groups of users who express their feedback on social networks and other websites. Better approaches are needed to tap crowd-generated data (e.g., feedback and opinions) to enable developers to consider them when developing their product's next version [7].

Stakeholder goals, environments, technology evolvement, and the emergence of new computing paradigms require continuous requirements elicitation. For example, social-network sites and mobile applications generate data that can be collected and analyzed for potential requirements [7]. The rise of social networks and mobile applications has enabled the collection of massively generated crowd data. Social-network users can contribute their feedback directly or indirectly regarding system improvements [19, 20]. Whereas social networks were not designed for the purpose of requirements engineering, many companies include social networks in their software development process for this purpose [21].

Understanding public opinion and demands is a time-consuming process because of the high volume of crowd-generated data that must be reviewed [22]. Thus, automatic approaches to elicit and analyze such data are needed to achieve faster response times [7]. Automation facilitates the identification and analysis of potential requirements that are otherwise challenging and unreachable using traditional RE [8].

An emerging theme within RE research is Crowd-based requirements engineering (CrowdRE). It is an overarching term for the employment of automated or semi-automated methods to elicit and explore data from a crowd to derive potential requirements [7]. Crowdsourcing in requirements elicitation would enable the continuous requirements elicitation process during the life cycle of the software product. Such a practice would facilitate a deeper, wider, and more up-to-date perspective of how users perceive systems and to understand how requirements evolve [5]. Typically, a crowd is a large and heterogeneous group of existing or prospective users [7]. CrowdRE captures and analyzes user needs regarding the evolution of existing software systems, and it monitors software system usage and experiences. Crowd users report on a variety of aspects, such as problems, improvements, or extension ideas, which are useful for software development teams [7].

Crowd-generated textual data should be retrieved and processed with NLP techniques. NLP concerns the application of computational techniques for automatic parsing, analysis, and representation of human language. Many techniques have been suggested to process raw text in natural languages. For example, tokenization is a technique used for splitting a stream of text into its basic elements (i.e., tokens) such as words and phrases and other symbols [23]. Part-of-speech (POS) tagging is used to assign labels (e.g., noun, verb) to each identified token in a given text [24].

Several studies have attempted to automate the requirements elicitation process using NLP techniques. For example, NLP was used to extract early requirements matching predefined patterns from user manuals and project reports. The text in these documents was tokenized and POS tagging was used to annotate the text. Then, topic modeling was applied to group-requirement items of similar content to avoid information overload. Whereas it is considered appropriate to reduce the burden of gathering requirements from scratch, some limitations have been reported, such as unclear extracted requirements and lack of comprehensive patterns [25].

Furthermore, an approach was developed to automate some requirements elicitation tasks using a tool that gathered stakeholder input in a centralized repository. Then, it used extended markup language and extensible stylesheet language transformations to render specifications [26]. In [27], a method was suggested to extract requirements from textual data in documents. NLP techniques (e.g., tokenization, POS tagging, and clustering) were used. Another study examined similar project documentation to extract potential requirements using NLP techniques (e.g., POS tagging) [28]. Another approach was proposed that used online customer reviews to extract needs and preferences regarding a specific product [29].

III. Proposed Continuous Requirements-Elicitation Methodology

This section outlines the proposed methodology and its supporting tool. This research provides an approach to automatically collect crowd-generated data via Twitter and process it using NLP techniques to find requirements. The proposed methodology is shown in Fig. 1. It enables engineers to elicit data from Twitter and analyze it using NLP techniques to find potential requirements. Twitter was selected because it is a popular microblogging social-media network and a potential data source to extract requirements [30, 31].

The methodology has four main steps: tweets collection and filtering, applying POS tagging, requirements generation, and requirements clustering. The following subsections illustrate the proposed methodology steps and its instantiation using AutoReq.

A. Tweet Collection (Pattern Matching) and Filtering

AutoReq enables requirements engineers to input a search keyword (e.g., the name of an existing system) and search twitter feed. The tool uses the Twitter application program interface (API) to retrieve real-time tweets matching the search criteria (i.e., predefined pattern). For example, if we were interested in finding the feedback of an existing system, X, the patterns added to AutoReq would include "X should …," "X could …," and "X lacks …".

In this study, the software system of interest is Snapchat. It is a global multimedia messaging application. It was selected because it is widely used and has very diverse user groups with constantly evolving requirements. Prior to this experiment, we noticed users tweeting potential requirements, additional features, complaints, and other issues about Snapchat. An AutoReq pattern search list was used. Then, tweets were filtered from unwanted noise (e.g., hashtags, user mentions, and universal resource locators). They were then saved to the AutoReq database. During the active stream retrieval of tweets, more than 350 tweets having the word "Snapchat" were retrieved, and only 47 matched the predefined pattern.

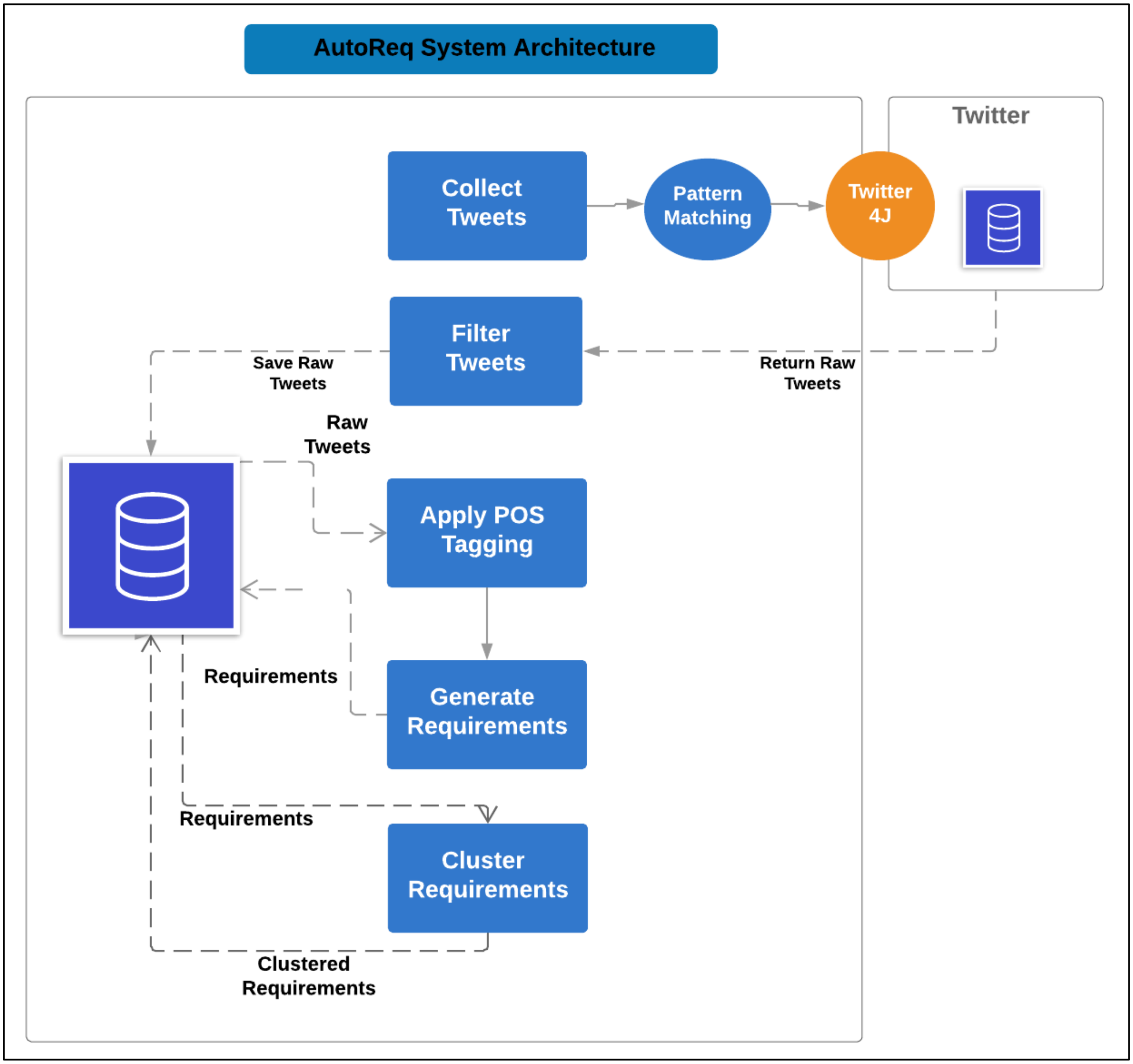


Fig. 1. Methodology and AutoReq System Architecture.

B. Part-of-Speech (POS) Tagging

Retrieved tweets were then tokenized by breaking them into tokens. Each tweet was then annotated using the Stanford POS tagger [24]. Tags were assigned to each word, depending on its role in the sentence. Still, there was some incorrect tagging. For example, the word “update” was incorrectly tagged in the tweet “snapchat should remove the last update.” It should have been tagged as a noun, but it was instead tagged as a verb. This phenomenon can lead to the generation of confusing requirements.

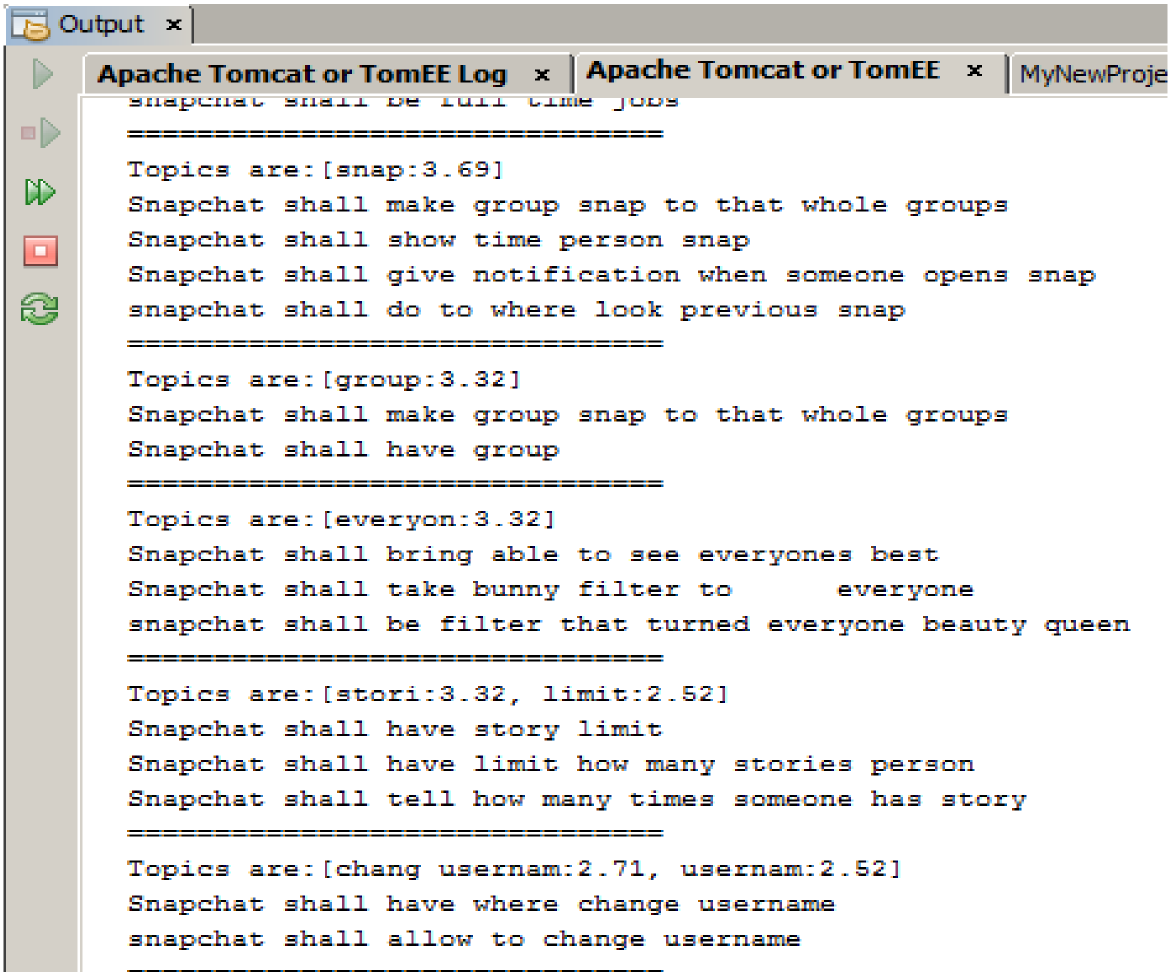
C. Requirements Generation

After tagging the words of each tweet, the first annotated verb and the closest three words were used to generate a requirement clause. Using a predefined requirement template within the tool, the requirement phrase was structured as “X shall + requirement clause.” In this experiment, generated requirements were structured as “Snapchat shall + requirement clause.” In some cases, a tweet contained more than one sentence. Thus, a recursion function of the tool was used to process the second part of the tweet. To find common conjunctions that potentially indicate the need for the use of the recursion capability, a qualitative analysis of the raw collected tweets was conducted. Then, the connection-words list was developed based on the qualitative analysis and the use of existing conjunction words in English [32]. Using recursion, tweets were split into parts using a conjunction word. Each part of the tweet was processed alone, and then both parts were combined as one requirement using the format “Snapchat shall + combined tweet output.”

D. Requirements Clustering

After requirements generation, clustering can be useful, particularly in cases where the retrieved tweets are large. Generated requirements were clustered to provide an aggregated perspective of common themes from the generated requirements. Clustering was conducted using the RxNLP sentence-clustering API [33]. It groups text tokens on a sentence level. It can be applied to short texts, or, in this research, tweets, to build logical and meaningful clusters with suggested topics for each cluster.

Generated requirements were clustered based on the most frequent topic themes, making it easier to find requirements of interest. The results, as shown in Fig. 2, contain the cluster topic, cluster score, and cluster tweets. The cluster topic is a suggested name of the cluster contents, whereas the cluster score describes the topic meaningfulness and cluster size. It facilitates cluster ranking and unwanted cluster pruning.



```
snapchat shall be full time jobs
===============================
Topics are:[snap:3.69]
Snapchat shall make group snap to that whole groups
Snapchat shall show time person snap
Snapchat shall give notification when someone opens snap
snapchat shall do to where look previous snap
===============================
Topics are:[group:3.32]
Snapchat shall make group snap to that whole groups
Snapchat shall have group
===============================
Topics are:[everyon:3.32]
Snapchat shall bring able to see everyones best
Snapchat shall take bunny filter to      everyone
snapchat shall be filter that turned everyone beauty queen
===============================
Topics are:[stori:3.32, limit:2.52]
Snapchat shall have story limit
Snapchat shall have limit how many stories person
Snapchat shall tell how many times someone has story
===============================
Topics are:[chang usernam:2.71, usernam:2.52]
Snapchat shall have where change username
snapchat shall allow to change username
===============================
```

Fig. 2. Requirements Clustering.

IV. Discussion

RE mostly uses traditional data sources (e.g., forms, reports, notes, workshops, and meetings) and manual approaches such as interviews for capturing stakeholder requirements. The wide use of social networks and mobile apps has contributed to a massive growth in online crowd-generated data. Crowd users report on a variety of issues based on software problems, desired improvements, and extension ideas, which are potentially useful for software development teams [7]. These data are often massive, unstructured, and manually inaccessible [22]. Hence, recent research has called for the development of automated approaches to capture and analyze these data to locate potential requirements. A rising opportunity for RE lies within the use of hidden and unused crowd-generated data [7].

This research endeavored to explore this research area and contributed as follows. First, this study is early research exploring the use of crowd-generated social-networks data to find new requirements for an existing software system. It proposed a methodology and a tool to capture and analyze crowd-generated data to identify potential requirements. Such an approach is needed to achieve fast responses to user needs and to explore the hidden, unused data generated by users on social networks [7]. The developed methodology and tool support requirements engineers in their tasks of monitoring and eliciting potential requirements from crowd-generated data using their reported feedback, comments, and experiences.

Second, this research used NLP techniques to automatically analyze the captured textual crowd-generated data (i.e., tweets). NLP techniques support requirements engineers by automating parsing, analysis, and representation of textual data. Manual inspection, filtering, and processing are time-consuming and resource-intensive. Thus, an automated approach of crowd-generated data retrieval and processing reduces time and resource utilization. Nonetheless, there were some issues with unclear generated requirements phrases from incorrect tagging when using a POS tagger. To overcome this, the developed tool was designed to show the original tweets and the generated associated requirement phrases to help requirements engineers trace and understand the generated requirements.

Third, this research emphasized the continuous requirements elicitation process over a software product life cycle using crowd-generated data [5]. Feedback and experiences of current or prospective users were continuously captured about new features, emerging needs, and other issues. This approach is not easily implementable with traditional data sources, such as manuals and reports.

Fourth, a sentence-clustering technique was used to cluster requirements based on their similarity [33]. In this research, every processed tweet was treated as a unique requirement and a genuine idea that may lead to redundant requirements. Thus, a sentence-clustering technique was used to enable requirements engineers to look at clusters when the generated requirements are large. This reduces requirements engineers' manual efforts. Previous research mostly used topic modeling techniques to detect the most frequent words in their data source to build requirements [25].Some studies used clustering to cluster the requirements based on predefined centroids, regardless of similarity [27].

V. Conclusion

Requirements elicitation is a crucial phase in the software development life cycle designed to fully understand users' needs. During the elicitation process, interviews, workshops, reports, and manuals are typically used to generate requirements. However, emerging computing paradigms and the massive growth of crowd-generated data require automated elicitation approaches. Crowds directly or indirectly express their feedback, comments, and opinions regarding an existing system on social networks and similar platforms. Gathering data using existing requirements elicitation techniques is an arduous process, particularly when dealing with large-scale systems.

This research proposed a methodology and proof-of-concept to automate the retrieval and analysis of crowd-generated data from Twitter using NLP techniques to find potential requirements of an existing software product. This is an early study investigating the use of crowd-generated data to find potential requirements. It employs NLP techniques to automatically analyze captured textual data, and it enables a continuous requirements elicitation process during the use of software products. It also uses a clustering-sentence technique to cluster requirements based on their similarity to automate grouping of similar tweets. This reduces manual RE efforts.

Because every research effort is limited, there are some limitations with this study. First, because we proposed a semi-automated tool, there needs to be an RE verification and evaluation of the generated requirements to assess their relevance and importance. Second, this study inherited some limitations of the applied NLP techniques, particularly POS tagging. In addition, automated text processing and analysis have their own limitations. For example, there were some generated requirements that were not meaningful because of either incorrect tagging or retrieval. Another limitation was inherited from the data source, owing restricted access to tweets using the Twitter API and the 140-character limitations at the time of the execution of the experiment.

In the future, extra efforts are needed to improve the suggested approach. For example, additional NLP techniques (e.g., collaborative filtering) should be included to extract relevant requirements. Furthermore, richer data sources are suggested, including Facebook and online app reviews, to collect richer requirements. In general, further research is needed to develop methods and tools that facilitate continuous requirements elicitation to retrieve and analyze online crowd-generated data during software systems use.

Acknowledgment

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this research.

References

[1] S. Gupta and M. Wadhwa, "Requirement Engineering: An Overview," International Journal of Research and Engineering, vol. 1, pp. 155-160, 2013.

[2] J. Vijayan and G. Raju, "A New Approach to Requirements Elicitation Using Paper Prototype," International Journal of Advanced Science and Technology, vol. 28, pp. 9-16, 2011.

[3] P. Rajagopal, R. Lee, T. Ahlswede, C. Chia-Chu, and D. Karolak, "A new approach for software requirements elicitation," in Proceedings of the 6th IEEE International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005.

[4] D. Pandey and V. Pandey, "Requirement Engineering: An Approach to Quality Software Development," Journal of Global Research in Computer Science, vol. 3, pp. 31-33, 2012.

[5] M. Hosseini, K. Phalp, J. Taylor, and R. Ali, "Towards Crowdsourcing for Requirements Engineering," in The 20th International Working Conference on Requirements Engineering: Foundation for Software Quality, 2014.

[6] J. A. Khan, L. Liu, L. Wen, and R. Ali, "Crowd Intelligence in Requirements Engineering: Current Status and Future Directions," in Requirements Engineering: Foundation for Software Quality, 2019, pp. 245-261.

[7] E. C. Groen, N. Seyff, R. Ali, F. Dalpiaz, J. Doerr, E. Guzman, et al., "The Crowd in Requirements Engineering: The Landscape and Challenges," IEEE Software, vol. 34, pp. 44-52, 2017.

[8] N. Mulla and S. Girase, "A New Approach to Requirement Elicitation Based on Stakeholder Recommendation and Collaborative Filtering," International Journal of Software Engineering & Applications (IJSEA), vol. 3, pp. 51-60, 2012.

[9] B. Nuseibeh and S. Easterbrook, "Requirements engineering: a roadmap," in Proceedings of the Conference on the Future of Software Engineering, 2000, pp. 35-46.

[10] S. Khan, A. B. Dulloo, and M. Verma, "Systematic Review of Requirement Elicitation Techniques," International Journal of Information and Computation Technology, vol. 4, pp. 133-138, 2014.

[11] O. I. A. Mrayat, N. M. Norwawi, and N. Basir, "Requirements Elicitation Techniques: Comparative Study," International Journal of Recent Development in Engineering and Technology, vol. 1, pp. 1-10, 2013.

[12] S. Sharma and S. Pandey, "Revisiting Requirements Elicitation Techniques," International Journal of Computer Applications, vol. 75, pp. 35-39, 2013.

[13] R. Snijders, Ö. Atilla, F. Dalpiaz, and S. Brinkkemper, "Crowd-centric requirements engineering: A method based on crowdsourcing and gamification," Master's Thesis, Utrecht University, 2015.

[14] M. G. Christel and K. C. Kang, "Issues in requirements elicitation," Technical Report CMU/SEI-92-TR-012., Software Eng. Inst., Carnegie Mellon University, 1992.

[15] M. S. Tabbassum Iqbal, "Requirement Elicitation Technique: - A Review Paper," International Journal of Computer & Mathematical Sciences, vol. 3, pp. 1-6, 2014.

[16] M. Yousuf, M. Asger, and M. U. Bokhari, "A Systematic Approach for Requirements Elicitation Techniques Selection: A Review," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, pp. 1399-1403, 2015.

[17] K. Wnuk, "Understanding and supporting large-scale requirements management," Licentiate Thesis, Department of Computer Science, Lund University, vol. 2010, 2010.

[18] U. Sajjad and M. Q. Hanif, "Issues and challenges of requirement elicitation in large web projects," School of Computing, Blekinge Institute of Technology, Ronneby, Sweden, 2010.

[19] D. T. Nguyen, N. P. Nguyen, and M. T. Thai, "Sources of misinformation in Online Social Networks: Who to suspect?," presented at the Military Communications Conference, 2012.

[20] D. S. Kim and J. W. Kim, "Public Opinion Sensing and Trend Analysis on Social Media: A Study on Nuclear Power on Twitter," International Journal of Multimedia and Ubiquitous Engineering, vol. 9, pp. 373-384, 2014.

[21] N. Seyff, I. Todoran, K. Caluser, L. Singer, and M. Glinz, "Using Popular Social Network Sites to Support Requirements Elicitation, Prioritization and Negotiation," Journal of Internet Services and Applications, vol. 6, pp. 1-16, 2015.

[22] M. Yousuf and M. Asger, "Comparison of Various Requirements Elicitation Techniques," International Journal of Computer Applications, vol. 116, pp. 8-15, 2015.

[23] T. Verma and D. G. Renu, "Tokenization and Filtering Process in RapidMiner," International Journal of Applied Information Systems (IJAIS)–ISSN, vol.7, pp. 2249-0868, 2014.

[24] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003.

[25] Y. Li, E. Guzman, K. Tsiamoura, F. Schneider, and B. Bruegge, "Automated Requirements Extraction for Scientific Software," Procedia Computer Science, vol. 51, pp. 582-591, 2015.

[26] N. W. Kassel and B. A. Malloy, "An approach to automate requirements elicitation and specification," in International Conference Software Engineering and Applications, 2003.

[27] S. Murugesh and A. Jaya, "A Generic Framework for Requirements Elicitation from Informal Descriptions," International Journal of Advanced Research in Computer Engineering & Technology, vol. 3, pp. 2545-2549, 2014.

[28] K. Li, R. Dewar, and R. Pooley, "Requirements capture in natural language problem statements," Heriot-Watt Technical Report HW-MACS-TR-0023, 2004.

[29] R. Rai, "Identifying key product attributes and their importance levels from online customer reviews," in ASME International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE2011), Paper No. DETC2012-70493, 2012.

[30] S. Arapostathis and S. Kalogirou, "Twitter data as a volunteered geographic information source: review paper of recent research analysis methods and applications," in Proceedings of the 1st Spatial Analysis Conference, 2013.

[31] Y.-k. Lee, N.-H. Kim, D. Kim, D.-h. Lee, and H. P. In, "Customer Requirements Elicitation Based on Social Network Service," KSII Transactions on Internet and Information Systems (TIIS), vol. 5, pp. 1733-1750, 2011.

[32] Smart Words. (20 Feb). Conjunctions. Available: https://www.smart-words.org/linking-words/conjunctions.html

[33] RxNLP. Sentence Clustering API. Available: http://www.rxnlp.com/api-reference/cluster-sentences-api-reference/

Augmented Reality App for Teaching OOP

Sana Rizwan[1], Arslan Aslam[2], Sobia Usman[3], Muhammad Moeez Ghauri[4]
Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Pakistan

***Abstract*—Now a days, there is demanding needs of developing interactive mediums of study. As our conventional methods of learning are not very effective. Programming has become one of the core subjects of every field of study due to their vast use. However, introducing computer programming to those students who's are not familiar with programming is a tough task. Use interactive learning through visual effects using AR (Augmented Reality) developed to provide a platform for new students to interact more in learning environment. As this learning environment becomes more effective it is easier for new comers to understand key concepts of programming more effective way.**

Keywords—Augmented reality; object-oriented programming; unity; visualization; human computer interaction; Vuforia; rendering; compiler

I. Introduction

Application will allow the people to visualize the OOP concepts and check their interactions and effects to understand them easily with efficiency rather than just sitting and thinking about them by using and visually seeing them their rate of progress will increase.

Although there is lot of e-learning application available on online platforms but there is no special application for object-oriented programming. This project consists of a web application and an android application. Web portion is consisting of textual based tutorials for learning and android application is for visual learning.

II. Literature Review

The importance of learning through computer starts from 21th century and their importance gradually increase with the passage of time. Programming is one of basic course required in any computer science related field. For most of the students, it is also one of the most challenging tasks that how to understand coding or how to write a computer program. For improving the novice experience with learning to program, visual environments and effects can be designed. At university level as well as the industry have placed increasing importance on the early experience of students to object-oriented programming (OOP).

There is a need to develop an interactive e-learning environment for students to overcome the challenge of learn programming. Incredible developments had been occurring in computer technology and its availability in that time duration. Computer technology used for learning in schools, universities, business market, industries and the many other professions in last ten years. So, the number of learning software's and applications has growth dramatically.

Traditional methods of study are more complex, boring and time taking. The field Human computer interaction directs us to how to make an interactive application? What are the requirements of user? What are the problems faced by users? What the user needed? There are some questions invoke at every developer while he/she build an application of software. HCI gives the better ways to develop an interactive application. Students' who learn in an e-learning environment would be independent of distance, time, computing platform as well as classroom size also.

The knowledge needed to write a computer program is an important part of literacy in modern society. While private coding classes are expensive and limited.

In programming some of the most difficult things to understand are concepts that we use to develop data structures and algorithms. Sometimes even for the more intelligent people they cannot understand some concepts due to the fact that each person has a different mind-set.

This application will allow such people to visualize such concepts and check their interactions and affects to understand them easily with efficiency rather than just sitting and thinking about them by using and visually seeing them their rate of progress will increase.

Although there is lots of e-learning application available on online platforms but there is no special application for object-oriented programming. This project consists of a web application and an android application. Web portion is consisting of textual based tutorials for learning and android application is for visual learning.

Originally this approach of learning started in the more developed countries of the world such as U.S.A and U.K in these countries to improve the education system e-learning was introduced now though with the advent of augmented and virtual reality, this system of e-learning is being taken a step further by making learning more interactive and hence more interesting by using such technologies. In such developed countries such application has started being widely used in young children [3], hence our aim is the same to increase education efficiency by using these technologies.

Moreover, The School Education Department of Punjab has been working closely with Punjab Information Technology Board (PITB) to develop and implement E-learning solutions for secondary and higher secondary students in which, PCTB textbooks for Grades 6 through 10 have been digitised and augmented with thousands of interactive multimedia components like Animations, Simulations, 3D models and videos. [2]

More related works would be those that follow the four steps [4], to teach students through e-learning, due to the fact that much of the increase in understanding comes from use of what is known as "visual literacy." Visual literacy is defined as the ability to understand data and information presented in a pictorial format or graphic images [1], Visual literacy has been proven to enhance learning especially in subjects where they do not have much prior knowledge. [5].

In fact, studies show that using graphics in instructional modules promotes greater understanding in general. It is in fact proven that less is more beneficial if it can be better understood [8], Logic building process is also a tough task for students and many students confused about different approaches to solve some programming problem. These aspects would be challenging for teachers and students [6], and the learning of the application of what is being learnt while it is being learnt is important [7], Researcher shows the object-oriented concept. Some concepts students understand properly but some concept cannot understand. It finally points difficulties in some areas and result, planning to give an easy and effective way to teaches OOP [9], by participating in applied task, the comprehension of the subject becomes better [10].

It is known through research that the programming skills of first year students are not up to the required level as most of them have not been exposed to how to solve tasks and do programming [11] hence in the past there have been many application that have focused on teaching beginners how to code but in most of these applications the user only learns the syntax and does not have an editor to practise it side by side and while it is important to know how to build a solution [12] it is also better to at least once make the solution as well because it cements our understanding, there have also been many lone editors for the java language but they are complicated and not very good for beginners to work on, while in the system we are proposing we will try to combine both an explainer as well as an editor with visual representations to allow the user to quickly understand concepts and learn the language efficiently as in interactive setting the user unknowingly tries harder [13].

III. Interfaces

Augmented Reality followed basic human interaction usability principles in user interface creation. Consistent interface and design present across all modules. By default, augmented reality opens in full screen mode in landscape orientation on some specific devices those who have sensors to show augmented objects. Interface is compatible with cross devices. Based on human computer interaction methods, we design interfaces very user friendly.

Following are prime interfaces components.

A. Explainer

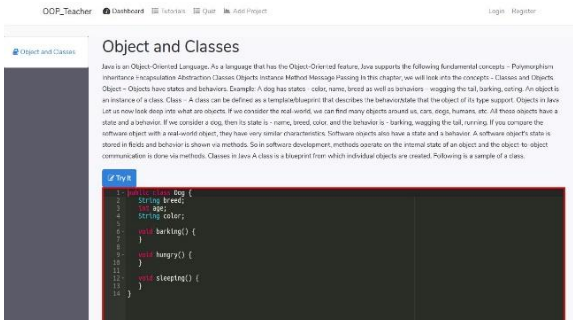
Explainer is mean interface component of project. This module consists of basic concepts tutorials of object-oriented programming. Every tutorial consists of three sections such as topic name, topic explanation and try yourself code as shown in Fig. 1. We provided the code examples of every topic which is easy to understand and make changes in practice code also available through compiler. User with be able to discuss or ask any question about the topic in comment section. Because it’s an interactive web application so tutorials are locked once someone take a tutorial about some topic the next task is to give a simple multiple-choice quiz. After successfully completed the quiz then the user moves towards next tutorial and so on.

B. Compiler

Compiler is basically for try yourself section that user pic example code of some specific topic and makes changes in it as shown in Fig. 2: Compiler able to work with multiple classes of object-oriented programming. As in case of any wrong code written by user Errors also shows with line number on console panel.

C. Quiz

Fig. 3 shows quiz section consists of multiple choice-based questions. Answers will be provided also for confirmation of selected choice. A progress bar will be saving the quiz data of users.

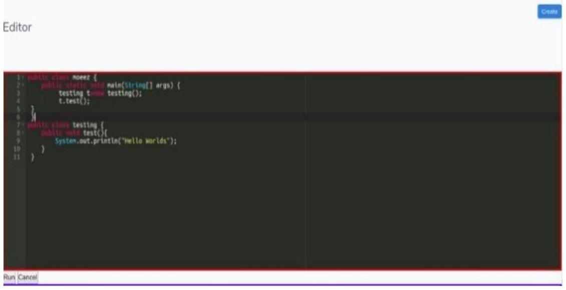


Object and Classes

Java is an Object-Oriented Language. As a language that has the Object-Oriented feature, Java supports the following fundamental concepts – Polymorphism Inheritance Encapsulation Abstraction Classes Objects Instance Method Message Passing In this chapter, we will look into the concepts - Classes and Objects. Object – Objects have states and behaviors. Example: A dog has states - color, name, breed as well as behaviors – wagging the tail, barking, eating. An object is an instance of a class. Class – A class can be defined as a template/blueprint that describes the behavior/state that the object of its type support. Objects in Java Let us now look deep into what are objects. If we consider the real-world, we can find many objects around us, cars, dogs, humans, etc. All these objects have a state and a behavior. If we consider a dog, then its state is - name, breed, color, and the behavior is - barking, wagging the tail, running. If you compare the software object with a real-world object, they have very similar characteristics. Software objects also have a state and a behavior. A software object's state is stored in fields and behavior is shown via methods. So in software development, methods operate on the internal state of an object and the object-to-object communication is done via methods. Classes in Java A class is a blueprint from which individual objects are created. Following is a sample of a class.

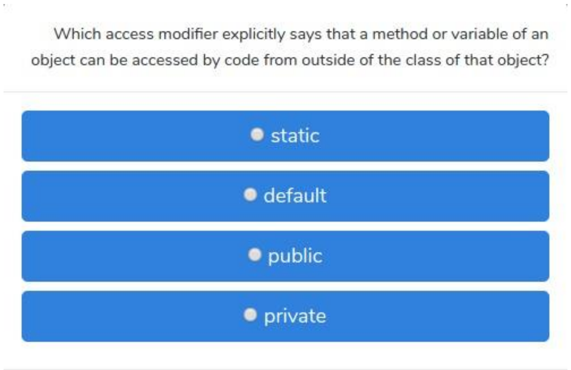
```
public class Dog {
    String breed;
    int age;
    String color;

    void barking() {
    }

    void hungry() {
    }

    void sleeping() {
    }
}
```

Fig. 1. Explainer



```
public class moeez {
    public static void main(String[] args) {
        testing t=new testing();
        t.test();
    }
}
public class testing {
    public void test(){
        System.out.println("Hello Worlds");
    }
}
```

Fig. 2. Compiler.



Which access modifier explicitly says that a method or variable of an object can be accessed by code from outside of the class of that object?

● static

● default

● public

● private

Fig. 3. Quiz.

D. *Rendering Screen*

Majority part of user interface cover by Rendering screen. Rendering screen helps us to seen Augmented reality (AR) components. One important thing that AR not supports with all the mobile devices, but some specific mobile devices shows AR object through camera to see elements present in surrounding.

IV. Methodology

Augmented Reality follows visual based interactive learning method to teach programming. The field of Human-Computer Interaction (HCI) or in other words show something by using visual effects is focused on enhancement the effectiveness as well as efficiency of human-computer interfaces through the development of both hardware and software designs to recognize human characteristics and behavior. Advancements in HCI technology can lead to enhanced Augmented reality (AR) experiences by providing more natural like environment and efficient methods for a user to interact with a real or virtual environment in an effective way.

A. *Software Architecture*

Fig. 4 shows software architecture. According to software architecture user interacts with web application or android application and moves towards explainer section for learning and editor section to explore learning concepts and tryout code examples with the help of compiler. All the learning data managed by an admin and only admin have authority to add, update or delete data in database. User might be suggesting some things in comments section as feedback.

B. *Compiler*

We make an API in java for makes an online compiler which code shows in Fig. 5 and integrate it to Laravel framework. We used library "Tools and Diagnostics" for multiclass compiler. This compiler saves all the classes into an array and separated with Java API basically is a JAR file which is run by using CMD.

C. *Augmented Reality (AR) apps with unity and Vuforia*

We can use AR to teach the concepts of OOP in more optimal manner. Augmented Reality can be used to teach OOP concepts. Inheritance can be taught using AR based mobile devices. Similarly, the concept of composition, aggregation, polymorphism etc. also taught by using AR app. We prepare the models of CAR which shows in Fig. 6, 7 and HUMAN which shows in Fig. 8. Firstly, we show that how to initialize an object in OOP then we move towards other major concepts of OOP. Make models of AR in unity 3D & Vuforia.

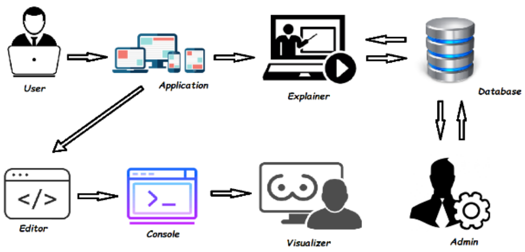
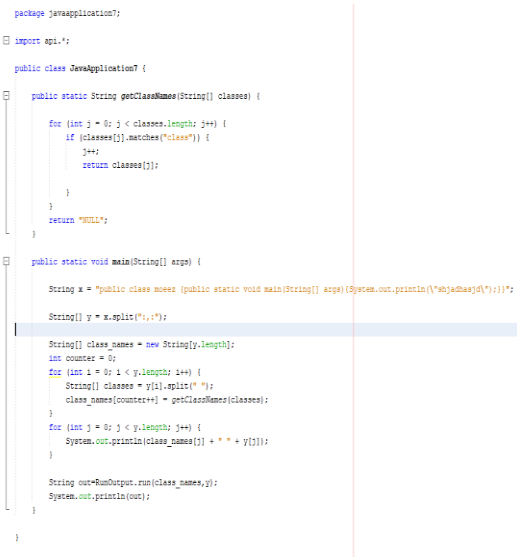
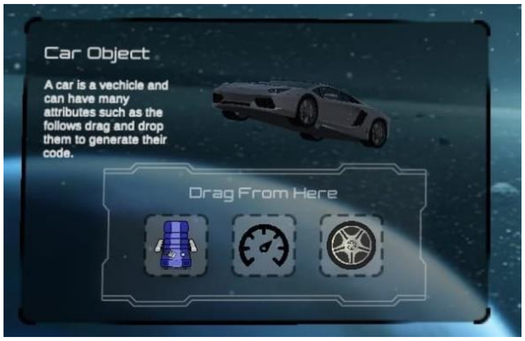


Fig. 4. Software Architecture.



```
package javaapplication7;

import api.*;

public class JavaApplication7 {

    public static String getClassNames(String[] classes) {

        for (int j = 0; j < classes.length; j++) {
            if (classes[j].matches("class")) {
                j++;
                return classes[j];

            }
        }
        return "NULL";
    }

    public static void main(String[] args) {

        String x = "public class moeez {public static void main(String[] args){System.out.println(\"shjadhasjd\");}}";

        String[] y = x.split(":,:");

        String[] class_names = new String[y.length];
        int counter = 0;
        for (int i = 0; i < y.length; i++) {
            String[] classes = y[i].split(" ");
            class_names[counter++] = getClassNames(classes);
        }
        for (int j = 0; j < y.length; j++) {
            System.out.println(class_names[j] + " " + y[j]);
        }

        String out=RunOutput.run(class_names,y);
        System.out.println(out);
    }

}
```

Fig. 5. Compiler.



Car Object

A car is a vechicle and can have many attributes such as the follows drag and drop them to generate their code.

Fig. 6. Augmented Reality App (Car Model (1)).

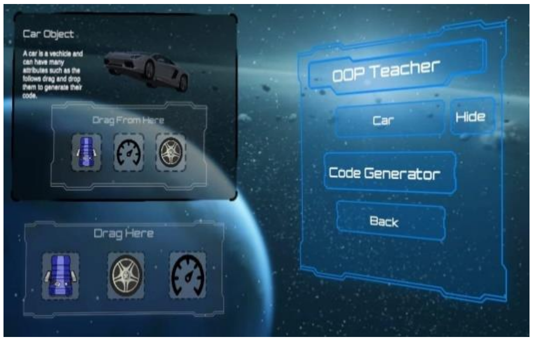


Fig. 7. Augmented Reality App (Car Model (2)).

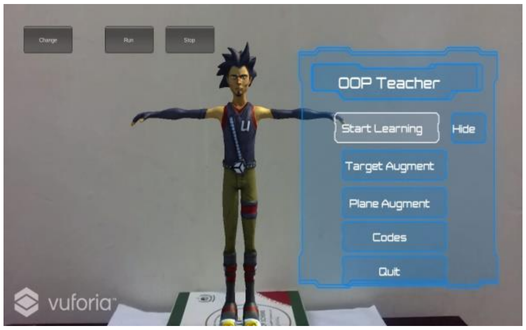


Fig. 8. Augmented Reality App (Human Model).

V. Results and Discussion

According to studies, it is observed that how augmented reality can be useful in teaching prospective such as for OOP concepts. It is more practical way to teach students in an artful environment. It is interesting and growing field. But lack of resources, awareness and devices that field not the part of studies even in modern institutes.

Augment reality gives us optimal solutions. By using methods and techniques of AR will make revolutionary changes in traditional learning process and enhance the modern learning techniques. The main purpose of this research is to find out the ways in which AR can be used in studies and we focus on OOP concepts and try out to develop a system which is more user friendly. We initially targeted small amount of population because everyone has not AR supported devices only few mobiles with specific sensors would be able to run an AR based application.

VI. Conclusion

Augmented reality is a new concept but can use generate optimal results. In the education field, it can be used in a very effective manner as it can be used to teach some concepts in a more practical way to students. Use of AR will make revolutionary changes in traditional teacher entered teaching process in future.

VII. Future Work

The project can be extended to theoretically all the programming languages as this is in essence a tool for e-learning using visual techniques, in the future, this application can be enhanced to follow “one size fits all” concept. Web Programming as well as Database Programming can also be included as in databases, we also require large amounts of visualizations and have to ponder upon the many what ifs. Machine learning techniques can be used to provide exercises to users based on their progress. This application can be made in game type theme in which user can progress only by completing a previous task to a satisfactory level. This will allow for more immersion and not let the user get bored. This application can be further extended to the professional level. This can also be really helpful for the beginners. More over the system can with some time include its own compiler that can visually show objects which will form basis for a simulator for testing running of the application.

References

[1] Suzanne Stokes. Visual Literacy in Teaching and Learning: A Literature Perspective. Electronic Journal for the Integration of Technology in Education, vol. 1, no. 1.

[2] https://elearn.punjab.gov.pk

[3] Tanvi Patel. Learning Object Oriented Programming Using Augmented Reality: A Case Study with Elementary School Students

[4] Hinterberger, H., 'E-Learnig: Make it as simple as possible, but not simpler'. Afr. Technol. Dev. Forum Journal, Vol. 4, Issue 2, July, 2007.

[5] Chanlin, L. (1997). The effects of verbal elaboration and visual elaboration on student learning.

[6] International Journal of Instructional Media, 24(4), 333-339. Retrieved December 26, 2001, from EBSCOhost database (Academic Search Elite).

[7] Heinich, R., Molenda, M., Russell, J. D., & Smaldino, S. E. (1999). Instructional media and technologies for learning (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

[8] Mayer, R. E., Bove, W., Bryman, A., Mars, R., & Tapangco, L. (1996). When less is more: Meaningful learning from visual and verbal summaries of science textbook lessons. Journal of Educational Psychology, 88(1), 64-73.

[9] Kelleher, C. and Pausch, R. Lowering the barriers to programming: A taxonomy of programming environments and languages for novice programmers. ACM Computing Surveys 37, 2 (June 2005), 83–137.

[10] Tony J. A Participative Approach to Teaching Programming, ACM SIGCSE, Volume 30 Issue 3, Sept. 1998, Pages 125-129 1998.

[11] M. McCracken, V. Almstrum, D. Diaz, M. G. anD. ianne Hagan, Y. B. Kolikant, C. Laxer, L. Thomas, I. Utting, and T. Wilusz. A multi-national, multi-institutional study of assessment of programming skills of first-year CS students. SIGCSE Bulletin, 33(4):125– 180, 2001.

[12] Elliot Soloway. Learning to Program = Learning to Construct Mechanisms and Explanation,1986.

[13] Malone, T. (1980). What makes things fun to learn? Heuristics for designing instructional computer games. Proceedings of the 3rd ACM SIGSMALL Symposium and the 1st SIGPC Symposium (pp. 162–169). Palo Alto, US.

A Readiness Evaluation of Applying e-Government in the Society: Shall Citizens begin to Use it?

Laith T. Khrais*[1], Mohammad Awni Mahmoud[3]
Department of Business Administration[1]
Department of MIS[3]
College of Applied Studies and Community Services
Imam Abdulrahman Bin Faisal University
Dammam, Saudi Arabia[1, 3]

Yara M. Abdelwahed[2]
Faculty of Commerce
Menoufia University, Menoufia, Egypt

***Abstract*—As people are in the era of the web, most of the society is using networks in their daily task, governments had found, it is crucial to build an electronic entity that was named e-government, to make transactions easier for citizens, and to make government nearer to society. The objective of this study is to assess the extent of e-government application on different countries, particular in Jordan; in addition, several experiences were displayed in this study. The examination was qualitative through interviewing governmental employees for extracting results based on their answers, focusing on the continuity of using e-government by users as a dependent variable. The conclusion was that the policies are trending to build their e-government entity, and to make it available for citizens to use. Further, this study recommends the government to concentrate on the path of building individuals' trust as well as using social influence to reinforce the idea of e-government service and evolve its usage.**

Keywords—e-Government; citizens; governmental transactions; Jordan

I. Introduction

The improvement of ICT is quick as of now significantly affect human life. In the line with this development, the worldwide versatile Internet clients' infiltration has achieved 4.68 billion by 2019 [1]. In Jordan, for instance, the penetration rate has come to 83 percent. This provoked an adjustment in procedures, capacities, and approaches in various areas of human life be found on ICT, including open segment administrations.

Furthermore, Jordan has a tremendous successfully mounting in telecommunication industry in the recent years. The Central Bank of Jordan was launched an electronic program so-called (E-fawateercom) since 2015 [2]. This program has some important benefits such as reduce the time, money and effort of paying bills compared to the traditional way.

Government aspects have also changed, including the departments by providing public service oriented to satisfy most people [3]. As a result, these Changes in the general society division are described by the improvement of electronic government or supposed e-government. In general, this study suggests dimensions of the most critical aspects towards the individual's choice to utilize e-government services in Jordan.

II. The Aspect of e-Government Structure

Governments are always searching to save costs in providing services chiefly along with the economic crisis existences. In this line, e-government possesses vital benefits to developing nations such as citizen empowerment by offering them with a variety of options online channels to improve the quality of service delivery [4]. A review of present literature exposes that Jordan has been heavily investing in enhancing its e-government facilities to the community.

E-government is planned the procedure of cooperation among government as well as society. According to [5], one critical aspect in the execution of e-government administrations is the acknowledgment and the readiness of society to utilize e-government administrations. From citizens' viewpoint, e-government allows persons to access public services to do their government transactions in a convenient way at anytime from anywhere.

E-government is an innovative form of advanced technology, with a series of set processes that the citizens' logs into the website by using a private username and password based on their selection and secret key with the end goal to do their various online transactions. An outline of the design of the e-government is illustrated in Fig. 1, which links to the essential beneficial components and their jobs inside the framework. Connection among citizens and specialist system for government frameworks is upheld by staggered exchanges.

Furthermore, the e-government framework bolsters correspondence with different servers, for example, Internet servers support the service administrations offering to clients. In turn, the (SWAN) provides secured as well as high-speed connectivity for Government operation between connecting State Headquarters and District Headquarters. The e-government framework bolsters correspondence with different servers, for example, Internet servers to support the service administrations offering to clients. In turn, the (SWAN) provides secured and high-speed connectivity for Government functioning between connecting State Headquarters and District Headquarters.

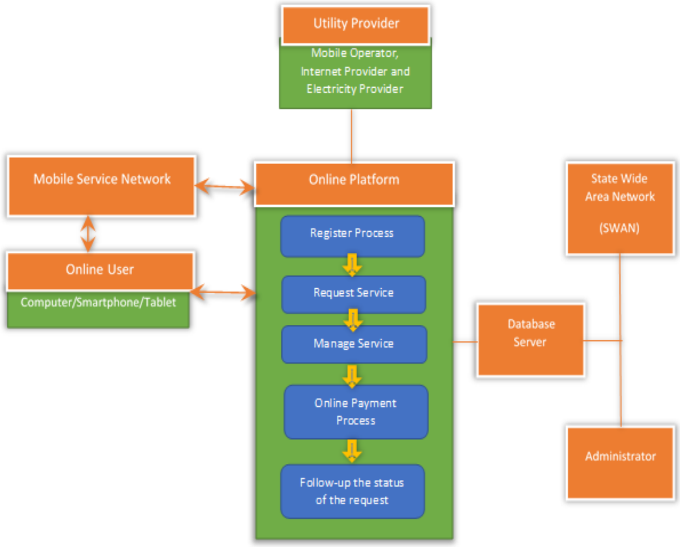


Fig. 1. e-Government Architecture.

In light of the data above, it very well may be seen that the network turns into a vital piece of the working of e-government, the cooperation between the national administration and the network can function admirably if there is investment from general society. In the event that there is lack of interest of the public in embracing e-government, it will never function and being vain presence.

One of the researcher's expresses that more consideration has been compensated to e-government administrations received from the point of view of the "supply" [6]. Then, simply little the researchers have been investigated for the civilian request's [7], and preparation [8]. The shortcoming of the investigations above does not possess a solid establishment to give a calculated model of the e-government benefit for a developing nation.

In [9] has portrayed e-government as the strategy for governments to use the many imaginative data innovation, for example, online web application. Such applications may outfit the citizens, inside associations with progressively accommodating obtainment to government, upgrade great administration, and offer more opportunities to law grounded establishments and methods, incorporates numerous issues, for instance, trust, security, insurance, transparency, care, and quality.

In addition, several initiatives of e-government in advancing nations failed because there is a wide gap between the e-government design as well as exciting condition in upcoming republics, like inadequate information, communication, and technology management and infrastructure, resulting in low acceptance of the online services by people as well as businesses.

The advancement of innovation is developing with time. This empowers a great deal of changes in procedures, capacities, and arrangements in different business exercises or in the general population segment. Changes that happen in the general population segment is set apart by the improvement of electronic-government.

There are four sorts of orders, first, to be specific Citizens, Government (G-to-C), and the second one, to be specific Government to Business (G-to-B), Government to Government (G-to-G), and the last one in particular Government to Employees (G-to-E) [3,10]. G-to-C is an e-government application that most popularly known. Whereby, the government here creates an expansive arrangement of data innovation with the fundamental target to reinforce cooperation with the citizens.

Many researches investigating what factors exist behind personal conduct towards choosing to utilize e-government services have been carried outside the Arab world. Thereupon, this research comes to assess the extent of e-government empowerment in different public sectors in Jordan, having into the picture of critical factors that may increase the dependence on e-government within the Jordanian community.

III. Related Work

The study of [8] revealed that an efficient e-government program needs a coherent set of strategies and policies to control the establishment of information systems, technical infrastructures, and the essential regulatory frameworks. E-government success is highly reliant on the strategies, which are incorporated into the perspective of various levels of the society. There is a lack of consensus within the literature regarding the significance of approach for e-government adoption [6]. Certain researches on e-readiness assessment tools emphasize that strategy of ICT is crucial for successful adoption of e-government [11]. While others hold that studies IT, approach does not possess a strong effect on e-government readiness [17]. Nevertheless, these findings as well as e-readiness evaluation tools consider the policy as the strategy of e-government not as administration organizational ICT approach. Scholars refer the portal as "a program of service transformation" since it serves as the umbrella for entire government administrations, authorities and departments [16].

The e-government program from the viewpoint of [19] can result in improved government effectiveness and efficiency in delivering proper services to different sectors of consumers via numerous delivery channels. What’s more, there exist three e-government portal complication levels that are transactions concerning incorporation of numerous organizations, single organization transactions, and information distributing as well as connecting of present web sites.

IV. Methodology

This study adopted qualitative research to adopt the strategy of in-depth focus interviewing employees working in different sectors of the governmental institutions. This research approach was chosen since it was suitable in obtaining detailed information concerning opinions, perceptions, and personal feelings on this topic. This research design allows collections of perceptions about e-government convenience process. The selection of the interviewees was based upon their willingness to participate from different sectors in Jordan such as: health, tele-communication companies, a social security, Jordan electricity company, municipality, and traffic department.

A pre-structured interview was done to assess the dependent variable that is going to be analyzed is the continuity of using e-government. This was conducted in a pleasant atmosphere. Other independent variables are going to be analyzed to see its effect on the dependent variable. These dependent variables are the government readiness to apply e-government for a hand, and on the other hand the citizen willingness to utilize e-government application with the cons and pros of using this application on both parties, the government and the citizens.

All questions were concerned with the e-government issues, like its benefits and online information privacy and security. Furthermore, the interviewees were also stimulated to make comparisons between online as well as offline government convenience. The strength of this research is that it provides detailed evaluation of the resources, services, and agency. Also, it backs strategy formulation at agency level. Conversely, it does not offer details required for effective strategy formulation at different levels of governance like community and local levels, of multi-level governance.

V. Result and Discussion

Assessing 16 participants did the research from different governmental sectors as it was mentioned in the methodology. However, the sample size of the participants was small but adequate for qualitative methods, since the primary concern of this research is placed on the deriving of considerable data rather than the verification of hypotheses [11].The textual outcomes of the focus group interviews were content assessed by the authors of this study.

All the differences were solved through meeting to capture significant facets of e-government aspects. The interview was constituted from five questions that assessed the dependent and the independent factors; they are going to be discussed as follows.

A. When do you Expect to Fully Rely on e-Government and the Traditional Paper Works Omitted?

All of the interviewees had the opinion that fully turning toward e-government will be soon, in line with the Jordanian Government's policy towards digitizing the transactions. Some of them had the opinion that the e-government process had already started as the employee in the traffic department, municipality, tele-communication companies and the Jordan electricity have stated:

"Paper works in our field of work cannot be omitted, however, the application of e-government has started to be applied and a website with usernames is being used for different services such as tax records and invoice numbers."

While another participant stated that it needs more time, as the employee of health department stated:

"As it is obvious from the updates to the system, and the new applied program of visits which is under the test now, I expect that within one year, and maximum for two years. Paperwork will be limited and the whole information will be saved on the computers such as the medications exist at the pharmacies, the medical prescriptions of the patients, and booking time with their doctors, etc."

Based on the employees' opinion, the project of e-government needs a short period may not exceed two years at their discretion.

B. Do you Encourage the use of e-Government? And why?

All of the participants encouraged the use of e-government; each one gave his/her perspective depending on the nature of his/ her employment.

The employee in health department suggesting the issue of honesty as he declared that:

"The access of data will be easier and the ability of manipulation of the data and stocks of medicines and other gadgets will be limited. Everything will be present before each single person in the department, so the censorship will be more intense. Each employee will be doing his/her job sincerely, and each patient will take his/her right in having the most proper health care."

One of the participants from the social security, however, had her doubts about the application of e-government as she remembered some bad experiences with the connection to the web:

"Sometimes I feel that it is a good step forward, until the connection with the network is disconnected, or the system when it gets down and transactions are stopped until reconstruction is done, I feel that this is a quiet endangering. However, we all hope that every bad and good expectation is given into consideration and a sort of pro-activeness is applied to overcome a bad circumstance before it occurs."

Moreover, saving time with a high level of accuracy and less chances to do mistakes were the points, which were repeated by the rest of the participants.

C. Do you Think that the Jordanian Society can Adopt e-Government in their Transactions?

This question had variations in answers among different participants, some of them said that it is possible; others said that it is possible but some obstacles might face the users.

Different economic levels are present in the Jordanian Society, and the vast majority of them are classified as a low economic level families and members.

From the perspective of an employee works in health stated that adopting e-government might be not easy to the Jordanian society. While employee in the Jordanian Electricity Company and municipality focused on the idea of change and the presence of willing on the society to adopt new methods, with the presence of doubts and fears to use these new methods as she stated:

"Change is hard, and the Jordanian Society is known to be reluctant to any new practice that would make them adopt new methods and train on the way of using them, even if it would make their lives easier.

In addition, the culture of doubt is present in most of the members of the society; they will be hardly convinced that this step will be for their own benefit. They will be worried about their safety and privacy, and convincing them will be hard."

From the side of tele-communication companies and traffic department were optimistic and believed that this awareness is spread to the members of the society. They also added:

"They can adopt e-government services; however, they will need to have an e-fawateercom account which needs to have a bank record which might be hard on some citizens"

D. Do you Think that the Government has the Appropriate Infrastructure for such a Change?

Most of them had the opinion that the government infrastructure is appropriate as a starting experience, however, it needs to be updated and more training to the employees is required. While one of the participants in the traffic department was very pleased regarding the government infrastructure, however, repairs are required based on the users' feedback as he stated:

"Yes and new gadgets are being spread to policemen to be directly linked to the net of traffic and vehicle departments' database. Based on this, we can say that the infrastructure is almost ready. He added that based on the feedback from users is the infrastructure needs to be updated continuously."

On the other hand, the view of employee in Communication Company was completely satisfied with the services of the e-government and the infrastructure of it as he stated:

"Yes, the system is working dynamically to be updated, to solve the problems that face users of the society. In addition, the government is trying to hire qualified employees with a good experience and a wide background in IT, so the hitches are going to be taken into consideration in advance."

E. What are the Factors that Affect e-Government Adoption?

After the analysis the factors can be separated into two groups, factors related to the government, and other factors related to the citizen.

The factors related to the government are dependent on the IT infrastructure and the employees' knowledge and willingness to follow-up and update it. While the factors related to the citizen are related to the economic situation of the citizen as well as the background and the culture of the families.

Table I summarizes forth the major dimensions of the critical facets of e-government derived from the results, and their associated descriptions.

Compared to our view of results, past researches about the selection of e-government has examined a few considers, for example, trust government, trust in web innovation, use of e-government value-based administration and others as a major issue in most recent five years [3,12,13]. In view of the related works over, the creators supported that those variables have generally proposed superior information on various sorts of e-government benefit from a few points of view.

In the meantime, previous research on e-government appropriation has for the most part centered on the developed nations; for example, learn concerning the acknowledgment as well as utilization of government Internet benefits in Netherland [14], a research focused on the client who uses e-government in Belgium [15], and assessment of government e-charge sites in Sweden [16].

Prominently, only a couple of research examined the selection in developing nations, for example, learn regarding e-government reception in Cambodia [17], and surveying resident appropriation of e-government activities in Gambia [18]. In outcome, little consideration was gotten to analyze e-government appropriation and utilize in the emergent nation as a rule.

Furthermore, [19] had focused on the application of e-government in Saudi Arabia. The qualitative research done and a survey were both done to find the presence of success factors such as presence of software and technology that support using e-government, presence of customer service, the level of education as well as computer skills present upon users, the extent of access between users, the level of privacy presented to the users in addition to security, factors related to religion and culture, financial ability, and other factors which might include age and gender.

The findings of this research had shown that the issues of e-government application success are present, and it is showing a bright future, as there is an extent of acceptance among the Saudi population.

Parallel with the study of [20] had focused on the opinion of e-government upon Jordanian citizens, and it was found that the perception of using e-government in Jordan is well due to seeing helpfulness, usability, social impact, respond, and similarity, yet the cost of this service is one of the suppression aspects that negatively affect using e-government. Another study was done earlier by [21], had assessed the satisfaction of the users using an electronic survey. The study demonstrated the significance to uncover the key drivers of e-fulfillment in order to give criticism in many proposals that will empower making e-Government gateway which are perfect with the people' needs and desires.

TABLE. I. DIMENSIONS OF E-GOVERNMENT

Dimensions	Description
Perceived Usefulness	Check-out process Payment method Reducing cost Saving time Saving energy Useful information
Perceived Ease to Use	Easy navigation Accessible anywhere Accessible anytime
Trust on Web	Personal data security
Social Influences	Word of mouth T.V. Radio Social media
IT Infrastructure	Internet coverage Computers and Tablets Websites design

VI. Conclusion

This study proposes dimensions of the most critical factors towards the individual's choice to utilize e-government services in Jordan. As so far, Jordan is going to follow the trend of adopting e-government, as it has positive effects on the quality of daily services transactions. In addition, spreading the awareness among the society members will encourage them to use this application.

In the end, the concept of e-government is constantly developing and changing to keep abreast of technological changes in term of designing and promoting services. In this context, there is an urgent necessity to find creative methods of evolving lasting beneficial services to fill gaps between service performance and people anticipations particularly in developing countries.

VII. Future Studies

The study needs to include more participants from different government sectors; in addition, it would give additional conclusions when citizens are included in the study. This study is not free of limitations, such as a qualitative technique is used in the study. Future studies therefore could use a quantitative study in seeking respondents' opinions. To put it plainly, A TAF model could be adopted to examine the most important factors affecting people's perception towards acceptance of e-government in Jordan.

References

[1] Laith T. Khrais. Toward A Model For Examining The Technology Acceptance Factors In Utilization The Online Shopping System Within An Emerging Markets, 9 (11), 2018, 1099- 1110.

[2] Laith T. Khrais. The Impact Dimensions of Service Quality on the AcceptanceUsage of Internet Banking Information Systems, American Journal of applied sciences, 15(4), 2018, 240-250.

[3] Alzahrani L, Al-Karaghouli W, Weerakkody V. Analysing the critical factors influencing trust in e-government adoption from citizens' perspective: A systematic review and a conceptual framework. International Business Review. 2017 Feb 1;26(1):164-75.

[4] Kacem A, Belkaroui R, Jemal D, Ghorbel H, Faiz R, Abid IH. Towards improving e-government services using social media-based citizen's profile investigation. InProceedings of the 9th International Conference on Theory and Practice of Electronic Governance 2016, 187-190. ACM.

[5] Hashim HS, Hassan ZB, Hashim AS. Factors influence the adoption of cloud computing: A comprehensive review. International Journal of Education and Research. 2015;3(7):295-306.

[6] Al-Khatib H, Lee H, Suh C, Weerakkody V. E-government systems success and user acceptance in developing countries: The role of perceived support quality. 2019.

[7] Zejnullahu F, Baholli I. Overview of researches on the influential factors of m-government's adoption. European Journal of Management and Marketing Studies. 2017 Aug 11.

[8] Moon M. J., Welch, E., W. & Wong, W. 2005. What Drives Global E-Governance? An Exploratory Study at a Macro Level. InProceedings of the 38th Hawaii International Conference on System Sciences. https://www.researchgate.net/publication/221179057_What_Drives_Global_EGovernance_An_Exploratory_Study_at_a_Macro_Level Date of access 2017 Mar (Vol. 12).

[9] Majdalawi YK, Almarabeh T, Mohammad H, Quteshate W. E-government strategy and plans in Jordan. Journal of Software Engineering and Applications. 2015 Apr 14;8(04):211.

[10] Witarsyah D, Sjafrizal T, Fudzee MD, Farhan M, Salamat MA. The critical factors affecting e-government adoption in Indonesia: A conceptual framework. International Journal on Advanced Science, Engineering and Information Technology. 2017 Feb 22;7(1):160-7.

[11] Izogo EE, Jayawardhena C. Online shopping experience in an emerging e-retailing market: Towards a conceptual model. Journal of consumer Behaviour. 2018 Jul;17(4):379-92.

[12] Alam MZ, Hu W, Barua Z. Using the UTAUT model to determine factors affecting acceptance and use of mobile health (mHealth) services in Bangladesh. Journal of Studies in Social Sciences. 2018 Dec 1;17(2).

[13] AL-Hujran, O., AL-Debei, M. M., Chatfield, A. and Migdadi, M. The imperative of influencing citizen attitude toward e-government adoption and use. Computers in human Behavior,53(1), 2015,189-203.

[14] Wirtz BW, Piehler R, Rieger V, Daiser P. E-government portal information performance and the role of local community interest. Empirical support for a model of citizen perceptions. Public Administration Quarterly. 2016 Apr 1:48-83.

[15] Wirtz BW, Kurtz OT. Local e-government and user satisfaction with city portals–the citizens' service preference perspective. International Review on Public and Nonprofit Marketing. 2016 Oct 1;13(3):265-87.

[16] Arias MI, Maçada AC. Digital Government for E-Government Service Quality: a Literature Review. InProceedings of the 11th International Conference on Theory and Practice of Electronic Governance 2018 Apr 4 (pp. 7-17). ACM.

[17] Rana NP, Dwivedi YK, Williams MD. A meta-analysis of existing research on citizen adoption of e-government. Information Systems Frontiers. 2015 Jun 1;17(3):547-63.

[18] Jung D. " Assessing citizen adoption of e-government initiatives in Gambia: A validation of the technology acceptance model in information systems success". A critical article review, with questions to its publishers. Government Information Quarterly. 2019 Jan 1;36(1):5-7.

[19] Basahel, A. and Yamin, M. Measuring success of e-government of Saudi Arabia. International Journal of Information Technology, 9(1**)**, 2017, 287-293.

[20] Abu-Shanab, E. and Haider, S. Major factors influencing the adoption of m-government in Jordan. Electronic Government. An International Journal, 11(1), 2015,223-240.

[21] Sachan A, Kumar R, Kumar R. Examining the impact of e-government service process on user satisfaction. Journal of Global Operations and Strategic Sourcing. 2018 Nov 19;11(3):321-36.

Generating and Analyzing Chatbot Responses using Natural Language Processing

Moneerh Aleedy[1]
Information Technology Department
College of Computer and Information Sciences
Princess Nourah bint Abdulrahman University
Riyadh, Saudi Arabia

Hadil Shaiba[2]
Computer Sciences Department
College of Computer and Information Sciences
Princess Nourah bint Abdulrahman University
Riyadh, Saudi Arabia

Marija Bezbradica[3]
School of Computing, Dublin City University, Dublin, Ireland

Abstract—**Customer support has become one of the most important communication tools used by companies to provide before and after-sale services to customers. This includes communicating through websites, phones, and social media platforms such as Twitter. The connection becomes much faster and easier with the support of today's technologies. In the field of customer service, companies use virtual agents (Chatbot) to provide customer assistance through desktop interfaces. In this research, the main focus will be on the automatic generation of conversation "Chat" between a computer and a human by developing an interactive artificial intelligent agent through the use of natural language processing and deep learning techniques such as Long Short-Term Memory, Gated Recurrent Units and Convolution Neural Network to predict a suitable and automatic response to customers' queries. Based on the nature of this project, we need to apply sequence-to-sequence learning, which means mapping a sequence of words representing the query to another sequence of words representing the response. Moreover, computational techniques for learning, understanding, and producing human language content are needed. In order to achieve this goal, this paper discusses efforts towards data preparation. Then, explain the model design, generate responses, and apply evaluation metrics such as Bilingual Evaluation Understudy and cosine similarity. The experimental results on the three models are very promising, especially with Long Short-Term Memory and Gated Recurrent Units. They are useful in responses to emotional queries and can provide general, meaningful responses suitable for customer query. LSTM has been chosen to be the final model because it gets the best results in all evaluation metrics.**

Keywords—Chatbot; deep learning; natural language processing; similarity

I. Introduction

With the arrival of the information age, customer support has become one of the most influential tools companies use to communicate with customers. Modern companies opened up communication lines (conversations) with clients to support them regarding products before and after-sales through websites, telephones, and social media platforms such as Twitter. This communication becomes faster and much easier with the support of the technologies that are being used today.

Artificial intelligence (AI) improves digital marketing in a number of different areas from banking, retail, and travel to healthcare and education. While the idea of using human language to communicate with computers holds merit, AI scientists underestimate the complexity of human language, in both comprehension and generation. The challenge for computers is not just understanding the meanings of words, but understanding expression in how those words are collocated. Moreover, a chatbot is an example of a virtual conversational service robot that can provide human-computer interaction. Companies use robotic virtual agents (Chatbot) to assist customers through desktop interfaces [1, 2].

Natural language processing (NLP) is a subfield of computer science that employs computational techniques for learning, understanding and producing human language content. NLP can have multiple goals; it can aid human-human communication, such as in machine translation and aid human-machine communication, such as with conversational agents. Text mining and natural language processing are widely used in customer care applications to predict a suitable response to customers, which significantly reduces reliance on call center operations [3].

AI and NLP have emerged as a new front in IT customer service chatbots. The importance of these applications appears when no technicians manage the customer service office due to the end of working time or their presence outside the office [4].

In this project, the main focus will be on the automatic generation of conversation "Chat" between a computer and a human by developing an interactive artificial intelligent agent using deep learning. This will provide customers with the right information and response from a trusted source at the right time as fast as possible.

This project aims to build an automated response system (Chatbot) that responds to customer queries on social networking platforms (Twitter) to accelerate the performance of the service. Also, to keep the simplicity in mind while designing the system to enhance its efficiency.

This project centers around the study of deep learning models, natural language generation, and the evaluation of the generated results.

We believe that this contribution can add improvement by applying the right preprocessing steps which may organize sentences in a better way and help in generating proper responses. On the other hand, we start with the existing text generative models CNN and LSTM and then try to improve them as well as develop a new model such as GRU to compare results. We focus on evaluating the generated responses from two aspects: the number of words matches between the reference response and the generated response and their semantic similarity.

The rest of this paper is organized as follows. Section II provides reviews of the related works. The methodological approach is described in Section III. Moreover, dataset collection and analysis in details are provided in Section IV. The implementation strategy and results of this project are discussed in section V. Finally, the conclusion of the project and its future work are provided in Sections VI and VII respectively.

II. Literature Review

Developing computational conversational models (chatbots) took the attention of AI scientists, for a number of years. Modern intelligent conversational and dialogue systems draw principles from many disciplines, including philosophy, linguistics, computer science, and sociology [5]. This section will explore the previous work of chatbots and their implementations.

A. Chatbots Applications and Uses

Artificial dialogue systems are interactive talking machines called chatbots. Chatbot applications have been around for a long time; the first well-known chatbot is Joseph Weizenbaum's Eliza program developed in the early 1960s. Eliza facilitated the interaction between human and machine through a simple pattern matching and a template-based response mechanism to emulate the conversation [6, 7].

Chatbot became important in many life areas; one of the primary uses of chatbots is in education as a question answering system for a specific knowledge domain. In [8], the authors proposed a system that has been implemented as a personal agent to assist students in learning Java programming language. The developed prototype has been evaluated to analyze how users perceive the interaction with the system. Also, the student can get help in registering and dropping courses by using a chatbot spatialized in student administrative problems, as mentioned in [9]. The administrative student's chatbot helps the colleges to have 24*7 automated query resolution and helps students have the right information from a trusted source.

On another hand, information technology (IT) service management is an important application area for enterprise chatbots. In many originations and companies, IT services desk is one of the essential departments that helps to ensure the continuity of work and solving technical problems that employees and clients are facing. This variability demands manual intervention and supervision, which affects the speed and quality of processes execution. IT service providers are under competitive pressure to continually improve their service quality and reduce operating costs through automation. Hence, they need the adoption of chatbots in order to speed up the work and ensure its quality [10].

On the medical side, the field of healthcare has developed a lot, lately. This development appears with the use of information technology and AI in the field. In [11], the authors proposed a mobile healthcare application as a chatbot to give a fast treatment in response to accidents that may occur in everyday life, and also in response to the sudden health changes that can affect patients and threaten their lives.

Customer services agent is an application of applying chatbot technologies in businesses to solve customer problems and help the sales process. As companies become globalized in the new era of digital marketing and artificial intelligence, brands are moving to the online world to enhance the customer experience in purchasing and provide new technical support ways to solve after-sales problems. Moreover, fashion brands such as Burberry, Louis Vuitton, Tommy Hilfiger, Levi's, H&M, and eBay are increasing the popularity of e-service agents [1].

B. Natural Language Processing

NLP allows users to communicate with computers in a natural way. The process of understanding natural language can be decomposed into the syntactic and semantic analysis. Syntactic refers to the arrangement of words in a sentence such that they make grammatical sense. Moreover, syntactic analysis transforms sequences of words into structures that show how these words are related to each other. On the other hand, semantic refers to the meaning of each word and sentence. The semantic analysis of natural language content captures the real meaning; it processes the logical structure of sentences to find the similarities between words and understand the topic discussed in the sentences [12].

As part of the text mining process, the text needs many modification and cleaning before using it in the prediction models. As mentioned in [13], the text needs many preprocessing steps which include: removing URLs, punctuation marks and stop words such as a, most, and, is and so on in the text because those words do not contain any useful information. In addition, tokenizing, which is the process of breaking the text into single words. Moreover, text needs stemming, which means changing a word into its root, such as "happiness" to "happy". For features extraction, the authors use Bag of Words (BoW) to convert the text into a set of features vector in numerical format. BoW is the process of transforming all texts into a dictionary that consist of all words in the text paired with their word counts. Vectors are then formed based on the frequency of each word appearing in the text.

Before entering the data into a model or a classifier, it is necessary to make sure that the data are suitable, convenient, and free of outliers. In [14], the authors explain how to preprocess the text data. The main idea was to simplify the text for the classifier to learn the features quickly. For example, the names can be replaced with one feature {{Name}} in the feature set, instead of having the classifier to learn 100 names from the text as features. This will help in grouping similar features together to build a better predicting classifier. On another hand, emoticons and punctuation’s marks are converted to indicators (tags). Moreover, a list of emoticons is compiled from online sources and grouped into categories. Other punctuation marks that were not relevant to the coding scheme are removed.

Chat language contains many abbreviations and contractions in the form of short forms and acronyms that have to be expanded. Short forms are shorter representations of a word which are done by omitting or replacing few characters, e.g., grp → group and can’t → cannot. The authors created a dictionary of these words from the Urban Dictionary to replace abbreviations by expansions. Spell checking is performed as the next step of the pre-processing pipeline on all word tokens, excluding the tagged ones from the previous steps [14].

Minimizing the words during the text pre-processing phase as much as possible is very important to group similar features and obtain a better prediction. As mentioned in [15], the authors suggest processing the text through stemming and lower casing of words to reduce inflectional forms and derivational affixes from the text. The Porter Stemming algorithm is used to map variations of words (e.g., run, running and runner) into a common root term (e.g., run).

Words can not be used directly as inputs in machine learning models; each word needs to be converted into a vector feature. In [4], the authors adopt the Word2vec word embedding method to learn word representations of customer service conversations. Word2vec's idea is that each dimension of inclusion is a possible feature of the word, which can capture useful grammatical and semantic properties. Moreover, they tokenize the data by building a vocabulary of the most frequent 100K words in the conversations.

C. *Machine Learning Algorithm and Evaluation*

A large number of researchers use the idea of artificial intelligence and deep learning techniques to develop chatbots with different algorithms and methods. As mentioned in [16], the authors use a repository of predefined responses and a model that ranks these responses to pick an appropriate response for a user’s input. Besides, they proposed topic aware convolutional neural tensor network (TACNTN) model to classify whether or not a response is proper for a message. The matching model used to select a response for a user message. Specifically, it has three-stages that include: pre-processing the message, retrieving response candidates from the pre-defined message-response pair index, then ranking the response candidates with a pre-train matching model.

In [17], the authors train two word-based machine learning models, a convolutional neural network (CNN) and a bag of words SVM classifier. Resulting scores are measured by the Explanatory Power Index (EPI). EPI used to determine how much words contribute to the classification decision and filter relevant information without an explicit semantic information extraction step.

The customer service agent is an important chatbot that is used to map conversations from request to the response using the sequence to sequence model. Moreover, a sequence to sequence models has two networks one work as an encoder that maps a variable-length input sequence to a fixed-length vector, and the other work as a decoder that maps the vector to a variable-length output sequence. In [4], the authors generate word-embedding features and train word2vec models. They trained LSTMs jointly with five layers and 640 memory cells using stochastic gradient descent for optimization and gradient clipping. In order to evaluate the model, the system was compared with actual human agents responses and the similarity measured by human judgments and an automatic evaluation metric BLEU.

As a conclusion of reviewing works concerned with the conversational system, text generation in English language and the collaboration of social media in customer support service, this paper proposes a work that aims to fill the gap of limited works in the conversational system for customer support field, especially in the Twitter environment. The hypothesis of this project was aiming to improve the automated responses generated by different deep learning algorithms such as LSTM, CNN, and GRU to compare results and then evaluate them using BLEU and cosine similarity techniques. As a result, this project will help to improve the text generation process in general, and customer support field in particular.

III. Methodological Approach

This section discusses the background of the implemented methods, explain why these methods are appropriate and give an overview of the project methodology.

A. *Text Generative Model*

Based on the nature of this project, which is generating a proper response to every customer query in social media, applying sequence-to-sequence learning are needed. Moreover, sequence-to-sequence means mapping a sequence of words representing the query to another sequence of words representing the response, the length of queries and responses can be different. This can be applied by the use of NLP and deep learning techniques.

Sequence-to-sequence models are used in many fields, including chat generation, text translation, speech recognition, and video captioning. As shown in Fig. 1, a sequence-to-sequence model consists of two networks, encoder, and decoder. The input text enters the encoder network in reverse order, then it is converted into a sequence of fixed length context vector, which is then used by the decoder to generate the output sequence [18].

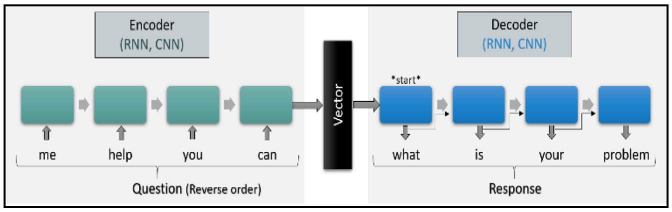


Fig. 1. Sequence to Sequence Model.

Before inserting the sequence of words into the encoder model, it needs to be converted into a numerical format; this can be applied by using NLP techniques. This project focused on Bag of Words, or BoW vector representations, which is the most commonly used traditional vector representation for text generating models. BoW is used to transforms all texts into a dictionary that consists of all words that appear in the document [13]. It then creates a set of features in real number inside a vector for each text.

B. *Deep Learning Models*

1) Convolutional Neural Network (CNN) Model: In this project, CNN is chosen mainly for its efficiency, since CNN is faster compared to other text representation and extraction methods [19]. The CNN consists of the convolution and pooling layers and provides a standard architecture that takes a variable-length sequence of words as an input and then passes it to a word embedding layer. The embedding layer maps each word into a fixed dimensional real-valued vector then passes it to the 1D convolutional layer. The output is then further down-sampled by a 1D max-pooling layer. Outputs from the pooling layers are then fed into the final output layer to produce a fixed-length feature vector [20]. CNN has been widely used in image and video recognition systems, and, lately, they have shown promising results in NLP applications [21]. Fig. 2 shows the standard architecture of the CNN model.

2) Recurrent Neural Network (RNN) Model: In a traditional neural network, all inputs and outputs are independent of each other, which is not useful when working with sequential information. Predicting the next word in a sentence requires knowing the sequence of the words in the sentence that come before the predicted word. Among all models for learning sentence representations, recurrent neural network (RNN) models, especially the Long Short Term Memory (LSTM) model, are the most appropriate models for processing sentences, as they have achieved substantial success in text categorization and machine translation [22]. Therefore, this project applies LSTM and Gated Recurrent Units (GRU) as a newer generation of Recurrent Neural Networks. Fig. 3 illustrates the basic architecture of RNN.

Hochreiter & Schmidhuber introduced Long Short Term Memory Networks in 1997. They solve the problem of vanishing and exploding gradient problem that is prevalent in a simple recurrent structure, as it allows some states to pass without activation. In 2014, Cho et al developed GRU networks in an effort to design recurrent encoder-decoder architecture [23]. They are relatively more straightforward than LSTM and retain a majority of its advantages.

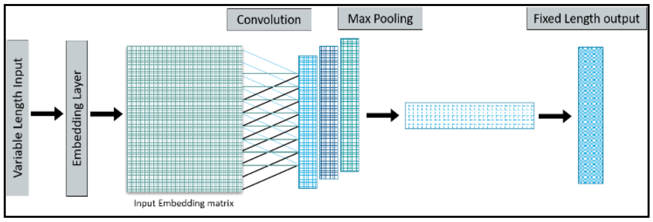


Fig. 2. The Architecture of CNN.

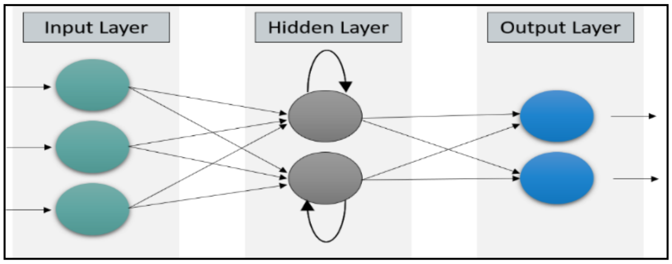


Fig. 3. The Architecture of RNN.

C. *Project Methodology*

In order to implement this project, several preprocessing and modeling steps are performed. First, split the original dataset into train and test sets. Then, prepare the dataset for modeling. The preparing process includes preprocessing steps and features extraction. After that, train models using train set with LSTM, GRU, and CNN. Finally, prepare the test set and use it for evaluating the models. Fig. 4 illustrates the methodology steps.

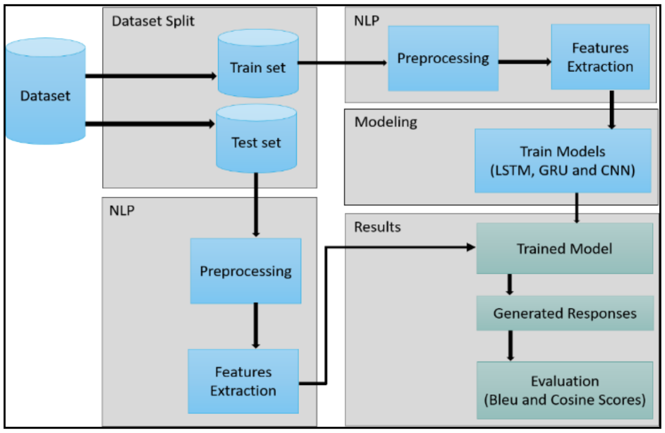


Fig. 4. The General Implementation Steps.

IV. Dataset Collection and Analysis

The dataset "Customer Support on Twitter" from Kaggle is used to develop and evaluate the models. The original dataset includes information such as: tweet_id, author_id, inbound, created_at, text, response_tweet_id and in_response_to_ tweet_id. The description of the original dataset is shown in Table I.

The original dataset contains 2,811,774 collections of tweets and replies from the biggest brands on Twitter as customer support (tweets and replies are in different rows). Moreover, the number of brands in the dataset is 108, and they responded to queries from 597075 users. Fig. 5 shows the top 10 customer support responses per brand.

While performing exploratory analysis on the dataset, it has been noticed, for instance, that Amazon customer support handles a lot of questions (around 84600 in seven months) which is a huge number to deal with if we consider the working hours and working days per week. Also, some of the questions have a delay in responding or had no responses at all. Fig. 6 shows the average delay in response to customers in hours per brand.

TABLE. I. DATASET FEATURES DESCRIPTION

Feature	Description	Datatypes
tweet_id	A unique, anonymized ID for the Tweet. Referenced by response_tweet_id and in_response_to_tweet_id.	int64
author_id	A unique, anonymized user ID. The real user_id in the dataset has been replaced with their associated anonymized user ID.	object
Inbound	Whether the tweet is "inbound" to a company doing customer support on Twitter. This feature is useful when reorganizing data for training conversational models.	bool
created_at	Date and time when the tweet was sent.	object
Text	Tweet content. Sensitive information like phone numbers and email addresses are replaced with mask values like __email__.	object
response_tweet_id	IDs of tweets that are responses to this tweet, comma-separated.	object
in_response_to_tweet_id	ID of the tweet this tweet is in response to, if any.	float64

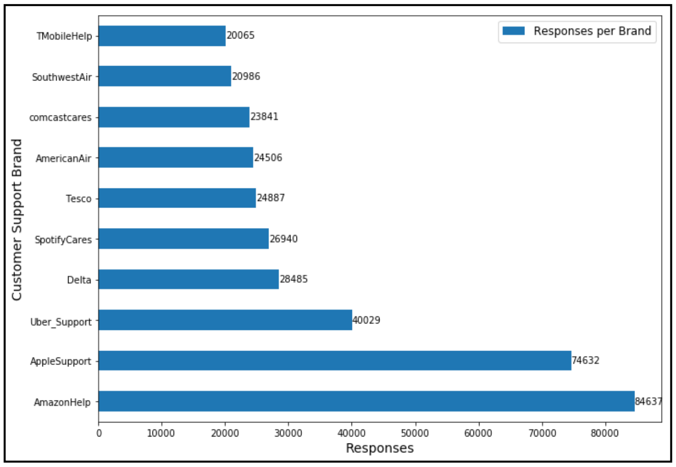


Fig. 5. Top 10 Customer Support Responses Per Brands.

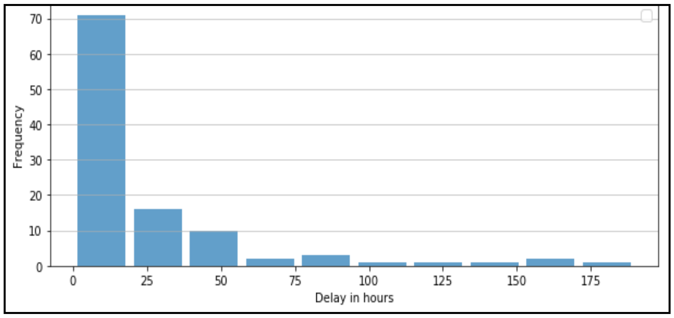


Fig. 6. The Average Delay in Response to Customers in Hours per Brand.

As shown in the above figure, around ten brands take more than two days (60 hours) to respond to customers queries, which may cause problems to customers, effect companies' reputation and the customers may start looking for other service providers.

A filtering process is used to convert the dataset records into a conversational dataset suitable for the experiments. The filtering is done as follows:

1) Pick only inbound tweets that are not in reply to any other tweet.

2) Organize each tweet with the corresponding reply by matching in_response_to_tweet_id with tweet_id features.

3) Filter out cases where reply tweets are not from a company based on the in inbound feature (if the inbound feature is False it means that the tweet is from a company; otherwise it is from a user).

However, when revising the dataset, it has been found that some of the tweets have no replies at all; they are from multiple languages, and some of them are just samples and emojis. For this type of tweets further preprocessing step is performed to remove non-English tweets by the use of the langdetect library which detects any non-English text [24]. Then, the non-responses English tweets are studied, as shown in the word cloud in Fig. 7, (which is a graph that illustrates the most words that appear in the text).

It can be observed that the words appear with no hint to a specific problem discussed, and most of the queries are thanking the customer support services for example:

- @AmazonHelp Thanks for the quick response
- @AppleSupport Awesome, thanks

Others asking for help in general:

- @Uber_Support Sent a DM Hope you could help soon.
- @O2 DM sent. Still no further forward!

The modified dataset contains 794,299 rows and 6 columns which are: author_id_x, created_at_x, text_x, author_id_y, created_at_y and text_y. X refers to the queries, and Y refers to the responses from customer support teams.

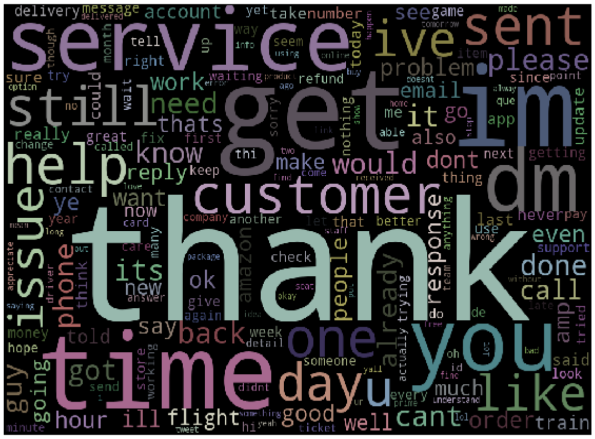


Fig. 7. Most Words used in the Queries without Responses Data.

V. Implementation Strategy

In this section, we are going to explain the methodology followed for this project. At first, prepare the dataset for modeling. The preparing process includes preprocessing step and features extraction then train the models using a training set and evaluate them with a test set.

A. Data Preprocessing

A data analyst cannot handle raw text directly to suit machine learning or deep learning methods. Therefore, it is necessary to work on texts' preprocessing from all existing impurities, for example, punctuation, expression code, and non-English words (Chinese, Spanish, French, and others). In order to do this, a number of python NLP libraries such as regular expression (RE), unicodedata, langdetect, and contractions are used.

In this project, the performed preprocessing steps include: remove links, images, Twitter ID, numbers, punctuation, emoji, non-English words and replace abbreviations with long forms. Table II illustrates the changes in the dataset before and after applying all the previous preprocessing steps.

The preprocessing steps are chosen carefully; not all preprocessing techniques are suitable for this kind of projects. For example, removing stopwords and text stemming cannot be applied because it will affect the sentences structures as well as the text generation process.

B. Feature Extraction

Before doing any complex modeling, the dataset needs to be transformed into a numerical format suitable for training. The Bag of Words (BOW) concept is applied to extract features from the text dataset. First, all of the texts in the dataset are split into an array of tokens (words). Then, a vocabulary dictionary is built with all of the words in the dataset and its corresponding index value. The array of words is then converted to an array of indexes. This process can be applied by the use of the sklearn' predefined method called CountVectorizer.

In order to handle variable length, the maximum sentence length needs to be decided. Moreover, all remaining vector positions should be filled with a value ('1' in this case) to make all sequences have the same length. On the other hand, words not in the vocabulary dictionary will be represented with UNK as a shortcut of unknown words. Moreover, each output text in the dataset will start with a start flag ('2' in this case) to help in training. Now the dataset is ready for training.

C. Modeling

The infrastructure used for experimentation involves google colaboratory and Crestle cloud services which are GPU-enabled Jupyter environments with powerful computing resources. All popular scientific computing and deep learning packages are pre-installed and configured to run on a GPU.

The experiments are applied using three different models LSTM, GRU, and CNN. The models use a training dataset of around 700k pairs of queries and responses and a testing dataset of 30k of unseen data. Training time is between 5 and 12 hours, depending on the model (see Table III).

TABLE. II. The Changes in Text before and after Applying Preprocessing Steps

Before preprocessing	After preprocessing
@115743 C91. Feel free to keep an eye on the PS Blog for news and updates: https://t.co/aLtfBAztyC	feel free to keep an eye on the ps blog for news and updates
@133100 We do our best to clear as many upgrades as we can, send us a DM with the reservation you're referring to and we'll take a look.	we do our best to clear as many upgrades as we can send us a dm with the reservation you are referring to and we will take a look
@129388 We'd like to look into this with you. To confirm, did you update to iOS 11.1? Please DM us here: https://t.co/GDrqU22YpT	we would like to look into this with you to confirm did you update to ios please dm us here

TABLE. III. Training Time in Hours

Model	Training Time in Hours
LSTM	12
GRU	8
CNN	5

In the experiments, multiple parameters are tested, and their effects are addressed. All models are tested with varying dimensionality of the word embeddings (100, 300 and 640), it was observed that models perform better and faster with 100-word embedding size.

The dataset is large, the number of vocabularies is 388,950 unique words, and our computers cannot handle it. So, only the frequent words appeared in the dataset should be used. The most frequent words are decided by the max_features parameter in the CountVectorizer function which sort words by its frequency then choose the most frequent words. The first vocabulary size in the experiments is 8000 and then it increases, taking into consideration memory limitation. A slight improvement has been recognized in all models and because of the memory limitation, only 10,000 of the vocabularies are used. Moreover, the GRU model was trained for eight epochs but without significant improvement. The three models are all trained under the same conditions. Table IV shows the common parameters used in all models.

TABLE. IV. The Common Parameters used in LSTM, GRU and CNN Models

Parameter	Value
Word embedding dimension size	100
Vocabulary size	10,000
Context dimension size	100
Learning rate	0.001
Optimization function	Adam
Batch size	1000 (the max that our computer can handle)
Max message length	30

The following are the common layers used in the models, starting from inserting the sequence of words into the model to generating the responses:

- Last Word Input Layer: Inputs the last word of the sequence.
- Encoder Input Layer: Inputs sequence data and pass it to the embedding layer.
- Embedding Layer: Used to create word vectors for incoming words.
- Encoder Layer (LSTM, GRU, CNN): Creates a temporary output vector from the input sequence.
- Repeated Vector Layer: Used like an adapter to fit the encoder and decoder parts of the network together. It can be configured to repeat the fixed-length vector one time for each time step in the output sequence.
- Concatenate Layer: Takes inputs and concatenates them along a specified dimension.
- Decoder Layer (LSTM, GRU, CNN)(Dense): Used as the output for the network.
- Next Word Dense Layer: Takes inputs from the previous layer and outputs a one vector representing the target word.
- Next Word softmax Layer: Applies a softmax function that turns the dense layer output into a probability distribution, from to pick the most likely next word.

D. Generating Responses

After training the models, the generating responses process is started using the 30k test set. The following are samples of the generated responses from all models (see Fig. 8 and 9).

E. Evaluation

The Bilingual Evaluation Understudy and cosine similarity evaluation metrics are used to compute the similarity between the generated response and the reference response.

1) Bilingual Evaluation Understudy (BLEU): BLEU was originally created to measure the quality of machine translation with respect to human translation. It calculates an n-gram precision (An n-gram is a sequence of n words that appear consecutively in the text) between the two sequences and also imposes a commensurate penalty for machine sequence being shorter than human one. A perfect match score is 1.0, whereas a perfect mismatch score is 0.0.

The computation of BLEU involves various components: n-gram precisions (Pn) and BLEU's brevity penalty. Those measures are calculated as shown in the following steps:

- Calculate n-gram precision (Pn): measures the frequency of the n-gram according to the number of times it appears in the generated response and reference response. Pn must be calculated for each value of n, which usually ranges from 1 to 4. Then the geometric average of Pn should be computed with a weighted sum of the logarithms of Pn.
- Calculate brevity penalty (equation 1): a penalization is applied to short answers, which might be incomplete.

$$\mathrm{BP}=\begin{cases}1 & \text{if } c > r \\ e^{(1-(r/c))} & \text{if } c \le r\end{cases} \qquad (1)$$

, where c is the length of generated response and r is the length of reference response.

- Then, calculate the BLEU score (equation 2) [23]:

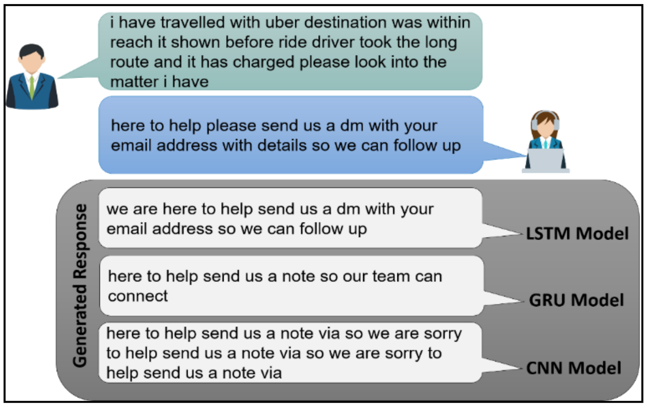
$$\mathrm{BLEU}= BP\ .\ e^{\sum_{n=1}^{N}(Wn\ .\ log\ (\mathrm{Pn}))} \qquad (2)$$

, where Wn = 1/N.

2) Cosine Similarity: On the other hand, cosine similarity also used to compute the similarity between the generated response and the reference response in vector representation. If there is more similarity between the two vectors, the cosine similarity value is near to one; otherwise, it is near to zero.

3) In order to implement the cosine similarity, the pre-trained model word2vec are used. The word2vec model is in gensim package, and it has been trained on part of Google News dataset (about 100 billion words) [25]. The model contains 300-dimensional vectors for 3 million words and phrases.

The word2vec model used to represent words in a vector space [26]. Words are represented in the form of vectors and placement is done in such a way that similar meaning words appear together, and different words are located far away.



i have travelled with uber destination was within reach it shown before ride driver took the long route and it has charged please look into the matter i have

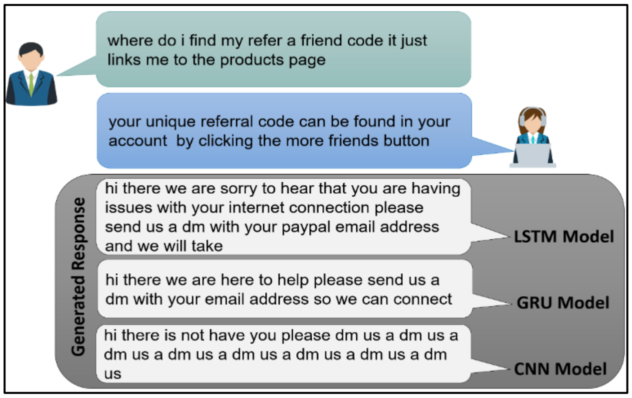
here to help please send us a dm with your email address with details so we can follow up

we are here to help send us a dm with your email address so we can follow up

here to help send us a note so our team can connect

here to help send us a note via so we are sorry to help send us a note via so we are sorry to help send us a note via

Fig. 8. Good Result Example.



where do i find my refer a friend code it just links me to the products page

your unique referral code can be found in your account by clicking the more friends button

hi there we are sorry to hear that you are having issues with your internet connection please send us a dm with your paypal email address and we will take

hi there we are here to help please send us a dm with your email address so we can connect

hi there is not have you please dm us a dm us a dm us a dm us a dm us a dm us a dm us

Fig. 9. Bad Result Example.

Gensim is a topic modeling toolkit which is implemented in python. Topic modeling is discovering the hidden structure in the text body. Word2vec model is imported from Gensim toolkit and uses a built-in function to calculate the similarity between the generated response and reference response.

F. *Result and Discussion*

Before discussing and reviewing the results, the most important features of the baseline model are discovered to have a rich discussion with clear comparisons. Table V shows the baseline model implementation.

In this project, the process of generating responses take around 6 hours for each model to be accomplished. Moreover, calculating BLEU and cosine similarity scores takes around 4 hours.

The models are evaluated automatically based on the words using BLEU score. The BLEU is applied for 1-gram, 2-gram, 3-gram, and 4-gram in order to explore the strength of the models. It can be seen that LSTM and GRU models outperform the official baseline LSTM model [4] with respect to the 4-gram BLEU score. Fig. 10, shows in details the performance of models in each n-gram.

Hence it can be seen that LSTM achieves the highest evaluation scores for all grams, but it takes a long time in training. Moreover, the GRU model has very close evaluation scores to LSTM. In the other hand, the CNN model has the lowest evaluation scores compared with all RNN models but achieves high-speed performance, which can be useful in application trained on large datasets.

TABLE. V. BASELINE MODEL IMPLEMENTATION

Preprocessing	Remove non-English queries, queries with images and @mentions.
Feature extraction	Word2vec
Model	LSTM with five layers.
Embedding size	640
Optimization Function	Stochastic gradient descent and gradient clipping.
Evaluation	BLEU with the best score achieved 0.36.

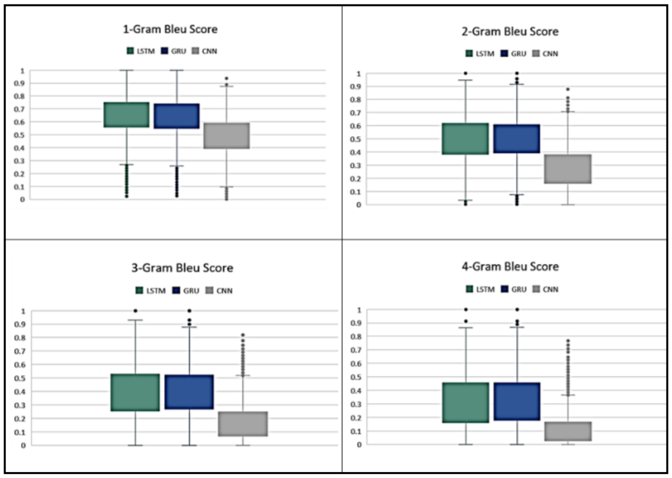


Fig. 10. The BLEU Scores for 1, 2, 3 and 4 Grams.

Furthermore, another evaluation metric cosine similarity are applied to captures the semantics beyond responses and gives similarity scores. It has been found that RNN models capture the semantics in the responses and they are more effective in improving the reply quality than the CNN model. Fig. 11 shows the similarity scores for each model.

After exploring the generated responses and get in-depth in the good and bad results, it has been found that RNN models, in general, are good in responses to emotional queries more than an informative one. The models can provide general, meaningful responses suitable for customer query. Table VI shows an example of an emotional query.

On the other hand, the queries that are more informative and ask about specific information are hard to generate, and the generated responses become less efficient. Table VII shows an example of an informative query.

By looking at the different responses from different models, it has been noticed that LSTM is generating better sentences that make sense and it is hard to say if the response is from a human or machine whereas GRU responses are not as good as LSTM.

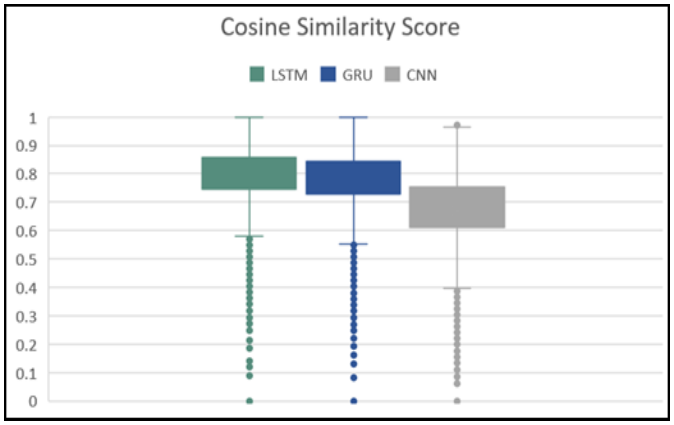


Fig. 11. The Cosine Similarity Scores.

TABLE. VI. EXAMPLE OF EMOTIONAL QUERY AND RESPONSES FROM ALL MODELS

Customer Query	my package is days late and i am leaving tomorrow on holidays could you please help it is extremely
Customer Support Response	sorry to hear this please dm us your tracking and phone number
LSTM Generated Response	i am sorry for the trouble with your order please report this to our support team here and we will check this
GRU Generated Response	i am sorry for the trouble with your order please reach out to us here and we will look into this for you please do not provide your order details
CNN Generated Response	hi there is not provide your order number and we can you please dm us a dm us a dm us a dm us a dm us

TABLE. VII. EXAMPLE OF INFORMATIVE QUERY AND RESPONSES FROM ALL MODELS

Customer Query	guys when are you going to open your services in middle east
Customer Support Response	hulu store is only available in the us at this time but we will share the interest in bringing our service to the middle east
LSTM Generated Response	hi there we are sorry to hear about this please dm us with your email address so we can connect
GRU Generated Response	hi there i am sorry to hear about this please dm me the details of the issue you are having with your services
CNN Generated Response	hi there is not have you are you

VI. CONCLUSION

In this project, we build customer support chatbot that helps companies to have 24 hours of automated responses. After analyzing the dataset and understanding the importance to have automated responses to customers and companies, we start exploring existing techniques used for generating responses in the customer service field. Then, we attempt to try three different models LSTM, GRU, and CNN. The experimental results show that LSTM and GRU models(with modified parameters) tend to generate more informative and valuable responses compared to CNN model and the baseline model LSTM. Besides, we used a BLEU score and cosine similarity as evaluation measures to support the final decision.

VII. FUTURE WORK

In future work, we plan to incorporate other similarity measures such as soft cosine similarity. Also, we plan to improve the experiments by increase the vocabulary size and try to increase the epoch parameters to reach 100 after providing proper infrastructure. We further can add more data for the training by taking benefits from the queries without responses and translate non-English queries.

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] M. Chung, E. Ko, H. Joung, and S. J. Kim, “Chatbot e-service and customer satisfaction regarding luxury brands,” J. Bus. Res., Nov. 2018.

[2] J. Hill, W. Ford, I. F.-C. in H. Behavior, and undefined 2015, “Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations,” Elsevier.

[3] J. Hirschberg and C. D. Manning, “Advances in natural language processing,” Science (80-.)., vol. 349, no. 6245, pp. 261–266, Jul. 2015.

[4] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, “A New Chatbot for Customer Service on Social Media,” in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI ’17, 2017, pp. 3506–3510.

[5] S. Oraby, P. Gundecha, J. Mahmud, M. Bhuiyan, and R. Akkiraju, “Modeling Twitter Customer ServiceConversations Using Fine-Grained Dialogue Acts,” in Proceedings of the 22nd International Conference on Intelligent User Interfaces - IUI ’17, 2017, pp. 343–355.

[6] H. Shah, K. Warwick, J. Vallverdú, and D. Wu, “Can machines talk? Comparison of Eliza with modern dialogue systems,” Comput. Human Behav., vol. 58, pp. 278–295, May 2016.

[7] R. DALE, “The return of the chatbots,” Nat. Lang. Eng., vol. 22, no. 05, pp. 811–817, Sep. 2016.

[8] M. Coronado, C. A. Iglesias, Á. Carrera, and A. Mardomingo, “A cognitive assistant for learning java featuring social dialogue,” Int. J. Hum. Comput. Stud., vol. 117, pp. 55–67, Sep. 2018.

[9] S. Jha, S. Bagaria, L. Karthikey, U. Satsangi, and S. Thota, “STUDENT INFORMATION AI CHATBOT,” in International Journal of Advanced Research in Computer Science, 2018, vol. 9, no. 3.

[10] P. R. Telang, A. K. Kalia, M. Vukovic, R. Pandita, and M. P. Singh, “A Conceptual Framework for Engineering Chatbots,” IEEE Internet Comput., vol. 22, no. 6, pp. 54–59, Nov. 2018.

[11] K. Chung and R. C. Park, “Chatbot-based heathcare service with a knowledge base for cloud computing,” Cluster Comput., pp. 1–13, Mar. 2018.

[12] J. Savage et al., “Semantic reasoning in service robots using expert systems,” Rob. Auton. Syst., vol. 114, pp. 77–92, Apr. 2019.

[13] S. T. Indra, L. Wikarsa, and R. Turang, “Using logistic regression method to classify tweets into the selected topics,” in 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2016, pp. 385–390.

[14] A. Shibani, E. Koh, V. Lai, and K. J. Shim, “Assessing the Language of Chat for Teamwork Dialogue,” 2017.

[15] A. Singh and C. S. Tucker, “A machine learning approach to product review disambiguation based on function, form and behavior classification,” Decis. Support Syst., vol. 97, pp. 81–91, May 2017.

[16] Y. Wu, Z. Li, W. Wu, and M. Zhou, “Response selection with topic clues for retrieval-based chatbots,” Neurocomputing, vol. 316, pp. 251–261, Nov. 2018.

[17] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek, “&quot;What is relevant in a text document?&quot;: An interpretable machine learning approach,” PLoS One, vol. 12, no. 8, p. e0181142, Aug. 2017.

[18] S. Sen and A. Raghunathan, “Approximate Computing for Long Short Term Memory (LSTM) Neural Networks,” IEEE Trans. Comput. Des. Integr. Circuits Syst., vol. 37, no. 11, pp. 2266–2276, Nov. 2018.

[19] Z. Wang, Z. Wang, Y. Long, J. Wang, Z. Xu, and B. Wang, “Enhancing generative conversational service agents with dialog history and external knowledge I,” 2019.

[20] J. Zhang and C. Zong, “Deep Neural Networks in Machine Translation: An Overview,” IEEE Intell. Syst., vol. 30, no. 5, pp. 16–25, Sep. 2015.

[21] R. C. Gunasekara, D. Nahamoo, L. C. Polymenakos, D. E. Ciaurri, J. Ganhotra, and K. P. Fadnis, “Quantized Dialog – A general approach for conversational systems,” Comput. Speech Lang., vol. 54, pp. 17–30, Mar. 2019.

[22] G. Aalipour, P. Kumar, S. Aditham, T. Nguyen, and A. Sood, “Applications of Sequence to Sequence Models for Technical Support Automation,” in 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4861–4869.

[23] J. Singh and Y. Sharma, “Encoder-Decoder Architectures for Generating Questions,” Procedia Comput. Sci., vol. 132, pp. 1041–1048, 2018.

[24] N. Shuyo, “Language Detection Library for Java.” 2010.

[25] R. Rehurek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, 2010, pp. 45–50.

[26] Y. Zhu, E. Yan, and F. Wang, “Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec.,” BMC Med. Inform. Decis. Mak., vol. 17, no. 1, p. 95, Jul. 2017.

The Criteria for Software Quality in Information System: Rasch Analysis

Wan Yusran Naim Wan Zainal Abidin[1], Zulkefli Mansor[2]
Faculty of Information Science and Technology
National University of Malaysia, Jalan Bangi
43000 Bangi, Selangor DE, Malaysia

***Abstract*—Most of the organization uses information system to manage the information and provide better decision making in order to deliver high quality services. Due to that the information system must be reliable and fulfill the quality aspect in order to accommodate organization's need. However, some of the information system still facing problems such as slow response time, problem with accessibility and compatibility issues between hardware and software. These problems will affect the acceptance and usage of the information system especially for non-computing users. Therefore, this study was aimed to investigate the factors that significantly contribute to the quality of software for information system. A survey was carried out by distributing a set of questionnaires to 174 respondents who are involved in development of software for information system. The data was analyzed using Rasch Measurement Model since it provides reliability of respondents and instruments. The result indicates that 30 factors had significantly contributed to the quality of software for information system and of these, six factors are under functionality, five for reliability, ten for usability, five for efficiency, two for compatibility and two for security. It is hoped that by identifying these factors, system developers can seriously consider of enhancing the quality of software for information system projects. In future, these factors can be used to develop an evaluation tool or metrix for quality aspects of software for information system projects.**

***Keywords*—Information system; quality of software; Rasch measurement model; evaluation; factors**

I. Introduction

Information system is important in developing successful and competitive organizations that can deliver high quality products and services to customers [1,2]. It helps improving the flow of information and work processes in organisation, thus can enhance the decision-making processes. Even though the information system provides benefits to the organization, it has been highlighted some weaknesses that contributed to the failure of fulfilling the quality aspects of a system such as slow response time, access problem, difficulties in using system, unavailability and incompatibility between hardware and software [3,4].

As a result, if these problems continuously faced by the users, it will cause less acceptance and usage of information systems. Thus, it leads to the poor delivery of services and products and finally will damage the organization's reputation. Therefore, system developers should focus to strengthen the software quality aspect of information system.

There are many software quality models currently being used to evaluate quality of software products. So, this paper will firstly show the comparisons between these models in the literature review section. The advantages and disadvantages of these models are also investigated. Based on these analyses, the ISO 25010 model was adopted in this study. Next, the paper discusses the methodology used in the study, which includes the descriptions of constructs and the explanation of the Rasch model used to perform data analysis. The paper also shows and discusses findings based on main assumptions of the Rasch Model for selection of items such as item fit, unidimensionality and local independence.

The study is very important in that it enables the improvement towards the information system development by having a guideline on factors that significantly improve the software quality aspect. It also serves as an additional reference towards the improvement of software quality in information systems.

However, this study only focuses on the human resource information system widely used in the planning and management of human resource. It also considers the software quality factors from the users' perspective only because software quality issues are usually related to this perspective.

II. Literature Review

Various models have been developed to measure software quality for information system such as the McCall, Boehm, FURPS, Dromey, ISO 9126 and ISO 25010 models. Each model was developed based on a certain unique principal or concept. These models explain about different aspects of software characteristics [5]. These models can be viewed from a user perspective, a manufacturing perspective or a product perspective. Table I below shows the comparison between these models.

These models also have the advantages and disadvantages of their own as stated in Table II below.

A. Factors Influencing Software Quality

This study also analyzed previous studies to identify the quality dimensions and factors that were used to measure the software quality. Table III shows the type of information systems that were analyzed.

TABLE. I. COMPARISON OF SOFTWARE QUALITY MEASUREMENT MODELS

Model	McCall	Boehm	FURPS	Dromey	ISO 9126	ISO 25010
Author	Jim McCall	Barry W. Boehm	Hewlett Packard	R. Geoff Dromey	ISO	ISO
Year	1977	1978	1992	1995	2001	2011
Description	• bridge the gap between user and system developer • Consider users' view and developer priorities • Focus on accurate measurement of high-level characteristics • based on 3 perspectives – Product Revision, Product Operation and Product Transition	• define software quality through a set of qualitative characteristics and metrics • based on hierarchy arranged according to characteristic level – high, moderate and primitive	• represent abbreviation for *Functionality, Usability, Reliability, Performance and Supportability* • categorized into two (2) types of requirement – functional and non-functional	• based on product quality perspective • focus on relationship between software product characteristics and software quality attributes	• developed based on McCall and Boehm models • to align the evaluation of software or system product using ISO quality model • list of internal and external characteristics of a software product	• improvement to ISO 9126 Model • two (2) additional quality factors – compatibility and security
Measurement Factor	• Correctness • Reliability • Usability • Efficiency • Integrity • Maintenance • Testability • Flexibility • Portability • Reusability • Interoperability	• Portability • Reliability • Efficiency • Usability • Testability • Understandability • Flexibility	• Functionality • Usability • Reliability • Performance • *Supportability*	• Efficiency • Understandability • Reliability • Functionality • Process Maturity • Maintenance • Portability	• Functionality • Reliability • Usability • Efficiency • Maintenance • Portability	• Functionality • Reliability • Usability • Efficiency • Compatibility • Security • Maintenance • Portability

TABLE. II. ADVANTAGES AND DISADVANTAGES OF SOFTWARE QUALITY MEASUREMENT MODELS

No.	**Model**	**Advantage**	**Disadvantage**
1.	McCall	• Having evaluation criteria	• Overlapping of components • Software quality measured subjectively, as it is based on responses of Yes or No • The model does not consider the functionality so that the user's vision is diminished • No consensus about what high level quality factors are important • Each quality factor is positively influenced by a set of quality criteria, and the same quality criterion impacts several quality factors. If an effort is made to improve one quality factor, another quality factor may be degraded. • No standards, no methods and no tools to measure the quality factors.
2.	Boehm	• Including factors related to hardware • Easy to understand and learn	• Lack of criteria • Very difficult to apply in practice
3.	FURPS	• Separating functional and non-functional requirements • Can be used as both product requirements as well as in the assessment of product quality	• Not considering portability
4.	Dromey	• Applicable to different systems	• Incomprehensiveness
5.	ISO 9126	• Having evaluation criteria • Separating internal and external quality • Developed in agreement among all country members of ISO • Unify and quantify different views of quality requirements • Having a single universal model makes it easier to compare one product with another. • The characteristics defined are applicable to any kind of software while providing consistent terminology for software product quality. • covers all crucial characteristics such as hierarchical structure; criteria for evaluation; comprehensive expression and terms; simple and accurate definitions; and one to between various layers of model • widely used in the software engineering community and has been adapted to different domains and contexts • easy to use and understand by its users.	• The traceability of the software and the consistence of the data are not represented in the model • The model does not include measurements methods • There is no consensus regarding what is a top-level quality-factor and what is more concrete quality criterion
6.	ISO 25010	• the most recent and updated model • Improvement to ISO 9126 • Additional of security attribute	

TABLE. III. TYPES OF INFORMATION SYSTEM

No.	**Type of System**	**Author**	**Reference**
1.	Enterprise Information System	Hu and Wu (2016) Green and Robb (2014) Esaki (2013)	A B C
2.	Knowledge Management System	Wu and Wang (2006)	J
3.	e-Government System	Cohen and Eimicke (2003)	M
4.	e-Learning System	Hassanzadeh, Kanaani & Elahi (2012) Bhuasiri et al. (2012) A.K.M. Najmul Islam (2011) Padayachee et al. (2010) Lin (2007)	D E F G I
5.	Generic Information System	Gable et al. (2008) Hellstén and Markova (2006) Iivari (2005) Poon & Wagner (2001) Goodhue, Thompson & Goodhue (1995) Srivinasan (1985)	H K L N O P

TABLE. IV. SOFTWARE QUALITY MEASUREMENT FACTORS

No.	Quality Factor	Author																Σ
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	Easy to use		●	●	●	●		●			●	●						7
2	Response		●			●		●		●	●		●					6
3	Reliability	●	●	●	●	●											●	5
4	Easy to access				●							●	●		●			4
5	Functionality	●		●		●												3
6	User friendly				●						●				●			3
7	Fulfill user requirement				●				●			●						3
8	Interactive				●	●									●			3
9	Security		●		●													2
10	Maintenance			●	●													2
11	Integration				●								●					2
12	Flexibility				●								●					2
13	Structural design				●			●										2
14	Updated information							●	●									2
15	Recoverability							●					●					2
16	Learnability											●				●		2
17	Data and System Accuracy											●					●	2
18	Portability			●														1
19	Efficiency			●														1
20	System Stability										●							1
21	Data Integration													●				1
22	Aesthetic				●													1
23	Personalization				●													1
24	Attractive				●													1
25	System Speed				●													1
26	Internet quality					●												1
27	Memory space						●											1
28	System Integration							●										1
29	Standard Compliance							●										1
30	Fault tolerance							●										1
31	Parallel terms							●										1
32	Consistent terms							●										1
33	Understandability							●										1
34	Information arrangement							●										1
35	Access by many users							●										1
36	Accurate solution									●								1
37	Data integration								●									1
38	Usability											●						1
39	Language												●					1
40	Availability															●		1
41	Response to user															●		1
42	Solution alternative																●	1

Based on the analysis, it was found that various quality factors used differently according to the types and functions of an information system. Several factors are widely used by researchers such as easy to use, response and reliability. This may be because these three factors reflect the basic features required to ensure the quality of the system. Table IV shows the list of quality factors.

III. Methodology

In this study, a quantitative approach was used by conducting a survey to achieve its objectives and questions.

A. *Participants*

A total of 174 civil servants in Public Service Department (PSD), Putrajaya participated in this study. They comprise of 67 males (39%) and females (61%). They were divided into 2 categories of services, where 79 or 45% were in the Professional and Management category, and 95 (55%) were in the Support Services category.

B. *Instrument*

This study employs a self-developed 39-items questionnaire consisting of six constructs that represent quality factors, namely the Functionality, Reliability, Usability, Efficiency, Compatibility and Security (Table V). The scale is 5 points Likert-type, where participants are required to give their response on a Strongly Disagree‒ Disagree – Slightly Agree – Agree – Strongly Agree pattern.

As stated above, the constructs and items (quality factors) are gathered based on the analysis of previous studies related to software quality. To ensure that the constructs and items are valid and can be used to collect data effectively, the development of survey is very important. It must be done systematically to ensure it fulfils the study objectives. After identifying constructs and items, a suitable scale is selected and the testing of item validity and instrument reliability are performed. Expert view is obtained and content validation is also done since they are also important elements in making sure the instrument is relevant.

C. *Data Analysis*

Rasch Model is used to analyze data from the respondents. The model refers to an idea, principal, guideline or technique that enables measurement of the latent trait [6]. It basically separates individual capabilities and instrument's quality. This model assumes that individual response towards an item is only influenced by individual capabilities and item difficulties [7]. The ability of the Rasch Model as an analytical instrument is proved by its application in various research areas including management and social science. This model prevents researchers in social sciences area from making a raw and blurred observation and undertakes definitive actions with realistic accuracy and clear quality control [8]. In this study, the WinStep software is used to perform the Rasch analysis.

TABLE. V. Description of Constructs

Construct	Description	No. of Items
Functionality	Software capability to provide functions that fulfil the user requirement of human resource information system.	8
Reliability	Software capability to maintain the performance level for a period to support the human resource management.	7
Usability	Software capability to be understood, learned and used to implement the human resource management.	13
Efficiency	Software capability to produce desired performance in assisting user to perform human resource management functions effectively.	7
Compatibility	Software capability to ensure efficient performance while sharing common environment and resources and/or exchange information with other products that perform the same functions.	2
Security	Software capability to ensure secure transactions while performing human resource management functions effectively.	2
Total		39

IV. Findings

This study considers main assumptions of the Rasch Model for selection of items such as item fit, unidimensionality and local independence. It also considers other aspects such as reliability index, separation index, and the respondent-item map.

A. *Item Fit*

Two criteria are used to measure the good fit-ness of items, namely the *Outfit MNSQ* = y, where 0.5 < y < 1.5 and *Outfit* Z-*standard* (*Zstd*) = z, where -2 < z < 2 for acceptance of items [9]. Although the analysis shows that 6 items do not fulfil the criteria, they are considered as important items. Thus, these items are reviewed and modified to suit the measurement objective.

B. *Unidimensionality*

The raw variance explained by measures is 53.8 percent, which is close to the expected model value of 57.5 percent. This exceeds the minimum value required of 40 percent. The unexplained variance in the first contrast is 7.5 percent, less than the maximum value of 15 percent. This shows that the instrument can measure in one standard dimension and thus is able to achieve its measurement objective.

C. *Local Independence*

Ten pairs of items with the largest standardized residual correlations values have been identified. One pair of items, A1 – *Provides accurate human resource information* and A2 – *Provides updated human resource information* has a value of 0.72, indicating these items are overlapped. However, since these items measure different characteristics (accurate and updated information) and both characteristics are important, both are retained in the actual questionnaire.

D. Reliability Index

The Cronbach Alpha index is valued at 0.96. This shows that the instrument is highly reliable. The respondent reliability index is 0.93, and item reliability index is 0.98. This result indicates that there is enough sample and the instrument is suitable to measure the respondents' capabilities and item difficulties.

E. Separation Index

The respondent separation index is 3.67, and the item separation index is 7.70. Index value between three and four indicates good value while the value more than five indicates excellent value [6]. This shows that the items can separate the respondents based on their capabilities and items based on the difficulties.

F. Respondent – Item Distribution Map

The respondent-item map shows the distribution of items based on difficulties with the distribution of respondents' capabilities along the logits scale (see Fig. 1).

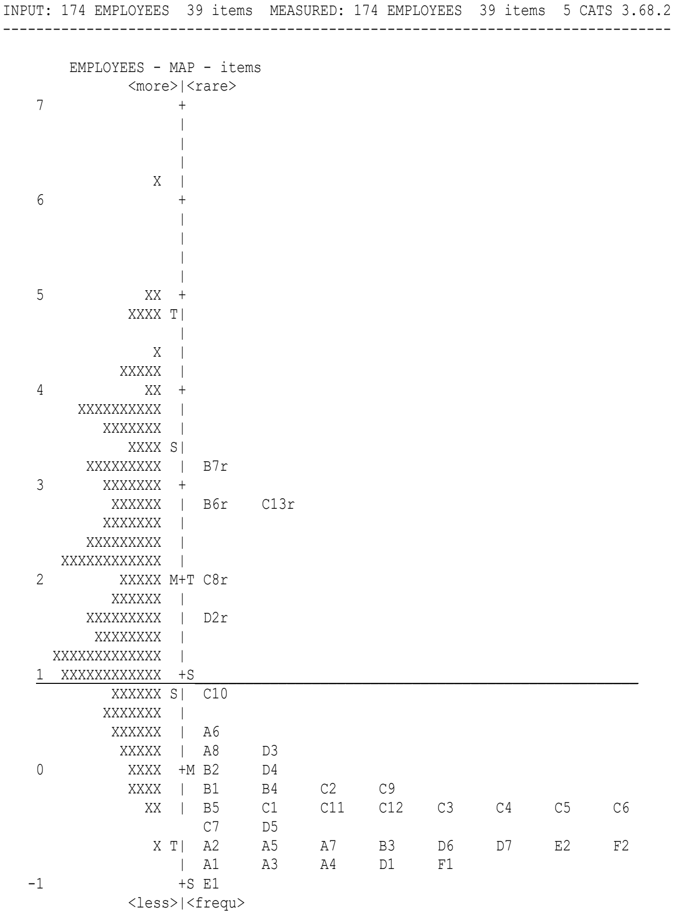
Overall, most respondents are above the $\text{Mean}_{\text{item}}$ = 0.00 logits. This indicates that almost all of them can answer the questionnaire. There are also a larger number of respondents above the $\text{Mean}_{\text{respondent}}$ = 2.08 logits. It also indicates that the respondents are competent enough to determine the criteria influencing the software quality of an information system.

The difficulty measurement value ranges between +3.12 logits and -0.96 logits. The item distribution shows item B7r - *Software problem affects the system performance* as the most difficult to be agreed item while item E1 – *All data must be integrated with each other* as the easiest item to be agreed.

Based on the analysis, there were 30 items that are under the $\text{Mean}_{\text{item}}$ = 0.00, indicating the number of criteria that contribute significantly to the software quality of the human resource information system. Table VI shows the significant items.

The gap in the item distribution map is also examined to determine whethe.r the items are enough to evaluate the capabilities of all respondents. The result shows there are gaps between items B6r (2.75 logits) and C8r (2.10 logits) and D2r (1.56 logits) and C10 (0.75 logits). These gaps indicate that there are not enough items to measure higher level of respondents' capabilities. This is also highlighted by the respondents and item distribution above the line, where the number of items is comparatively smaller compared to the number of employees.

Thus, further study is required to develop more difficult items that can measure respondents with higher capabilities.



```
INPUT: 174 EMPLOYEES  39 items  MEASURED: 174 EMPLOYEES  39 items  5 CATS 3.68.2
--------------------------------------------------------------------------------


    EMPLOYEES - MAP - items
           <more>|<rare>
 7               +
                 |
                 |
                 |
              X  |
 6               +
                 |
                 |
                 |
                 |
 5           XX  +
           XXXX T|
                 |
              X  |
          XXXXX  |
 4           XX  +
     XXXXXXXXXX  |
        XXXXXXX  |
           XXXX S|
      XXXXXXXXX  |  B7r
 3      XXXXXXX  +
         XXXXXX  |  B6r     C13r
        XXXXXXX  |
      XXXXXXXXX  |
   XXXXXXXXXXXX  |
 2        XXXXX M+T C8r
         XXXXXX  |
      XXXXXXXXX  |  D2r
       XXXXXXXX  |
  XXXXXXXXXXXXX  |
 1 XXXXXXXXXXXX  +S
         XXXXXX S|  C10
        XXXXXXX  |
         XXXXXX  |  A6
          XXXXX  |  A8      D3
 0         XXXX  +M B2      D4
           XXXX  |  B1      B4      C2      C9
             XX  |  B5      C1      C11     C12     C3      C4      C5      C6
                    C7      D5
              X T|  A2      A5      A7      B3      D6      D7      E2      F2
                 |  A1      A3      A4      D1      F1
-1               +S E1
           <less>|<frequ>
```

Fig. 1. Respondent-Item Distribution Map.

TABLE. VI. SIGNIFICANT ITEMS FOR EACH QUALITY FACTOR

Quality Factors	Significant Items
Functionality	A1, A2, A3, A4, A5, A7
Reliability	B1, B2, B3, B4, B5
Usability	C1-C7, C9, C11, C12
Efficiency	D1, D4, D5, D6, D7
Compatibility	E1, E2
Security	F1, F2

V. DISCUSSION

This section will discuss the findings from the respondent and item point of view.

A. Respondent

Overall, most respondents are placed above the $\text{Mean}_{\text{Item}}$ value. This shows that almost all respondents can answer the questionnaire to determine the criteria for information system capabilities. Number of respondents above the $\text{Mean}_{\text{Respondent}}$ are also bigger than that of below the mean value. This shows that the respondents possess experience in utilizing information system in their work thus are aware of the criteria required to ensure good software quality in information system.

Apart from that, there is also a group of respondents that can be considered the most competent. They are the experts and possess huge experience in information system development. Based on their experience, they agreed that all software quality criteria are important and should be considered when developing an information system.

B. Item Distribution

Based on Fig. 1, the item distribution map shows that item B7r – The system performance affected when there is software problem is the least agreeable item. This indicates the respondents' view that if the software quality is good, the information system will not be affected and can maintain its performance although there might be problem with the software. Meanwhile, the item E1 – All data in the system must be integrated is the most agreeable one, indicating that the respondents agreed the criteria is very important in ensuring the best software quality in information system.

Besides, there are 30 items below the $\text{Mean}_{\text{Item}}$ value, placed under the category of most agreeable items. It shows that these criteria are widely used dan emphasized in determining software quality in an information system. It also shows that the respondents agreed these items significantly contribute to the information system software quality.

For Construct A (Functionality), there are six items under the category of most agreeable. This shows that these items are deemed very important and significant in ensuring the best software quality in information system. Two (2) items, A6 – Capable to perform prediction towards a certain scenario and A8 – Capable to perform what-if analysis on various scenarios are placed under the moderate category. This shows that both criteria are becoming more important today and system developers should consider these aspects when developing or modifying an information system to further enhance its capabilities in assisting stakeholders to ensure accurate and comprehensive decisions are made. This is in line with the study done by [10] that discussed and suggested the framework for intelligent human resource information system.

As for Construct B (Reliability), five items are under the most agreeable category, indicating that these items are very important in determining the software quality. Respondents also agreed that item B2 – Suggest accurate solution based on user requirement is becoming more important in assisting the decision-making process. The least agreeable items are item B6r – Information visualization in certain formats only and B7r - The system performance affected when there is software problem. It shows that the respondents agreed an excellent information system should be capable to visualize information in various formats, depending on the alternative and requirement of users. This is because there is no one format that can fulfil different requirements of the users comprehensively. This is in line with the information technology advancement, where Business Intelligence (BI) is considered when developing an information system. Apart from that, an excellent system should be capable of recovering quickly from any software problem and maintain its performance. In an era where time is essence, decision making must be made quickly thus it is very critical to have an information system that is always stable and maintain the best performance. In the context of software reliability, different models are needed to evaluate the software reliability at different levels of development [11]. This is because software reliability is a very critical factor and should be evaluated thoroughly. Reliability issues do not only affect the system performance but can also cause complete failure of the system.

Under Construct C (Usability), ten items are placed in the most agreeable category. This shows that these criteria are very important and widely used in determining information system software quality. This finding is in line with the study by [12] which states that the usability factor is very important in ensuring the efficiency and effectiveness of health information system. However, two (2) items are the least agreeable. One of them is item C8r – User spends a long time in learning to use the system, indicating that the respondents do not agree if users require a long time to learn how to use a system. Another item is C13r – User needs ICT knowledge to use the system. With the advancement in ICT resulting in the development of more intelligent information system, the system should be easier to use by anybody, including those without detailed technical and ICT knowledge.

For Construct D (Efficiency), six items are in the agreeable category. This indicates that the respondents agreed with most items under this construct. It is very important to have an efficient system that provides fast service and has information presentation design that is understood by the users. An efficient system should also be capable to maintain its performance while being accessed by many users at the same time. For example, in public sector there are information systems that are developed for the benefit of the whole civil servants. In this construct, only item D2r – Takes a long time to process users' request such as generating report or analyzing information that is the least agreeable. This is because an information system is designed to help users doing their work. Therefore, they need a system that would be able to give response to a request and process it within a short period of time.

As for Construct E (Compatibility), both items are most agreed by respondent. This shows that both items are very important in the development of information system. The capability of data and system integration with different environment or system is among the basic characteristics of a good and efficient system and can ensure real time accurate information [13]. Information system integration is also very important for the success of e-Government [14] For example, integration of several systems in different agencies will enable business license applications to be processed and approved in a short time.

Lastly for Construct F (Security), both items F1 and F2 are also in the most agreeable category. This shows that the security aspect is very important in an information system. Nowadays, the introduction and application of Internet of Things (IoT) in intelligent devices enables the connectivity of the devices to the Internet [15]. But this has also exposed the devices to security risks such as information leakages or theft. So, it is very important to ensure an information system is secure and prevent system or process failure.

VI. Conclusion

Overall, this study has successfully identified the factors and criteria influencing the software quality of information system. By identifying these criteria, it will serve as a guideline to improve the software aspect of an information system and ensure that the system has the capability and quality at its best.

However, technological advancements today has brought about the needs for intelligent information systems that are able to perform more complex functions. Thus, there is a need for future research to look into other software quality factors and criteria that will give advanced capabilities to the system.

System developers also play an important role in ensuring that an information system is able to perform as required. Their inpu is equally important so as to enable the software quality to be assessed comprehensively. Therefore, it is also recommended for future research to take into account the perspective of system developers.

Information systems are being used in various fields to assist in decision making and strategic planning. Each field requires different capabilities and technical requirements. So it is also recommended that future works should look into different types of information system. This will further contribute to the improvement of the system and software quality.

Acknowledgment

This paper was funded by Fundamentals Research Grant Scheme (FRGS), Ministry of Education Malaysia under project code FRGS/1/2018/ICT04/UKM/02/11.

References

[1] Oprea, M., "MAS_UP-UCT: A multi-agent system for university course timetable scheduling," International Journal of Computers, Communications & Control, 1, 2007, 94–102. doi:10.1590/S1415-65552003000100014.

[2] Masrek, M. N., " Reinventing Public Service Delivery : The Case of Public Information Systems Implementation in Malaysia," International Journal of Public Information Systems, 1, 2009, 9–15.

[3] Poon, P. and Wagner, C., "Critical success factors revisited: Success and failure cases of information systems for senior executives," Decision Support Systems, 30(4), 2001, 393–418. doi:10.1016/S0167-9236(00)00069-5

[4] Elkadi, H., "Success and failure factors for e-government projects: A case from Egypt," Egyptian Informatics Journal, 14(2), 2013, 165–173. doi:10.1016/j.eij.2013.06.002

[5] El-far, I. K. and Whittaker, J. A., "Model-based Software Testing," 1–22, 2001.

[6] Azrilah Abdul Aziz, Mohd Saidfudin Masodi, and Azami Zaharim, "Asas Model Pengukuran Rasch," Penerbit Universiti Kebangsaan Malaysia. 2013.

[7] Bond, T. and Fox, C. M., "Applying the Rasch Model Fundamental Measurement in the Human Science," hlm.Second Edi. 2007.

[8] Wright, B. D. and Mok, M. M. C., "An Overview of the Family of Rasch Measurement Models. Introduction to Rasch Measurement," 2004, 1–24.

[9] Nopiah, Z. M., Rosli, S., Baharin, M. N., Othman, H., and Ismail, A., "Evaluation of pre-assessment method on improving student's performance in complex analysis course," Asian Social Science, 8(16), 2012, 134–139. doi:10.5539/ass.v8n16p134

[10] Masum, A.-K., Beh, L.-S., Azad, A.-K., and Hoque, K., "Intelligent human resource information system (i-HRIS): A holistic decision support framework for HR excellence," International Arab Journal of Information Technology, 15(1), 2018, 121–130.

[11] Anila, M., Sirisha, N., and Karthik, R., "Software reliability models - A comparative study. Proceedings of the International Conference on Intelligent Sustainable Systems," ICISS 2017, (Iciss), 2018, 1152–1154.

[12] Alshamari, M., "Usability Factors Assessment in Health Information System," Intelligent Information Management, 08(06), 2016, 170–180.

[13] Wiemann, S., Brauner, J., Karrasch, P., Henzen, D., and Bernard, L., "Design and prototype of an interoperable online air quality information system," Environmental Modelling and Software, 79, 2016, 354–366.

[14] Siti Istianah Mahdzur; and Juhana Salim, "Information Systems Integration Factors In Organization: Towards Government Information Systems Sustainability," Journal of Theoretical and Applied Information Technology, 71(2), 2015, 235–250.

[15] Kumar, S. A.; Vealey, T.; and Srivastava, H., "Security in internet of things: Challenges, solutions and future directions," Proceedings of the Annual Hawaii International Conference on System Sciences, 2016–March (January), 2016, 5772–5781.

Chemical Reaction Optimization Algorithm to Find Maximum Independent Set in a Graph

Mohammad A. Asmaran[1], Ahmad A. Sharieh[2], Basel A. Mahafzah[3]
Department of Computer Science, The University of Jordan, Amman, Jordan

***Abstract*—Finding maximum independent set (MIS) in a graph is considered one of the fundamental problems in the computer science field, where it can be used to provide solutions for various real life applications. For example, it can be used to provide solutions in scheduling and prioritization problems. Unfortunately, this problem is one of the NP-problems of computer science, which limit its usage in providing solution for such problems with large sizes. This leads the scientists to find a way to provide solutions of such problems using fast algorithms to provide some near optimal solutions. One of the techniques used to provide solutions is to use metaheuristic algorithms. In this paper, a metaheuristic algorithm based on Chemical Reaction Optimization (CRO) is applied with various techniques to find MIS for application represented by a graph. The suggested CRO algorithm achieves accuracy percentages that reach 100% in some cases. This variation depends on the overall structure of the graph along with the picked parameters and colliding molecule selection criteria during the reaction operations of the CRO algorithm.**

Keywords—Chemical reaction optimization; graph; maximum independent set; metaheuristic algorithm; modified Wilf algorithm; optimization problems

I. Introduction

In this paper, a metaheuristic Chemical Reaction Optimization (CRO) algorithm has been utilized to find out maximum independent set (MIS) in a graph. In this approach, computational steps are formulated as a set of molecules reactions that leads toward approximated solution. CRO approach considers two types of collisions that could happen: On-Wall collision and Inter-molecular collision as illustrated in [1]. These collisions could be effective or ineffective depending on the nature and the type of the problem to be implemented or solved. The effective On-Wall collision is called decomposition, where the colliding molecule is supposed to be decomposed into several parts (mainly two parts). Effective inter-molecular collision is called synthesis, which involves merging the colliding molecules together.

In [2], Independent Sets (ISs) are described to be some of the useful information that can be concluded from graphs and used in real life applications; such as project scheduling and social network analysis, while they are important concept in building bipartite graphs [3,4] which are fundamental in many computing areas; such as coding theory and projective geometry. Independent set can be defined as a set of nodes in a graph that are not connected. Note that a graph may contain several independent sets, and finding the maximum one is the best goal to achieve. The IS with maximum size is referred to as MIS.

Finding an MIS in a graph is considered very useful approach for many real life applications and problems, such as optimization problems, job scheduling, and social network analysis. MIS can be determined using brute force approach in O($N^2 \times 2^N$) run time units, where N is the number of vertices in a graph. This approach takes a lot of time to find an MIS for large N as described in [5]. Many approaches and algorithms are proposed to find out an MIS of a graph, but with very long run time. So, many algorithms have been proposed to find out an approximation to actual exact MIS solution with less time complexity as in [6,7,8,2,9].

Here are some definitions related to MIS and CRO:

- An undirected graph is G(V, E), where V is a set of vertices and E is a set of edges in G. The set of vertices is a collection or group that contains the vertices (nodes) in the graph and these vertices (nodes) are connected to each other by links that are called Edges. The collection or group that contains all graph edges is called Edge Set noted by E.
- An Independent Set (IS) in a graph G(V,E) is defined in [10,11] to be a set V’, where V' ⊆ V and there is not exist an edge that connects v_s and v_e, where e ∈ E, v_s∈ V' and v_e∈ V' (i.e. either v_s or v_e∈ V');Where a Maximum Independent Set (MIS) is defined to be the IS of the largest size among all available ISs in G.
- Chemical Reaction Optimization (CRO) is defined in [1,15] as a metaheuristic approach that mimics the process of chemical reactions in the field of Computer Science. It relays on minimizing the potential energy to the minimal value without sticking in local minima. This algorithm defines an objective function that is used to calculate potential energy of the current state of reaction (execution) process. Just like genetic algorithms, this is done by iterating for a predefined number of iterations or meeting optimal objective value.
- On-wall ineffective collision is a CRO operation that involves colliding the molecule on the wall without any effective restructure of the colliding molecule.
- Decomposition (On-wall effective collision), is a CRO operation that involves colliding the molecule on the wall effectively so that colliding molecule is decomposed (divided) into multiple molecules.
- Inter-molecular ineffective collision is a CRO operation that involves colliding two molecules

together ineffectively so that no major structural change would occur.

- Synthesis (Inter-molecular effective collision) is a CRO operation that involves colliding two molecules together effectively so that a new molecule of the merged collided molecules will be generated.

The solution for a problem based on CRO is represented as a molecular structure noted as (ω), which has a minimum potential energy that is determined by a problem specific objective function noted as PE_{ω}, which is determined by an objective function $f(\omega)$. Each molecule has a kinetic energy that illustrates the tolerance of having worse solutions and noted as *KE*.

In this paper, different techniques are implemented over CRO algorithm to provide an approximate solution for the MIS problem. In these techniques, an implementation of CRO is provided to solve MIS problem to provide near optimal solution.

In the remaining sections, a review of related work is presented in Section 2. A description of the proposed algorithms will be explained in Section 3. This is followed by experimental results in Section 4 and discussion in Section 5. Section 6 presents the conclusion and intended future research.

In this paper, a new approach is applied to find Maximum Independent Set to explore its ability in order to find better approximation results than previous approaches that cannot be applied on huge graphs which may contain millions of nodes. Finding a maximum independent set with near optimal results would be used to provide a solution of many real-life applications; such as prioritization and scheduling applications.

II. Related Work

In their research for finding solution of the MIS problem, researchers have handled the issue using different approaches based on type of final result or the nature of graph, such as degree of nodes, as illustrated in [10,11,12] where such approaches have been used. In general, finding MIS can be done using one of the following three approaches: using brute force algorithm, approximation algorithms, and exact algorithms for special type of graphs.

The first approach is using an exact (brute force) algorithm. The direct way to solve such problem is to check all possible solutions by representing the presence of node in the solution by 1 and the absence by 0 as mentioned in [11]. So, we can represent the solution by a binary number with length N, where N is the number of nodes in the given graph. This involves checking 2^N numbers that represent all possible subsets of the original set of nodes. For each solution (binary number), all nodes must be checked to ensure disconnection of nodes (N^2). So, final run time complexity would be $O(N^2 \times 2^N)$. Nevertheless, some researchers have produced exact algorithms with better runtime. In [14], the authors proposed an algorithm, which achieved an exact solution in $O(1.2132^n)$ time for a graph of size n vertices, while in [15], the authors provided an algorithm with running time complexity of $O(1.2114^n)$ to find an exact solution. Such a reduction in the run time could have high influence in case of critical run time applications like process scheduling on a CPU.

The second approach is using approximation algorithms based on heuristics to provide approximate solution in polynomial-time. According to [12], “Most polynomial-space algorithms for MIS use the following simple idea to search a solution: branch on a vertex of maximum degree by either excluding it from the solution set, or including it to the solution set. In the first branch, we will delete the vertex from the graph and in the second branch we will delete the vertex together with all its neighbors from the graph”. Algorithms that use such heuristics can be found in [7,16]. More evolutionary heuristic approaches can be found. For example, in [6,17,18,19], genetic algorithm was used to find an approximate solution for the MIS problem. In [20], a swarm intelligence approach based on ant-colony optimization was used to find a solution. Note that, approximation algorithms are used in real applications just like in [21], where genetic algorithm is used to generate data for testing PLSQL (Procedural Language extension to Structured Query Language) program units. This generated data is a sub-set of the actual data range that can’t be covered in some extreme cases, where data to be tested is huge and can’t be tested using normal brute-force concept. A more generic test data generation for software testing is proposed in [22] to generate test data using genetic algorithm for software testing purposes rather than using normal brute-force test data generation.

The third approach is using exact algorithms to find exact solution in polynomial-time, but for graphs of special classes, such as designing a polynomial run time algorithm that finds an exact solution in graphs with vertex of degree 2 at maximum. Such algorithms are case sensitive ones and can’t be generalized to find exact solutions to graph of random shape and arbitrary degree. Examples of this form of algorithms can be found in [23], where an exact algorithm is provided for graphs with vertices of maximum degree of 3, or in [12,24], where, in addition to an exact solution provided for any random graphs, the authors provided a $O(1.1571^N)$, $1.1737^N \times N^{O(1)}$, $1.1893^N \times N^{O(1)}$, and $1.1970^N \times N^{O(1)}$, for graphs of maximum degree of 4, 5, 6, and 7, respectively.

As mentioned before, all exact solutions attempts consume a very large amount of time to execute. Such algorithms would decrease the feasibility of e solutions. So, a new paradigm of computing near optimal solution has been proposed, such as in [7,8,9,16]. As illustrated in [8], this is done using heuristic or metaheuristic techniques. Combining of exact and meta-heuristic algorithms can provide near optimal solution in a shorter time like in [25] where better execution time has been achieved. Moreover, there are some known strategies to do parallel implementation of metaheuristic approaches. By parallelizing these algorithms such as in [26], an enhanced version with better performance could be achieved.

In [27], CRO has been used to find optimal solution for task scheduling and resource allocation in grid computing. They propose several versions of CRO to solve task scheduling problem. These versions have been experimented and tested against four metaheuristic approaches. The results show that CRO outperforms the other approaches in terms of accuracy and performance, especially in case of large test instances.

In [1], CRO has been used to provide a solution for quadratic assignment problem described in [28]. CRO implementation has been tested against various evolutionary approaches. Test results show that CRO implementation outperforms other implementations in many cases. Parallel implementation of CRO has been used to solve the same problem in [29], where test results show that parallel CRO implementation provides better performance along with solution quality in comparison with sequential one.

In [1], CRO algorithm has been used to solve resource-constrained project scheduling problem described in [30] as planning the project milestones according to predefined priorities. In real life, project is divided into fixed time slots. Project activities are assigned to time slots according to available resources that are limited, while activities could be dependent on each other. CRO is used to find best scheduling of tasks that minimizes project lifetime. Test results show that CRO implementation can achieve better results for known benchmarks.

In [1], CRO has been used to provide a solution for channel assignment problem in wireless mesh networks described in [31] to assign available channels to multiple wireless networks. It is used for wireless communication channel selection to be used in the communication between neighboring mesh routers without suffering any interference or communication problems. The results show that CRO has improved current solutions of the problem.

In [32], CRO has been used to solve population transition problem in peer-to-peer live streaming. In this problem, network live streaming has been improved by grouping peers into multiple colonies according to delay. Peers with less delay can act as service providers for longer delay ones. So, the system is said to be in universal streaming when all peers are served with sufficient streaming data. Test results show that evolutionary approach of CRO outperforms existing non-evolutionary approaches.

In [33], CRO has been used to find a solution for network coding optimization problem described in [34] to provide coding mechanism for network with minimum number of digits. In this problem, network coding has been used to enhance transmission rate between routers on certain interfaces. This strategy of coding specific interfaces could increase transmission rate without avoidance of extra computational overhead by coding all available interfaces. Test results show that CRO outperformed existing algorithms.

In [35], CRO has been used in Artificial Neural Network (ANN) training. ANN is composed of layers that contain multiple computational units called neurons. Neurons must be assigned weights to provide best results. Tuning is done by training the network with set of training data. Test results show that CRO trained ANN has better testing error.

In [36], CRO has been used to solve Set Covering Problem (SCP) while in [37] a strengthened version of clique covering has been investigated. SCP can be formulated as the following:

- Given a set M, $M_j \subseteq M$, j = 1,...,n are n subsets of M, and weights of the subsets, c_i, j = 1,...,n; and set cover is a collection $T \subseteq \{1,...,n\}$ such that $\bigcup_{j \in T} M_j$ = M. SCP tries to minimize the cost of covering the entire set using a subset of the original set. There are two types of set covering problem, unicost, and non-unicost. CRO outperformed the accuracy of other algorithms in case of non-unicost SCP, where optimal solution has been determined in 65 experiments. In case of unicost SCP, CRO shows outstanding performance in comparison to other approaches.

In [38], a version of CRO called Greedy CRO (CROG) has been proposed and implemented to solve 0-1 Knapsack Problem. Experimental results show that CROG outperforms other metaheuristic approaches, such as genetic algorithms, ant-colony, and quantum-inspired evolutionary algorithms.

In [39], enhanced version of CRO has been used to find optimal road network design that takes into consideration the cost along with noise and vehicles emissions. Proposed CRO was tested against Genetic Algorithm (GA) for comparison. Test results show that CRO outperformed GA in most cases.

In [40], Objective Power Flow (OPF) problem has been solved using CRO algorithm. OPF aims to minimize power generation cost by considering many constraints, such as the balance of the power, bus voltage magnitude limits, transmission line flow limits, and transformer tap settings. The results show that CRO can provide the best results among other algorithms on the IEEE-30 test case. Note that best result is the one with lowest power flow cost.

In [27], CRO implementation has been extended using parallel approach to solve the Quadratic Assignment Problem (QAP). QAP seeks to optimally assign facilities to locations in a way to minimize transportation cost of facilities, as they are required in multiple locations. Parallel CRO has been compared with sequential one in solving QAP, experimental results show that parallel CRO reduces computational time with more accurate results.

In [41,42], CRO implementation has been done to solve Max Flow problem (MFP) in a way that is close to Ford-Fulkerson algorithm. In [42], the results have been compared with GA in term of accuracy and performance. The results show that the problem is solvable by CRO and GA; however, the GA one outperforms the CRO one.

In this research, we provide adapted versions of CRO to find a solution of the MIS problem. Several scenarios are investigated when a molecule (subset of the graph) is selected randomly among available molecules, and a molecule is selected according to certain criteria. The selected criteria are the minimum connectivity. The adapted CRO algorithm with its implementation and performance are presented.

III. CRO Algorithm for MIS

In CRO, a molecule is represented by a node in a graph. Thus, an MIS has set of not connected nodes, or set of none neighboring nodes. In such representation, the CRO considers each molecule a candidate solution (i.e. Independent Set). Molecule potential energy is defined as the number of remaining graph nodes that are not contained in the molecule. So, if the number of graph nodes is 50 and the molecule contains 5 nodes, the potential energy is 50-5=45. Fig. 1 shows the flowchart of the CRO algorithm.

Initially, there are *N* molecules manipulated by the algorithm, since every node is considered as one molecule, which is the minimum solution (each node is an independent set). A molecule is selected for the purpose of collision in each iteration. Collision type is selected according to the initial inter-molecular to on-wall collisions ratio.

In case of inter-molecular collision, effectiveness of the collision depends on whether selected molecules can be merged together or not. This is done by checking the confliction between the two molecules, so that each node in the second molecule is checked with the conflicting (i.e. neighbors) nodes of the first molecule. If the node is found among the conflicting nodes of the first molecule, the collision is defined to be none effective collision and nothing would happen because the two molecules are not eligible to be merged. This is because each molecule is assumed to be an independent set, and it is not allowed to contain conflicting nodes. On the other hand, if the entire nodes of the second molecule are not exist among the conflicting nodes of the first molecule, the collision is defined to be effective, so that selected molecules are merged together and new molecule is formulated. This new molecule contains the whole nodes of the collided molecules.

Table I shows the mapping of chemical notations to their corresponding mathematical representation defined in [1,13]. The solution is represented by a molecular structure noted as (ω).

Fig. 1 was adapted from [1,13], shows a flowchart of the CRO algorithm, which indicates that the first step of the algorithm is the initialization, as described in [1]. Initialization includes pre-processing (e.g. preparing the data in appropriate data structure, and removing unnecessary data), and initial values calculations (e.g. algorithm variables and constants). This step is followed by the iteration checking condition, which examines stopping criteria condition to avoid infinite calculations or iterations. If stopping criteria condition is met, the algorithm execution is finished, and no more iteration is done. On the other hand, if the condition is not satisfied, no more iteration is done. In each iteration, a collision must be performed, which could be either on-wall or inter-molecular collision. This involves determination of which action to be taken in the next iteration. If the collision type is selected, the next step is to decide whether the selected collision type is effective or ineffective according to the selected collision molecules. In case of intermolecular collision, effective collision is called synthesis, which indicates that collided molecules should be merged. In case of on-wall collision, effective collision is called decomposition, which indicates that collided molecule should be decomposed into two molecules. Regardless of collision type or its effectiveness, all affected molecules potential energy should be calculated and checked with previously registered minimum value of the molecules.

In case of on-wall collision, effectiveness of the collision depends on how many times a molecule collision did happen without any improvement in the solution. So, if a predefined number of iterations are reached without any improvement in its minimum value, the collision is defined to be an effective on-wall collision. In this case, the molecule is divided into two molecules, where each molecule contains the same number of original molecule's nodes. For example, if the collided molecule contains nodes {1, 10, 19, 50}, this molecule will be divided into two molecules one molecule contains {1, 10}, while the other one contains {19, 50}.

Another main factor of the proposed algorithm is molecule selection, which indicates to how a molecule is selected for further processing, such as on-wall collision or inter-molecular collision. In this proposed algorithm, multiple scenarios are tested, as the following:

1) A molecule is selected randomly among available molecules.

2) A molecule is selected according to certain criteria. The selected criteria are the minimum connectivity.

TABLE. I. Mapping Chemical Reaction to Mathematical Meaning

Chemical Meaning	**Mathematical Meaning**	**Mathematical Representation**
Molecular structure	Solution	Ω (e.g. MIS)
Potential energy	Objective function value	$PE_\omega = f(\omega)$ (e.g. number of remaining nodes in a graph that are not selected as in the solution)
Kinetic energy	Measure of tolerance of having worse solutions	KE_ω(e.g. the same value determined by the original algorithm)
Number of hits	Current total number of moves	(e.g. number of iterations)
Minimum structure	Current optimal solution	(e.g. the best solution found during the execution of the algorithm)
Minimum value	Current optimal function value	(e.g. the potential energy of the minimum structure)
Minimum hit number	Number of moves when the current optimal solution is found	(e.g. number of iterations "hits" till finding the minimum structure)

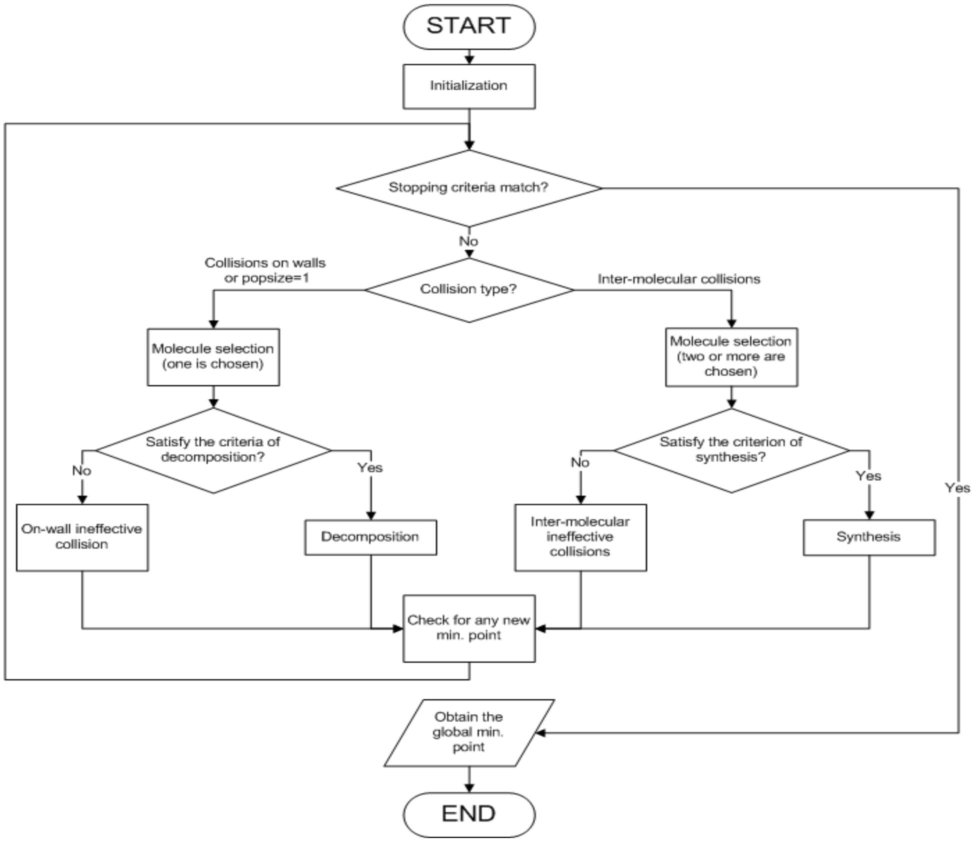


Fig. 1. A General Flowchart of the CRO Algorithm.

According to the above criteria of molecule selection, multiple combinations are tested to find out whether things go better or not, as follows:

1) Multiple random molecules: In this scenario, molecules are selected randomly for collision. In each iteration, a random molecule is selected for collision with another random selected molecule, or to collide with wall.

2) Single random molecule with random molecules: In this scenario, a random molecule is selected as a main molecule. In each iteration, this molecule is selected as the main molecule. In case of inter-molecular collision is performed, the second molecule is selected randomly. So, in this scenario, all iterations are done on the same molecule, but the variation appears in the second molecule only.

3) Single random molecule with minimum degree molecules: This scenario appears to be the same as the previous one, where a single random starting molecule is selected and used for every iteration in the reaction life cycle. But, the variation is that second molecule in case of inter-molecular collision is selected according to the criteria that is not random. Instead of that, the second molecule is selected according to its connectivity degree, where minimum connectivity degree molecule is selected to collide with fixed starting random molecule.

4) Single minimum molecule with random degree molecules: In this scenario, minimum connectivity degree molecule is selected at the beginning and used for every iteration. In case of inter-molecular collision iteration, the second molecule is selected randomly.

5) Single minimum molecule with minimum degree molecules: In this scenario, the same behavior of the previous scenario (4) is done with a difference that the second molecule in case of inter-molecular collision is selected according to its connectivity degree, so that minimum connectivity degree is selected to collide with initial minimum connectivity degree molecule.

If the collision is defined to be an effective inter-molecular collision, the components of the molecule are merged together and the conflicting nodes are computed with redundant nodes removal (no redundancy in conflicting nodes). The old molecules are removed from the pool of available molecules, while the resultant molecule is added to the pool.

In an iteration, potential energy is updated according to equation (1).

$$f(\omega) = N - \mathrm{Size}(\omega) \quad (1)$$

Where ω denotes a molecule, Size(ω) denotes number of nodes in a molecule, and N denotes the number of nodes in the graph.

Kinetic energy doesn’t affect the process of CRO in this proposed algorithm, since each molecule is assumed to be effective and capable of reacting with other molecules at any moment, regardless of its situation or kinetic energy.

A. CRO Algorithm

CRO algorithm relies on two major operations. These operations determine the way of finding the final solution. These operations are on-wall collision and inter-molecular collision. On-wall collision divides the molecule into two equally size molecules in case if it is effective. Otherwise (i.e. ineffective), the original molecule remains without any change. This operation is presented in Fig. 2. The other operation is to collide with another molecule (i.e. inter-molecular collision). If the collision is effective, the molecules are merged together (synthesized); otherwise, nothing happens as illustrated in Fig. 3. A function called next is called to determine whether to continue in executing next iteration or not, as it is illustrated in Fig. 4. Choosing collision molecules is done by one of two functions "chooseMinimumConflecting Molecule" or "chooseRandom Molecule", which is called once in case of on-wall collision and twice in case of inter-molecular collision to select necessary molecule(s) for collision. There are two implementations for this function, because there are two cases for the second selected molecule that are random or minimum connected molecules; as it is in Fig. 5. There are two cases for passing parameters to both functions: null value or non-null value. In case of null value, it means that this is the first selected molecule in the iteration regardless of collision type.

First selected molecule is fixed by selecting the same node all the time. So, the algorithm seeks the molecules looking for the molecule containing the initially selected node at the beginning of the algorithm execution. In case of non-null value, the algorithm has two cases: the second molecule is selected randomly or the molecule with minimum neighbors is selected.

In the code in Fig. 5, random molecule is selected. In the code in Fig. 6, the minimum neighbor molecule is selected.

At the beginning of execution, initror initmc function is called, where it is responsible of initializing molecules pool by adding created initial molecules to it and selecting the base molecule for reactions, as in Fig. 7 and Fig. 8.

Name: collideOnWall
Input: Molecule object that should collide on wall, and Boolean value to specify if the collision is effective.
Output: array of molecules either with size 1 (in case of ineffective collision) or 2 (in case of effective collision).

```
Function collideOnWall(Molecule molecule, boolean effective) {
    MISMolecule results[] = null;
    if(effective == true){
      Molecule results[] = new Molecule[2];
      int mid = (Number Of Nodes in molecule)/2;
      results[0] = new molecule of colliding molecule nodes indexed
      between 0 and mid-1.
      results[1] = Create Molecule of molecule Nodes indexed from mid to
      the end of the list.
    }
    else{
      results = new MISMolecule[1];
      results[0] = molecule;
    }
    return results;
}
```

Fig. 2. Collideonwall Function that Performs the Collision on Wall of the Selected (Parameterized) Molecule According to the Selected Effectiveness.

Name: collideWithMolecule
Input: Two molecule objects that should collide together, and Boolean value to specify if the collision is effective.
Output: Molecule object that represents the synthesized molecule (effective collision) or null (ineffective collision).

```
Function collideWithMolecule(Molecule molecule1, Molecule
molecule2, boolean effective) {
    if(effective==true){
        MISMolecule result = create molecule of nodes contained in
        molecule1 and molecule2
        return result
    }
    return null
}
```

Fig. 3. Collidewithmolecule Function that Performs the Collision between two Selected (Parameterized) Molecules According to the Selected Effectiveness.

Name: next
Input: number of hits (iterations), and maximum number of hits (iterations).
Output: Boolean value that indicates whether CRO algorithm should continue or stop its operations.

```
Function next() {
    if(TotalNumberOfHits<NoOfIterations){
        return true
    }
    return false
}
```

Fig. 4. Next Function Decides whether CRO Algorithm should Perform Further Steps or Stop its Work.

Name: chooseRandomMolecule
Input: Molecule object, and available molecules list.
Output: Chosen Molecule object (Random Selection).

```
Function chooseRandomMolecule(Molecule molecule) {
    Molecule pickedMolecule= null
    if(molecule == null){
        for each molecule in the available molecules{
            if(the molecule contains the default selected node){
                pickedMolecule = current molecule
                Break
            }
        }
    }else{
        int index = random number between 0 and number of available
        molecules
        pickedMolecule = select molecule located at the random index in the
        available molecules list
    }
    return pickedMolecule
}
```

Fig. 5. ChooseRandomMolecule Function that Chooses a Molecule from Available Molecules.

Note that in the code of Fig. 8, the initial selected node that would be selected during CRO life cycle is determined according to the number of neighbors, where it is the node with minimum number of nodes. While in the other implementation, the node is selected randomly among graph nodes regardless of its connectivity, as in Fig. 7.

```
Name: chooseMinimumConflectingMolecule
Input: Molecule object, and available molecules list.
Output: Chosen Molecule object (Minimum Connectivity Degree Molecule).
Function chooseMinimumConflectingMolecule(Molecule molecule){
   Molecule pickedMolecule=null
   if(molecule == null){
      for each molecule in the available molecules {
         if(the molecule contains the default selected node){
            pickedMolecule = current molecule
            Break
         }
      }
   }
   else {
      Molecule minimum = null
      int min = Number of nodes in graph
      for each molecule in the available molecules {
         if (currentMolecule != molecule){
            int temp = number of conflecting nodes in currentMolecule
            if ((minimum == null) or (temp < min)){
               minimum = choosenMolecule
               min = temp
            }
         }
      }
      pickedMolecule = minimum
   }
   return pickedMolecule
}
```

Fig. 6. Chooseminimumconflectingmolecule Function that Chooses Minimum Connectivity Degree Molecule from Available Molecules.

```
Name: initr
Input:Graph Nodes
Output: initializing molecules (conversion of graph nodes into CRO
molecules).
Function initr(){
   noOfIterations = number of graph nodes
   minimumNoOfIterations = 0
   minimumSize = number of graph nodes
   intselectedIndex= pick random number between 0 and number of nodes-1
   foreach node in the graph nodes{
      MISMolecule molecule = create molecule containing current graph node
      only
      molecule.PotentialEnergy = number of graph nodes-1
      molecule.NumberOfHits = 0
      molecule.MinimumHitNumber = 0
      molecule.MinimumStructure  = molecule;
      molecule.MinimumValue = molecule.PotentialEnergy
      add molecule to the available molecules
      if (node index=selectedIndex){
         selectedNode = node
      }
   }
   remove molecules that contain selected node neighbors nodes from the
   available molecules
   noOfIterations = noOfIterations  - number of removed molecules
}
```

Fig. 7. Initr Function that Initializes the Execution of CRO Algorithm and Chooses Starting Molecule Randomly.

B. Example

In this section, an example of the algorithm execution is provided by considering (**M**inimum initial node &**M**inimum iteration node) algorithm. Consider the graph in Fig. 9. The algorithm will initialize CRO molecules by representing each graph node by a single molecule. The potential energy equals to the number of remaining nodes not included in the molecule. So, initially, there are 5 molecules where these molecules contain nodes 1, 2, 3, 4, and 5; while potential energy for each of them is 4. This is because there is a graph node in the molecule, and the remaining graph nodes are not included in the molecules.

The algorithm will pick molecule with minimum conflicting nodes first, and do all reactions on that molecule. In this example, the algorithm can pick one of the molecules containing nodes 1, 4, and 5, as each has minimum number of conflicting nodes, which equals to 2.

```
Name: initmc
Input: Graph Nodes.
Output: initializing molecules (conversion of graph nodes into CRO
molecules) and picks minimum connected node molecule as initial starting
solution.
function initmc(){
   noOfIterations = number of graph nodes
   //the algorithm will iterate exactly the number of nodes
   minimumNoOfIterations = 0
   //initial minimum number of iterations to find solution is 0
   minimumSize = number of graph nodes
   /*minimum solution initially is same number of graph nodes (maximum
   excluded nodes in worst case)*/
   int minimumLinks = number of graph nodes + 1
   /*initial minimum number of node linkes is the number of graph nodes +1
   note that this variable is used to keep track of discovered minimum no of
   node neighbors*/
   foreach node in the graph nodes{
      MISMolecule molecule = create molecule containing current graph node
      only
      //each node in the graph would be represented as a unique molecule.
      molecule.PotentialEnergy = number of graph nodes-1
      /*initial molecule potential energy is the no of remaining graph nodes
      not included in the molecule which is the number of graph nodes-1*/
      molecule.NumberOfHits = 0
      //initial no of hits is 0 where no collisions have occurred.
      molecule.MinimumHitNumber = 0
      //minimum no of hits to find best solution is initially 0
      molecule.MinimumStructure  = molecule;
      /*minimum structure (best solution) is the initial one which is the current
      molecule structure (one node)*/
      molecule.MinimumValue = molecule.PotentialEnergy
      /* minimum value of potential energy (best solution value) is the initial
      one which is the initial potential energy of molecule*/
      add molecule to the available molecules
      //adding molecule to the molecules pool.
      if (minimumLinks> number of node Neighbors)
      {
         selectedNode = node
         minimumLinks = number of node Neighbors
      }
      /*check the number of current node neighbors so that if it is less than
      minimum observed links, then its corresponding molecule will be
      selected to be initial colliding molecule and its number of neighbors is
      saved in minimumLinks to keep track of it and compared to remaining
      nodes*/
   }
   remove molecules that contain selected node neighbors nodes from the
   available molecules
   /*selected node neighbors should be excluded from the molecules pool
   since they won't be part of the solution (IS) since their neighbor node is
   selected to be initial part of the solution*/
   noOfIterations = noOfIterations  - number of removed molecules
   //number of  iterations decreased by the number of removed molecules
}
```

Fig. 8. Initmc Function that Initializes the Execution of CRO Algorithm and Chooses Starting Molecule with Minimum Connectivity Degree.

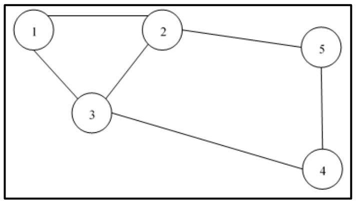


Fig. 9. Example of a Graph of 5 Nodes; Initially, each Node is Considered a Molecule and a Potential Energy for Each is 4.

Pick one of these three nodes randomly; and assume that molecule of node 5 is selected. The algorithm will iterate 5 times (number of nodes in graph). In the first iteration, the algorithm will choose another molecule (molecule that contains node 1) and do the collision with previously selected one (contains node 5). The algorithm will check the effectiveness of collision by checking whether the nodes in the two molecules are conflicting (neighbors) or not; which in this case, they are not. So, the collision is effective, and the molecules should merge (synthesized). This will produce new molecule that contains nodes 1 and 5, and the number of conflicting nodes is 3, and the potential energy is modified to be 3 instead of 4. Assume a molecule that contains node 4 is selected in the next iteration, the collision with the molecule that contains nodes 1 and 5 won't be effective, and nothing would happen because the node 4 conflicts with node 5 included in the molecule.

In case on-wall collision is decided to be performed, the molecule that contains nodes 1 and 5 would be divided into two molecules: a molecule would contain node 1, and another molecule would contain node 5. At the end, the result will be the molecule that achieves lower potential energy, which is in our simple iteration is the molecule that contains nodes 1 and 5. So, the Maximum Independent Set is {1, 5}. In case the molecule that contains node 4 has been chosen to collide with the original molecule (i.e. molecule of node 1), the collision will be effective, and the two molecules will synthesize and form new molecule that contains both nodes. While in case the same molecule that contains node 4 has been chosen to collide with the molecule that contains nodes {1, and 5}, the collision will be marked as ineffective and nothing would happen. This is because node 4 conflicts with node 5 contained in the main molecule.

C. Analytical Evaluation

Given a graph with N nodes, the algorithm will iterate exactly N iterations (Stopping criteria is the iteration of N iterations, where N is the number of nodes in graph). Within each iteration, first of all, the collision type is chosen and defined to be on-wall or inter-molecular collision. If the collision is defined to be on-wall collision, one of the followings will be done according to the effectiveness of collision:

1) Effective on-wall collision: The original molecule is divided into two molecules containing the halves of the original molecule. The original molecule is removed from the molecules pool, and the resultant molecules are added to that pool. In this case, the run time complexity of dividing the molecule is $O(N/2) \approx O(N)$.

2) Ineffective on-wall collision: The original molecule remains with same structure and nothing happens at all, since the original molecule is not affected by the collision. See *collideOnWall* in Fig. 2. In this case, a constant number of steps $O(K)$ is performed, where K is constant number that represents the number of steps needed to check effectiveness flag and going forward to the next step.

On the other hand, if the collision is defined to be inter-molecular collision, one of the followings will be done according to the effectiveness of collision:

1) Effective inter-molecular collision: This is referred to as "collideWithMolecule" function in Fig. 3. Note that effective is a Boolean parameter that indicates whether the collision is effective or not. If the collision is effective, the algorithm will iterate through first molecule, and the second molecule will create new molecule that contains all the nodes contained by the two molecules. So, in worst case, the first molecule contains half of the graph nodes, and the second one contains the other half of the graph nodes. The merge process will iterate with run time cost of $O(N/2)$ to add first molecule nodes; while in the addition of the second molecule it will check every node to prevent adding the same node twice. Every node in the second molecule will be checked across first molecule nodes with run time cost of $O(N/2)$, and this will be done for each node in the second molecule. So, the overall run time complexity is $O([N/2]+[N/2] \times [N/2]) \approx O(N^2)$.

After merging the molecules, the conflicts will be computed by adding first molecule conflicting nodes list to the second molecule conflicting nodes list with redundancy removals. In the worst case, first molecule conflicting nodes are N-2 (e.g. all graph nodes except itself and the merging node), and the second node conflicting nodes are N-2 (e.g. all graph nodes except itself and the merging node). The algorithm will iterate (N-2) to add first conflicting nodes and will iterates (N-2) to add second molecule conflicting nodes. But, while adding second molecule conflicting nodes, it will check the list of the first molecule conflicting node to prevent duplication of the nodes. In this case, all the conflicting nodes in the second molecule will be found in the first molecule. So, the second molecule conflicting nodes will be found in the list of size (N-2) added by the first molecule. This involves finding all the conflicting nodes of the second molecule in the first molecule conflicting list by iterating 1, 2, 3,…, (N/2) iterations. So, the algorithm will iterate [1+2+3+…+(N/2)] iterations to add second molecule conflicting nodes. Thus, the overall complexity is $O(N^2)$.

2) Ineffective inter-molecular collision: In this case, the resultant run time complexity of collision execution will be $O(2N^2)$. In this case, the algorithm won't do anything, refer to "collideWithMolecule" in Fig. 3, and nothing happens while the molecules are returned back to the molecules pool without any processing. So, constant number of steps is performed.

Complexity of collision effectiveness computation:

1) On-wall collision: As described in [1,13,32], the effectiveness of on-wall collision is determined by checking the number of ineffective iterations of the molecule. Ineffective iterations are the iterations that have been done on the molecule after minimum value is found without any improvement. If the number of iterations exceeds a predefined constant value, the collision will be defined to be an effective one; otherwise it is not. So, the run time complexity of determining effectiveness of the collision is constant.

2) Inter-molecular collision: The effectiveness of the collision is determined by checking the readiness of molecules to be merged together. This is done by checking the existence of any of the second molecule nodes within the first molecule conflicting nodes. If the check determines that any of the second molecule nodes exists in the conflicting nodes of the first molecule, the collision is defined to be effective; otherwise, it is not. In the worst case, first molecule contains one node and (N-1) conflicting nodes, while the second molecule contains all the remaining graph nodes so that its size is (N-1). To check the existence of second molecule nodes in the conflicting nodes of the first molecule, the whole list of the first molecule nodes should be iterated for every node in the second molecule, until finding the checked node or reaching the end of the list and the node is assumed to be not conflicting. So, the iterations are, [1,2,3,…,N-1] and the run time complexity is [1+2+3+…+(N-1)].In this case, the run time complexity of the effectiveness calculation is O((N-2)(N-3)/2)=O(N^2).

Complexity of molecule selection types:

1) Random selection: In the random selection, the algorithm will pick a random molecule from the list of available molecules to perform intended operation. So, in this case, no processing is done, and a constant number of steps (K) is performed.

2) Minimum connectivity degree node selection: In this case, the algorithm will iterate through the available molecules to select the molecule with minimum number of conflicting nodes. In the worst case, the number of molecules is equal to the number of graph nodes (N). So, the algorithm will iterate through N molecules to find out the one with minimum number of conflicting nodes. The complexity of finding minimum connectivity degree among N nodes is O(N). The run time complexity of finding the same initial molecule is O(N).

One of the main constants to be defined prior to algorithm execution is Inter-Molecular to On-Wall collisions Ratio (R). According to the value of R, the number of inter-molecular collisions equals to R× (Number of CRO iterations) and on-wall collisions will be (1-R) × (Number of CRO iterations). So, equations (2) and (3) will hold.

$$\text{Number of Inter} - \text{Molecular Collisions} \;=\; \mathrm{O}(R \times N) \qquad (2)$$

$$\text{Number of On} - \text{Wall Collisions} \;=\; \mathrm{O}((1 - R) \times N) \qquad (3)$$

The overall run time complexity of the collision is the complexity of collision effectiveness calculation, the molecule selection complexity, and the collision execution complexity according to its effectiveness; and is expressed as in equation (4).

$$\mathrm{O}(\text{Collision}) = \mathrm{O}(\text{Eff. Calculation}) + \mathrm{O}(\text{Mol. Selection}) + \mathrm{O}(\text{Col. Execution}) \qquad (4)$$

In case of On-Wall collision, there are two cases:

1) Ineffective collision: By applying equation (4), the resultant is equation (5) for On-Wall collision complexity

$$\mathrm{O}(\text{Collision}) = 1\;[\mathrm{O}(K)] + \mathrm{O}(\text{Mol. Selection}) + 1\;[\mathrm{O}(K)] \; (5)$$

Where O(Mol. Selection) depends on the molecule selection criteria. So, in case of random molecule selection, the resultant equation is represented in equation (6). While in case of minimum connectivity degree molecule selection is used, the collision run time complexity is as in equation (7).

$$\mathrm{O}(\text{Rand. Mol. Sel. Col.}) = \mathrm{O}(\text{constant}) \qquad (6)$$

$$\mathrm{O}(\text{Min. Con. Deg. Mol. Sel. Col.}) = \; \mathrm{O}(N) \qquad (7)$$

2) Effective collision: By applying equation (4), the resultant equation (8) of collision complexity is as in equation (7).

$$\mathrm{O}(\text{Col.}) = 1\;[\mathrm{O}(K)] + \mathrm{O}(\text{Mol. Sel.}) + N\;[\mathrm{O}(N)] \qquad (8)$$

Where O(Mol. Selection) depends on the molecule selection criteria. So, in case of random molecule selection, the resultant equation is as in equation (9). While in case of minimum connectivity degree molecule selection is used, the collision complexity is as in equation (10).

$$\mathrm{O}(\text{Rand. Mol. Sel. Col.}) = 1 + 1\;[\mathrm{O}(K)] + N = \mathrm{O}(N) \qquad (9)$$

$$\mathrm{O}(\text{Min. Con. Deg. Mol. Sel. Col.}) = 1 + N\;[\mathrm{O}(KN)] + N = \mathrm{O}(N) \qquad (10)$$

In case of Inter-Molecular collision, there are two cases:

1) Ineffective collision: By applying equation (4), the resultant collision complexity is as in equation (11). The ineffective collision does not perform any operation on the colliding molecule(s). So, the complexity of its execution is constant (K). But the calculation of collision effectiveness in worst-case would check the half of graph nodes against the second half of graph nodes that could be fully connected. So, the final equation would look like the following:

$$\text{Infc} = N^2 \; \times [\mathrm{O}(\textstyle\sum_{i=0}^{N-1} i) = \mathrm{O}(\frac{(N-2)(N-3)}{2})] + \mathrm{O}(\text{Mol. Selection}) + 1\;[\mathrm{O}(K)] \qquad (11)$$

Where O(Mol. Selection) depends on the molecule selection criteria. So, in case of random molecule selection, the resultant is as in equation (12). While in case of minimum connectivity degree molecule selection is used, the collision complexity is as in equation (13).

$$\mathrm{O}(\text{Col.}) = N^2 + 1[\mathrm{O}(K)] + 1 = \mathrm{O}(N^2 + 1) = \mathrm{O}(N^2) \qquad (12)$$

$$\mathrm{O}(\mathrm{Col.}) = N^2 + N[\mathrm{O}(KN)] + 1 = \mathrm{O}(N^2 + N + 1) = \mathrm{O}(N^2) \quad (13)$$

2) Effective collision: By applying equation (4), the resultant of collision complexity is as in equation (14).

$$\mathrm{O}(\mathrm{Col.}) = N^2 \times \left[\mathrm{O}(\textstyle\sum_{i=0}^{N-1} i) = \mathrm{O}\left(\frac{(N-2)(N-3)}{2}\right)\right] + \mathrm{O}(\mathrm{Mol.\,Sel.}) + N^2[\mathrm{O}(2N^2)] \quad (14)$$

In case of random molecule selection, the resultant is as in equation (15)

$$\mathrm{O}(\mathrm{Col.}) = N^2 + 1[\mathrm{O}(K)] + N^2 = \mathrm{O}(N^2) \quad (15)$$

While in case of minimum connectivity degree molecule selection is used, the collision complexity is as in equation (16).

$$\mathrm{O}(\mathrm{Col.}) = N^2 + N[\mathrm{O}(KN)] + N^2 = \mathrm{O}(N^2) \quad (16)$$

Overall complexity of on-wall collision is as in equation (17).

$$\mathrm{O}(\mathrm{on-wall\ collision}) = \mathrm{O}((1-R) \times N^2) = \mathrm{O}(N^2) \quad (17)$$

Overall complexity of inter-molecular collision is as in equation (18).

$$\mathrm{O}(\mathrm{inter-molecular\ collision}) = R \times N \times [\mathrm{O}(N^2 + N^2)] = \mathrm{O}(R \times N^3) \quad (18)$$

The overall run time complexity of CRO algorithm is as in equation (19).

$$\mathrm{O}(\mathrm{CRO}) = \mathrm{O}(\mathrm{on-wall\ collision}) + \mathrm{O}(\mathrm{inter-molecular\ collision})$$

$$\mathrm{O}(\mathrm{CRO}) = O(N^2 + N^3) = O(N^3) \quad (19)$$

IV. Experimental Results

The CRO variations are implemented using Java programming language and tested for comparison purposes using random generated graphs. The generated graphs are saved on permanent storage to insure the execution of various CRO versions on the same graphs for more accurate comparison. Moreover, multiple graph sizes have been generated with different connectivity degree percentages. As in [7], most of these graphs are manipulated using Modified Wilf algorithm to find out their exact MIS solution. These exact solutions are used to find the accuracy percentage of CRO results. CRO has been executed with different inter-molecular to on-wall collision ratio values (4 values) seeking for the best ratio value in term of time and accuracy. Moreover, these executions are repeated 5 times to calculate average execution time looking for more accurate measures. The tested ratio values are (0.25, 0.50, 0.75, and 0.95). The accuracy percentage is defined, as in equation (20).

$$\text{Accuracy Percentage} = \frac{100\times\text{Size(MIS by CRO)}}{\text{Size (MIS by Modified-Wilf)}} \quad (20)$$

The algorithms are tested on a laptop with the following specifications: CPU: Intel (R) Core(TM) i7-4510U CPU @ 2.00GHz 2.60GHz; Memory: 8.00 GB; and Operating System: 64-bit Operating System (Windows 10 Home).

After executing the CRO algorithm over a set of randomized graphs, the algorithm variations have been tested over a set of benchmark datasets to measure accuracy, where optimal solution has achieved in some cases, as shown later.

The resulting execution times are listed below for graphs of sizes range from 100 nodes to 1000 nodes with connectivity degrees (20%, 60%, 80%, and 90%) for the various versions of the proposed CRO algorithm.

Tables II to V show that when we increase the ratio of inter-molecule collision to on-wall collision ratio, the smoothness of chart increases, which indicates that the algorithm runtime much closer to theoretical analysis. On the other hand, the algorithms show unpredictable time results due to on-wall collisions that divide the molecules to two different parts. Note that those new molecules will start over collecting other molecules to formulate new solutions.

The results show that high collision ratio provides better performance for low connectivity degree. Note that this is not the case for graphs with higher connectivity degree, where the maximum collision ratio consumes the highest time. When the collision ratio is decreased, the execution time of the algorithm on highly connected graphs achieves the minimum run time, and for those with lowest connectivity degree it achieves the worst run time. This scenario is a result of checking the efficiency of collision between two molecules. As described in the algorithm code, in order to check whether the molecules can collide effectively the connections (neighbors) of the colliding molecules are checked to be sure of conflicting neighbors. So, in case of high connectivity graphs, this will be done by a higher number of iterations. As long as the ratio of collisions is low, the number of inter-molecule collisions is low, which decreases the number of checks between molecules that minimizes the time of execution.

After calculating average time for the different algorithm executions and for the different selected graphs with different sizes and different connectivity degrees, Fig. 10 shows that the best execution time is achieved when a random starting node and picking minimum connected node in each iteration, and when the inter-molecule to on-wall collision ratio is 75%. Moreover, the results show that when picking minimum connected molecules, in each iteration, it provides better in execution time performance than picking random molecule at each stage. This happens because the number of checks of conflicting nodes between colliding molecules is minimum when the two molecules are picked according to minimum connectivity. While in case of random molecules, there is no guarantee of the number of connections in the picked molecules which could be the highest, so that the number of checks of conflicting nodes is high.

TABLE. II. Experimental Execution Time in MSEC. for 20%, 60 %, 80%, and 90% Connectivity Degrees' Graphs with Initial Random Graph Node Selection and Random Molecule Selection (RR) in Each iteration with (Collision Ratios 0.25, 0.5, 0.75, and 0.95)

Connectivity Degree	**20%**				**60%**				**80%**				**90%**			
Size/Collision Ratio	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**
100	593	362	256	222	50	59	46	50	19	16	19	12	6	6	3	9
200	4619	3209	1912	1925	272	256	246	271	28	53	53	49	15	15	15	16
300	13675	9078	5958	6147	900	698	831	872	137	128	121	156	37	31	34	28
400	35356	22709	15665	14917	1791	1493	1725	2094	212	234	262	281	44	47	47	46
500	62235	47346	29810	30129	4531	2631	3256	4047	218	412	409	528	84	75	84	87
600	117545	72077	48263	51952	8056	4695	5490	6882	968	625	825	912	75	112	118	134
700	187109	126150	79830	82306	10953	8171	8324	11051	1250	953	1122	1424	140	140	165	193
800	269999	192701	107425	121437	12533	11324	12495	16707	831	1297	1787	2106	237	187	240	281
900	407757	277431	166954	177255	18281	14379	18175	22828	2515	1796	2368	2940	124	234	356	406
1000	551154	373748	232953	242983	26922	20995	24464	32371	2762	1965	3231	4069	281	381	431	544

TABLE. III. Experimental Execution Time in MSEC. for 20%, 60 %, 80%, and 90% Connectivity Degrees' Graphs with Initial Minimum Graph Node Selection and Random Molecule Selection (MR) in Each Iteration with (Collision Ratios 0.25, 0.5, 0.75, and 0.95)

Connectivity Degree	**20%**				**60%**				**80%**				**90%**			
Size/Collision Ratio	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**
100	606	409	287	251	109	81	65	69	28	25	21	18	19	12	6	10
200	4230	3035	2130	2057	553	379	296	367	90	78	95	91	43	28	34	26
300	14841	11393	7000	6832	1234	906	991	1198	212	250	253	280	59	53	59	57
400	32222	24763	14987	15814	3309	2185	2100	2555	496	303	381	451	77	87	93	106
500	67398	45029	31119	30814	4398	3722	4239	5010	543	510	793	925	168	122	178	183
600	121295	75942	49422	54483	6157	5507	7376	9413	700	862	925	1171	122	206	224	246
700	189972	137178	87550	89818	12060	10264	11134	13771	1034	1090	1503	1940	412	275	331	425
800	266150	188499	123894	130012	24717	15082	16855	21770	1847	2119	2647	3414	390	365	478	597
900	409454	267807	161301	187702	24989	19514	23766	31255	4853	2406	3150	3885	432	453	621	757
1000	545105	389631	229173	256270	35639	26005	27629	37977	2232	3428	4365	5817	409	568	797	1044

TABLE. IV. Experimental Execution Time in MSEC. for 20%, 60 %, 80%, and 90% Connectivity Degrees' Graphs with Initial Random Graph Node Selection and Minimum Molecule Selection (RM) in each Iteration with (Collision Ratios 0.25, 0.5, 0.75, and 0.95)

Connectivity Degree	**20%**				**60%**				**80%**				**90%**			
Size/Collision Ratio	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**
100	476	328	212	193	80	47	43	46	21	15	12	9	7	6	3	0
200	3917	2423	1525	1439	299	218	241	218	56	34	34	31	21	21	15	12
300	5952	5385	4818	4869	753	616	679	650	134	87	100	93	34	31	31	28
400	23999	16660	11689	11410	975	1113	1263	1486	362	228	196	203	56	44	34	37
500	52396	34185	24442	22005	3730	2803	2370	2844	478	359	325	387	62	53	56	68
600	34862	36307	38833	38067	4160	3831	4451	4900	375	353	515	669	68	81	106	97
700	94925	101125	54167	61317	5401	6825	7222	7741	356	750	903	968	81	97	128	149
800	233388	173169	94075	91432	2106	8976	10587	11448	1053	697	1262	1478	200	206	200	231
900	334125	149173	143130	129653	7993	16879	14036	16306	1284	1028	1750	2112	259	253	250	284
1000	309296	230572	168135	176303	8178	10025	19302	22409	1659	2281	2365	2872	331	337	322	365

TABLE. V. BEST ACCURACY RESULTS OF INITIAL RANDOM GRAPH NODE SELECTION AND RANDOM MOLECULE SELECTION (RR) IN EACH ITERATION ON 20%, 60%, 80%, AND 90% CONNECTIVITY DEGREE GRAPHS WITH (COLLISION RATIOS 0.25, 0.5, 0.75, AND 0.95)

Connectivity Degree	20%				60%				80%				90%			
Size/Collision Ratio	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**
100	32	58	63	74	57	71	86	71	67	67	67	67	50	75	75	75
200					44	56	56	67	80	80	100	80	75	75	75	75
300					42	58	50	67	67	67	67	83	60	60	80	80
400					45	55	64	55	57	57	57	71	60	60	80	80
500					50	70	60	70	43	57	57	57	60	60	60	80
600					36	43	57	50	57	57	57	71	80	80	80	80
700					45	64	64	64	57	57	71	71	60	80	60	80
800					36	50	57	57	38	63	50	63	80	80	80	80
900					46	54	62	62	50	63	75	63	50	50	67	83
1000					31	44	44	56	56	44	56	56	50	67	67	67

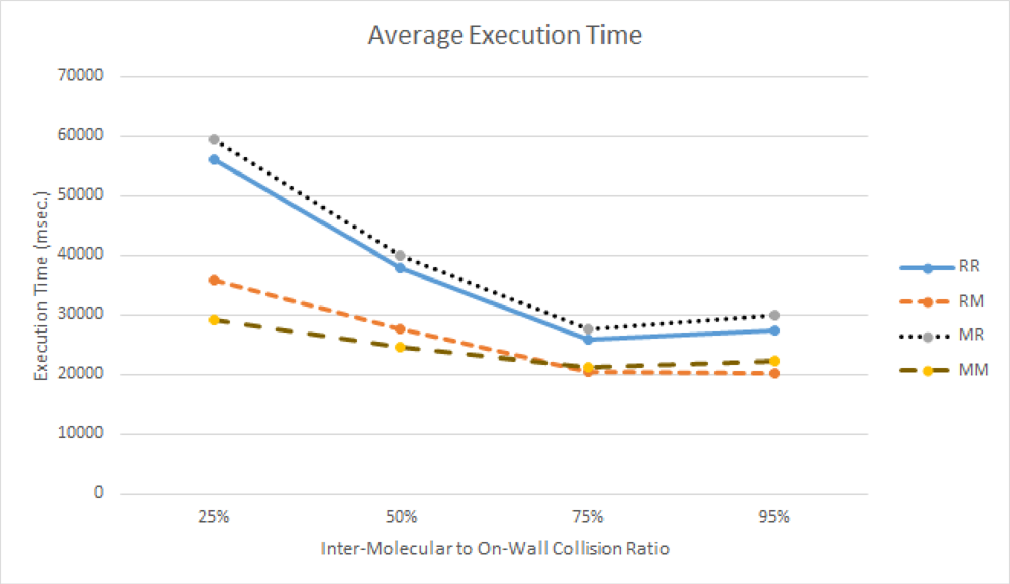


Fig. 10. Average Execution Time for All Tested CRO Algorithm Versions with (Collision Ratios 0.25, 0.5, 0.75, and 0.95).

In Table VI, we demonstrate the accuracy of CRO algorithm running using full random selection of initial and iteration molecules, where the accuracy is calculated using equation (20) according to Modified Wilf algorithm results. In 20% connectivity degree, Modified Wilf algorithm can obtain an MIS from a graph of up to 150 nodes in an acceptable time. While in the higher connectivity degrees (60%, 80%, and 90%) the solutions are obtained in a graph of up to 1000 nodes. The results show that the accuracy is dropping when the number of nodes is going up. When the connectivity degree increases, the accuracy becomes more stable and near to constant regardless of graph size. Moreover, when CRO algorithm is run using 95% inter-molecules to on-wall collision ratio, it provides better results. This is a result of performing more inter-molecule collisions, which provides more combinations of nodes (solutions), so that better solutions could be discovered.

Table VII demonstrates the accuracy of CRO algorithm using random selection of iteration molecules, while starting with minimum molecule (minimum connected node), where the accuracy is calculated using equation (20) according to Modified Wilf algorithm results. The results show that the accuracy is dropping when the number of nodes is growing up. When the connectivity degree increases, the accuracy becomes more stable and near to constant regardless of graph size. Moreover, when CRO algorithm is run using 95% inter-molecules to on-wall collision ratio, it provides better results.

Table VIII shows accuracy results of CRO algorithm using random selection of initial iteration molecules and picking minimum connectivity node in each iteration, where the accuracy is calculated using equation (20) according to Modified Wilf algorithm results. In the higher connectivity degrees (60%, 80%, and 90%), the solutions are obtained from graph of up to 1000 nodes. The results show that the accuracy is dropping when the number of nodes is growing up. This is a normal result of increasing the number of nodes, where the size of MIS becomes greater, so that the percentage won't be affected by low number of nodes, not like small solutions, where a single node could increase the percentage of accuracy by a significant value.

Table IX shows the accuracy results of CRO algorithm using minimum connectivity molecule and selecting minimum connectivity molecule in each iteration, where the accuracy is calculated using equation (20) according to Modified Wilf algorithm results. In the higher connectivity degrees (60%, 80%, and 90%), the solutions are obtained up to 1000 nodes graph size. The results show that the accuracy is dropping when the number of nodes is going up. Moreover, the algorithm shows almost identical accuracy regardless of inter-molecule to on-wall collisions ratio. This indicates that the best results are obtained early at the beginning of execution, so that it doesn't differ if the collisions between the molecules are increased or not. This indication can be used to decrease the number of iterations in case of higher ratio; but the problem is how to obtain stopping condition?

Fig. 11 shows the average accuracy of each type of algorithms along with inter-molecule to on-wall collisions ratio. The figure shows that random selection of molecules in CRO iterations provides better accuracy results, especially, when the ratio of inter-molecule to on-wall collisions increases. The algorithm performance on graph with95% ratio provides better results in case of random selection. These results represent the worst results in term of accuracy among all tests. This is a result of using minimum number of neighbors as selection criteria for initial base molecule and other molecules in each iteration, so that static selection of colliding nodes is performed, and less nodes combinations are discovered.

Extra experiments have been done to test proposed implementation on benchmark datasets, such as Graph50_10, Graph100_10, Hamming6_2, Hamming6_4, Hamming8_4, and Hamming10_4 obtained from [43,44,45]. The results listed in Table X show that the CRO algorithm provides optimal solution in some cases, specially, when the selection of molecules is done in random and the inter-molecular to on-wall collisions ratio is high, such as (75% or 95%). On the other hand, the results show that minimum degree molecule selection criteria provide lower accuracy, which tends to be the result of selecting special molecules each time of collision, which could deviate from the correct path of optimal solution that may contain higher degree nodes. The results show that optimal solution (Exact solution) of MIS could be achieved by CRO. But, the main problem is that this result is not guaranteed. CRO should be executed many times (in our case 10 times) to have more solutions that may contain the optimal one. So, if the execution of CRO is finished within 1 second, and the re-execution is done 10 times, this means that the total execution time is 10 seconds, which is the actual time to be compared with. This makes Modified-Wilf better choice and more worthy to use in case of small problems (lower graph size and higher connectivity), since the difference of achieved performance is low with guaranteed results.

TABLE. VI. BEST ACCURACY RESULTS OF INITIAL MINIMUM GRAPH NODE SELECTION AND RANDOM MOLECULE SELECTION (MR) IN EACH ITERATION ON 20%, 60%, 80%, AND 90% CONNECTIVITY DEGREE GRAPHS WITH (COLLISION RATIOS 0.25, 0.5, 0.75, AND 0.95)

Connectivity Degree	**20%**				**60%**				**80%**				**90%**			
Size/Collision Ratio	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**
100	42	58	74	68	43	71	86	100	50	67	67	67	50	75	75	100
200					44	56	67	67	60	80	80	80	75	75	75	75
300					42	50	58	58	67	83	83	67	80	60	80	80
400					45	55	64	64	57	71	71	71	60	60	60	80
500					50	60	70	70	43	57	71	71	60	80	80	80
600					36	50	50	57	57	57	71	71	60	60	80	80
700					55	55	73	73	57	57	71	71	80	60	60	80
800					43	43	50	57	50	63	63	63	80	80	60	80
900					38	54	62	62	50	75	63	75	50	67	50	67
1000					38	44	50	56	44	56	56	78	50	50	67	67

TABLE. VII. EXPERIMENTAL EXECUTION TIME IN MSEC. FOR 20%, 60 %, 80%, AND 90% CONNECTIVITY DEGREES' GRAPHS WITH INITIAL MINIMUM GRAPH NODE SELECTION AND MINIMUM MOLECULE SELECTION (MM) IN EACH ITERATION WITH (COLLISION RATIOS 0.25, 0.5, 0.75, AND 0.95)

Connectivity Degree	**20%**				**60%**				**80%**				**90%**			
Size/Collision Ratio	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**
100	602	385	266	209	127	84	75	50	27	18	12	12	16	15	15	6
200	1002	1493	1562	1616	622	427	293	262	54	46	62	68	36	31	21	22
300	13926	8450	6857	5362	299	403	719	912	384	275	206	200	53	40	43	43
400	38450	23482	13688	11688	2965	2118	1897	1776	600	440	331	319	87	62	75	81
500	49903	36941	26663	23738	972	1603	2972	3635	1228	837	671	653	215	200	128	131
600	113015	70506	47772	40013	15934	10204	7210	6739	1650	1290	937	802	87	121	159	190
700	90279	80873	63352	66949	2159	4441	7888	9837	362	794	1084	1406	153	187	284	293
800	138417	97888	83193	98044	29062	19616	15420	15520	3347	2821	2600	2434	221	268	322	447
900	384084	229520	160319	139371	3503	9079	16972	22223	772	1409	2290	2816	1187	737	615	525
1000	27344	79701	140458	187492	41688	37861	27133	26490	959	2088	3381	4256	225	400	603	718

TABLE. VIII. Best Accuracy Results of Initial Random Graph Node Selection and Minimum Molecule Selection (RM) in each Iteration on 20%, 60%, 80%, and 90% Connectivity Degree Graphs with (Collision Ratios 0.25, 0.5, 0.75, and 0.95)

Connectivity Degree	**20%**				**60%**				**80%**				**90%**			
Size/Collision Ratio	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**
100	37	37	42	26	57	43	57	43	50	50	67	50	75	100	100	75
200					56	44	44	44	60	80	60	40	75	100	100	100
300					33	33	33	33	67	50	33	50	60	60	60	80
400					27	36	36	36	43	43	43	43	60	60	60	60
500					30	40	30	40	43	43	29	43	40	60	60	80
600					36	21	29	21	43	29	29	43	60	60	60	60
700					27	27	36	27	43	57	57	43	40	40	80	60
800					21	21	29	21	38	25	50	38	60	60	60	40
900					23	23	31	23	38	25	50	50	50	50	33	33
1000					19	13	19	25	33	44	33	33	50	50	33	50

TABLE. IX. Best Accuracy Results of Initial Minimum Graph Node Selection and Minimum Molecule Selection (MM) in each Iteration on 20%, 60%, 80%, and 90% Connectivity Degree Graphs with (Collision Ratios 0.25, 0.5, 0.75, and 0.95)

Connectivity Degree	**20%**				**60%**				**80%**				**90%**			
Size/Collision Ratio	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**
100	32	32	32	21	43	43	43	43	33	33	33	33	75	75	75	75
200					33	33	33	33	40	40	40	40	75	75	100	75
300					17	17	17	17	50	50	50	50	40	40	40	40
400					27	27	27	27	43	43	43	43	40	40	40	40
500					20	20	20	20	43	43	43	43	60	60	60	60
600					29	29	29	29	57	57	57	57	40	40	40	40
700					18	18	18	18	29	29	29	29	40	40	40	40
800					21	21	21	21	50	50	50	50	40	40	40	40
900					15	15	15	15	25	25	25	25	83	83	83	83
1000					25	25	25	25	22	22	22	22	33	33	33	33

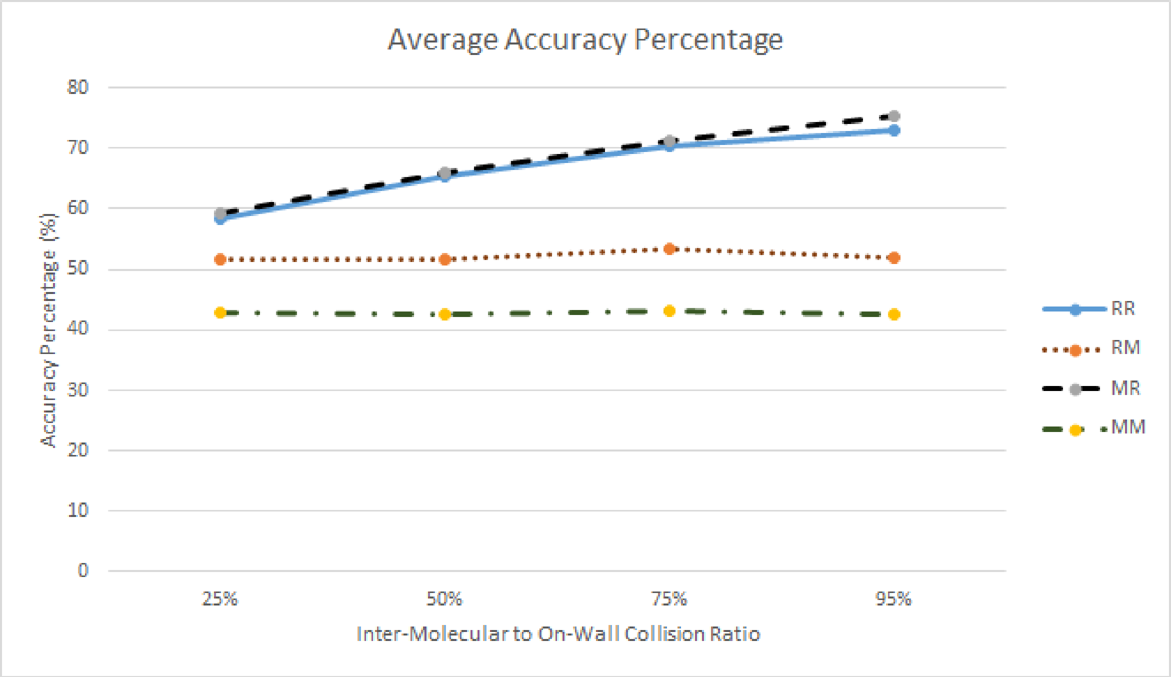


Fig. 11. Average Accuracy for the Tested CRO Algorithm Versions with (Collision Ratios 0.25, 0.5, 0.75, and 0.95).

TABLE. X. SIZES OF MIS RESULTED FROM EXECUTING CRO ALGORITHM ON A SELECTED SET OF BENCHMARK DATASETS

CRO Algorithm	**Optimal MIS**	**CRO-RR**				**CRO-RM**				**CRO-MR**				**CRO-MM**			
Benchmark/Collision Ratio		**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**	**0.25**	**0.5**	**0.75**	**0.95**
Graph50_10	15	7	11	14	15	14	15	15	15	8	12	14	15	14	14	14	14
Graph100_10	30	25	30	30	30	30	30	30	30	22	30	30	30	30	30	30	30
Hamming6_2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Hamming6_4	12	10	10	12	12	9	10	10	9	8	10	12	12	8	10	9	9
Hamming8_4	16	8	13	16	16	10	12	12	11	8	14	16	16	10	10	10	12
Hamming10_4	20	7	11	20	20	11	12	12	11	8	11	17	20	11	11	11	11

V. CONCLUSIONS AND FUTURE WORK

In this paper, scenarios of CRO algorithm have been implemented and applied to solve the MIS problem. The CRO algorithm is applied and evaluated on a set of randomized graphs and the accuracy of the results has been compared against exact solutions obtained by Modified Wilf algorithm.

The algorithm mainly converts graph nodes into a set of molecules. After that, it picks one of the molecules to be its main molecule for reaction. Then, it iterates while picking another interacting molecule to collide with initial one. In this paper, the selection of molecule was implemented using random one and minimum connected one.

The algorithm is tested with variety of parameters, and provides good results in some cases; such as the results of executing RR version with higher collision ratio of the algorithm, as shown in Tables VI and X. This shows that CRO technique can be used to find the solution of MIS problem, and it can be modified to provide better results. The random technique of selecting initial molecule and selecting molecules that are involved in the reaction iterations achieves better accuracy, while guided technique that depends on the degree of connectivity achieves better execution time. This can lead to look for a combination of both techniques to achieve better results in term of execution time and accuracy. Note that the average accuracy of random approach is about 75%, and achieves exact solution in some cases, while minimum approach outperforms random approach by decreasing the execution time by at least 25%.

As future work, this algorithm will be adapted to run on a parallel architecture, just like in [46,47], where a parallel heuristic local search is used to solve travelling salesman problem using four parallel architectures (e.g. “OTIS-Hypercube”, “OTIS-Mesh”, OTIS hyper hexa-cell, and OTIS mesh of trees optoelectronic architectures) and testing against other architectures, such as “The optical chained-cubic tree interconnection network” which is illustrated in [48].

REFERENCES

[1] A. Lam, V. Li, “Chemical-Reaction-Inspired Metaheuristic for Optimization”. IEEE Transactions on Evolutionary Computation, 2010, 14(3): pp: 381 – 399, https://doi.org/10.1109/TEVC.2009.2033580.

[2] S. Butenko, “Maximum Independent Set And Related Problems, With Applications”. PhD dissertation, University Of Florida, 2003.

[3] Y. Shang, "Groupies in random bipartite graphs". Applicable Analysis and Discrete Mathematics, 2010, 4(2), pp: 278–283, DOI: 10.2298/AADM100605021S.

[4] Y. Shang, "On the Hamiltonicity of random bipartite graphs". Indian Journal of Pure and Applied Mathematics, 2015, 46(2), pp: 163–173, DOI: 10.1007/s13226-015-0119-6.

[5] Y. Liu, J. Lu, H. Yang,X. Xiao, and Z. Wei, “Towards maximum independent sets on massive graphs”. in Proceedings of the VLDB Endowment, 2015, 8(13), pp:2122-2133, https://doi.org/10.14778/2831360.2831366.

[6] S. Abu Nayeem and M. Pal Madhumangal, “Genetic algorithmic approach to find the maximum weight independent set of a graph”, Journal of Applied Mathematics and Computing, 2007, 25(1), pp: 217-229. https://doi.org/10.1007/BF02832348.

[7] A. Al-Jaber and A. Sharieh, “Algorithms Based on Weight Factors for Maximum Independent Set”. DIRASAT , 1999,27(1), pp: 74-90.

[8] D. Andrade, M. Resende, and R. Werneck, “Fast local search for the maximum independent set problem”. Journal of Heuristics, (2012), 18(4), pp: 525-547. https://doi.org/10.1007/s10732-012-9196-4.

[9] T. Chan and S. Har-Peled , “Approximation Algorithms for Maximum Independent Set of Pseudo-Disks”. Discrete & Computational Geometry,2012, 48(2), pp: 373-392, https://doi.org/10.1007/s00454-012-9417-5.

[10] J. Robson, “Algorithms for maximum independent sets”. Journal of Algorithms, 1986, 7(3), pp:425-440, https://doi.org/10.1016/0196-6774(86)90032-5.

[11] Wikipedia. “Independent set (graph theory)”.2016, [online] Available at: https://en.wikipedia.org/wiki/Independent_set_(graph_theory) [Accessed 01 Nov. 2018].

[12] A. Lam, Li Victor, “Chemical Reaction Optimization: a tutorial. Memetic Computing”, 2012, 4(1), pp: 3-17, https://doi.org/10.1007/s12293-012-0075-1.

[13] M. Xiao and H. Nagamochi. “Exact Algorithms for Maximum Independent Set”. Algorithms and Computation, 2013, pp: 328-338. Berlin, Heidelberg: Springer, https://dx.doi.org/10.1016/j.ic.2017.06.001.

[14] J. Kneis, A. Langer, and P. Rossmanith, “A Fine-Grained Analysis of a Simple Independent Set Algorithm, ” in Proceedings of FSTTCS 2009, 2009, pp: 287–298, https://doi.org/10.4230/LIPIcs.FSTTCS.2009.2326.

[15] N. Bourgeois, B. Escoffier, V. Paschos, and J. Van Rooij, (2012) “Fast algorithms for Max Independent Set,” Algorithmica, 62(1).382–41, https://doi.org/10.1007/s00453-010-9460-7.

[16] A. Sharieh, W. Al-Rawagepfeh, M. Mahafzah, and A. Al-Dahamsheh, “An Algorithm for finding Maximum Independent Set in a Graph”. European Journal of Scientific Research, 2008, 23(4), pp: 586-596.

[17] T. Back and S. Khuri, “An evolutionary heuristic for the MIS problem, Evolutionary Computation” IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on Computational Intelligence, 1994, pp: 531-535, https://doi.org/10.1109/ICEC.1994.350004.

[18] X. Liu, A. Sakamoto, T. Shimamoto, “A genetic algorithm for maximum independent set problems”. in Proceedings of 1996 IEEE International Conference on Systems, Man and Cybernetics, Beijing, China, 14-17 October 1996, pp: 1916 - 1921 vol.3, https://doi.org/10.1109/ICSMC.1996.565404.

[19] S. Mehrabi, A. Mehrabi, and A. Mehrabi, “A New Hybrid Genetic Algorithm for Maximum Independent Set Problem”. In Proceedings of the 4th International Conference on Software and Data Technologies, (ICSOFT 2009), Sofia, Bulgaria, July 26-29, 2009, pp: 314 – 317, https://doi.org/10.5220/0002253403140317.

[20] L. Youmei and X. Zongben, “An Ant Colony Optimization Heuristic for Solving MIS Problems, “Computational Intelligence and Multimedia Applications, ICCIMA 2003. Proceedings. Fifth International Conference, pp: 206-211, 2003, http://dx.doi.org/10.1109/ ICCIMA.2003.1238126.

[21] M. Alshraideh, B. Mahafzah, H. Salman, and I. Salah, “Using genetic algorithm as test data generator for stored PL/SQL program units”, Journal of Software Engineering and Applications,2013, 6(2), pp: 65-73, http://dx.doi.org/10.4236/jsea.2013.62011.

[22] M. Alshraideh, B. Mahafzah, and S. Al-Sharaeh, “A multiple-population genetic algorithm for branch coverage test data generation”, Software Quality Journal, 2011, 19(3), pp: 489-513, https://doi.org/10.1007/s11219-010-9117-4.

[23] I. Razgan, “Faster Computation of MIS and Parameterized Vertex Cover for Graphs with Maximum Degree 3,” Journal of Discrete Algorithms, 2009, 7(2), pp: 191-212, https://doi.org/10.1016/j.jda.2008.09.004.

[24] M. Xiao and H. Nagamochi, “An exact algorithm for maximum independent set in degree-5 graphs”. Discrete Applied Mathematics 199, 2016, pp: 137–155, https://doi.org/10.1016/j.dam.2014.07.009.

[25] J. Puchinger and G. Raidl, “Combining Metaheuristics and Exact Algorithms in Combinatorial Optimization: A Survey and Classification”. Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach (pp” 41-53). Berlin, Heidelberg: Springer,2005, https://doi.org/10.1007/11499305_5.

[26] V. Cung, S. Martins, C. Ribeiro, and C. Roucairol. “Strategies for the Parallel Implementation of Metaheuristics”. Essays and Surveys in Metaheuristics, US: Springer,2002, (pp. 263-308). https://doi.org/10.1007/978-1-4615-1507-4_13.

[27] H. Kim, H. Lam, and S. Kang, “Chemical Reaction Optimization for Task Scheduling in Grid Computing”. IEEE Transactions on Parallel and Distributed,2011, 22(10), pp: 1624 – 1631, https://doi.org/10.1109/TPDS.2011.35.

[28] E. Loiola, N. de Abreu, P. Boaventura-Netto, P. Hahn, and T. Querido, “A survey for the quadratic assignment problem”. Eur J Oper Res, 2007, 176(2), pp:657–690, https://doi.org/10.1016/j.ejor.2005.09.032.

[29] J. Xu, A. Lam, and V. Li, “Parallel Chemical Reaction Optimization for the Quadratic Assignment Problem”. Proceedings of the 2010 International Conference on Genetic and Evolutionary Methods, GEM 2010, July 12-15, 2010, Las Vegas Nevada, USA.

[30] E. Demeulemeester and W. Herroelen,“Project scheduling: a research handbook”. Academic Publishers, Boston, MA, USA, 2002, https://doi.org/10.1007/b101924.

[31] A. Subramanian, H. Gupta, S. Das, and J. Cao, “Minimum interference channel assignment in multiradio wireless mesh networks”. IEEE Trans Mobile Comput, 2008, 7(12), pp:1459–1473, https://doi.org/10.1109/TMC.2008.70.

[32] A. Lam, J. Xu, and V. Li, “Chemical reaction optimization for population transition in peer-to-peer live streaming”. Proceedings of the IEEE congress on evolutionary computation. Barcelona, Spain, 2010, https://doi.org/10.1109/CEC.2010.5585933.

[33] B. Pan, A. Lam, and V. Li, “Network coding optimization based on chemical reaction optimization”. Proceedings of the IEEE global communications conference. Houston, TX, USA, 2011, https://doi.org/10.1109/GLOCOM.2011.6133697.

[34] M. Kim, M. Medard, V. Aggarwal, U. OReilly, W. Kim, and C. Ahn, “Evolutionary approaches to minimizing network coding resources”. Proceedings of the 26th annual IEEE conference on computer Communications, Anchorage, AK, USA,2007, https://doi.org/10.1109/INFCOM.2007.231.

[35] P. Palmes, T. Hayasaka, and S. Usui, “Mutation-based genetic neural network”. IEEE Trans Neural Network, 2005, 16(3), pp:587–600, https://doi.org/10.1109/TNN.2005.844858.

[36] J. Yu, A. Lam, and V. Li, “Chemical reaction optimization for the set covering problem”. in Proceedings of 2014 IEEE Congress on Evolutionary Computation (CEC 2014), Beijing, China, 6-11 July 2014, In IEEE CEC Proceedings, 2014, pp: 512-519, https://doi.org/10.1109/CEC.2014.6900233.

[37] Y. Shang, "Poisson approximation of induced subgraph counts in an inhomogeneous random intersection graph model". Bulletin of the Korean Mathematical Society, in press.

[38] T. Truong, K. Li, and Y. Xu, “Chemical reaction optimization with greedy strategy for the 0–1 knapsack problem”. Applied Soft Computing, 2013, 13(4), pp: 1774–1780, https://doi.org/10.1016/j.asoc.2012.11.048.

[39] W. Szeto, Y. Wang, and S. Wong, “The chemical reaction optimization approach to solving the environmentally sustainable network design problem”. Computer-Aided Civil and Infrastructure Engineering, 2014, 29(2), pp: 140-158, https://doi.org/10.1111/mice.12033.

[40] Y . Sun, A. Lam, V. Li, J. Xu, and J. Yu, “Chemical reaction optimization for the optimal power flow problem”. The 2012 IEEE Congress on Evolutionary Computation (CEC 2012), Brisbane, Australia, 10-15 June 2012. In IEEE CEC Proceedings, 2012, pp: 1-8, https://doi.org/10.1109/CEC.2012.6253003.

[41] Y. Khanafseh, M. Surakhi, A. Sharieh, and A. Sleit, “A Comparison between Chemical Reaction Optimization and Genetic Algorithms for Max Flow Problem”, International Journal of Advanced Computer Science and Applications (IJACSA), 2017, 8(8), pp: 8-15, http://dx.doi.org/10.14569/IJACSA.2017.080802.

[42] R. Barham, A. Sharieh, and A. Sliet, “Chemical Reaction Optimization for Max Flow Problem”, (IJACSA) International Journal of Advanced Computer Science and Applications, 2016, 7(8), pp: 189-196.

[43] K. Xu,“Vertex Cover Benchmark Instances (DIMACS & BHOSLIB)”. IJEA (international journal of Experimental algorithms),2012, 3(1), pp: 1-18.

[44] Penn State Harrisburg University. Vertex Cover Benchmark Instances, 2019. [online] Available at: https://turing.cs.hbg.psu.edu/benchmarks/ vertex_cover.html [Accessed 27 March 2019].

[45] DIMACS. the Center for Discrete Mathematics and Theoretical Computer Science, 2019. [online] Available at: http://dimacs.rutgers.edu [Accessed 8 March 2019].

[46] A. Al-Adwan, B. Mahafzah, and A. Sharieh, “Solving traveling salesman problem using parallel repetitive nearest neighbor algorithm on OTIS-Hypercube and OTIS-Mesh optoelectronic architectures”, Journal of Supercomputing, 2018, 74(1), pp: 1-36, https://doi.org/10.1007/s11227-017-2102-y.

[47] A. Al-Adwan, A. Sharieh, and B. Mahafzah, "Parallel heuristic local search algorithm on OTIS hyper hexa-cell and OTIS mesh of trees optoelectronic architectures" Applied Intelligence, 2018, 49(10), pp: 1-28, https://doi.org/10.1007/s10489-018-1283-2.

[48] B. Mahafzah, M. Alshraideh, T. Abu-Kabeer, E. Ahmad, and N. Hamad, “The optical chained-cubic tree interconnection network: Topological structure and properties” Computers & Electrical Engineering,2012, 38(2), pp: 330-345, https://doi.org/10.1016/j.compeleceng.2011.11.023.

DLBS: Decentralize Load-Balance Scheduling Algorithm for Real-Time IoT Services in Mist Computing

Hosam E. Refaat[1]
Dept. of Information System
Faculty of Computers and Informatics
Suez Canal University, Egypt

Mohamed A.Mead[2]
Dept. of Computer Science
Faculty of Computers and Informatics
Suez Canal University, Egypt

***Abstract*—Internet of Things (IoT) has been industrially investigated as Platforms as a Services (PaaS). The naive design of these types of services is to join the classic centralized Cloud computing infrastructure with IoT services. This joining is also called CoT (Cloud of Things). In spite of the increasing resource utilization of cloud computing, but it faces different challenges such as high latency, network failure, resource limitations, fault tolerance and security etc. In order to address these challenges, fog computing is used. Fog computing is an extension of the cloud system, which provides closer resources to IoT devices. It is worth mentioning that the scheduling mechanisms of IoT services work as a pivotal function in resource allocation for the cloud, or fog computing. The scheduling methods guarantee the high availability and maximize utilization of the system resources. Most of the previous scheduling methods are based on centralized scheduling node, which represents a bottleneck for the system. In this paper, we propose a new scheduling model for manage real time and soft service requests in Fog systems, which is called Decentralize Load-Balance Scheduling (DLBS). The proposed model provides decentralized load balancing control algorithm. This model distributes the load based on the type of the service requests and the load status of each fog node. Moreover, this model spreads the load between system nodes like wind flow, it migrates the tasks from the high load node to the closest low load node. Hence the load is expanded overall the system dynamically. Finally, The DLBS is simulated and evaluated on truthful fog environment.**

***Keywords*—Cloud computing; fog computing; mist computing; IoT; load balancing; reliability**

I. Introduction

Cloud computing is presented as an ongoing innovation, which is totally dependent on the web. The engineering of the Cloud computing depends on a focal server that keep up a tremendous measure of sharing database, various assets and an enormous number of business applications. Then again, a colossal number of remote customers that has a place with various associations can profit by the various administrations given by the focal server. Every remote client has its own, working framework and internet browser that work autonomously on the substance of the cloud server [1, 2]. The association of the client to the web is the main prerequisite from the client to use the cloud server capacities. Along these lines, the IT business and any little association can get these services from the cloud without spending tremendous measure of cash in equipment or software. As a matter of fact, the execution of the cloud introduces a few related ideas. These ideas manage virtualization, resource allocation, computing distribution, utilization of bandwidth, load balancing, fault tolerance, high availability and dynamic scalability for various classifications of data and applications. The administration of the operations identified with every one of these concepts is performed by the cloud service provider.

The cloud providers allocate the resources to the end clients as a service relying upon the uniqueness of the service models and furthermore dependent on the client needs. The service models may incorporate Software as a service known as SAAS, Platform as a service known as PAAS, Infrastructure as a service known as IAAS. These services are inclined on one another and in a pool way.

By and large, the executions of the various procedures on the cloud present a few advantages to the end clients. At First, the data is shared more than one stage, so better services are conveyed to every user. Also, the end user can get the services resources on-demand, flexible, reliable and portable way as indicated by his need as it were.

In spite of these advantages that can be offered by cloud computing to enormous applications, it faces a lot of challenges [3]. The first challenge happens when the number of the clients is increased. For this situation, the requests are broadened to increase the number of services than the cloud capacities. As client requests is increased, as the responses time is increased unless the available resources and the available bandwidth are upraised to acquire all the extra requests. The second challenge happens when the created data by the cloud services is migrated through a long distance from the cloud to the clients. The far distance creates additional challenge about the data security. Moreover, an unpredictable abundance in the workload may cause the need to create a novel load balancing strategy. The load balancing is the reasonable assignment of the task among the parallel resources such as networking, hard drives and computers [4]. In this way, it will be required to achieve the improvement in the distribution of the computation resources and storage devices. So as to beat these challenges, another innovation of profoundly virtualized processing model has been displayed known as Fog computing. The model [5] is proposed by CISCO to be held as cloud edge of an enterprise organizes. The control of the fog computing isn't a substitution of the cloud computing. In reality, it fills in as a steady domain that can give high QoS to the diverse client requests of the close distances. In this way, the entire fog-cloud colony comprises of a set of fogs computing servers and a set of the clouds computing centers.

For the most part, the activities of the Fog processing are like the cloud computing with two fundamental differences. The principal difference is identified with the area of the fog computing that is put near the clients. Subsequently, the fog computing can be envisioned as a nearby cloud. The second difference related to the resources capacities of the fog that have fewer capacities contrasted with the capacities of the cloud assets. In any case, each fog computing incorporates its very own server that is bolstered by its own resources. Furthermore, each fog server is involved by the vital software or firmware to set up the required VMs, for example, the hypervisor. Whilst Cloud computing exhibits big data processing at the data center level, fog computing provides data processing and actuation capability at the network edges [6, 7]. Also, Fog computing expand the same capability in the middle at edge gateways. In another word, fog computing provides the closed resources to many services, which cannot be realized with alternative strategies [8, 9].

The scattered IoT devices create the need to spread the fog nodes to cover the IoT environment. As of late, mist computing has been rise to capture a more extreme edge [10]. In other words, the nodes in the fog environment are classified as mist and middle edge node, as shown in Fig. 1. The mist computing model depicts scattered computing at the extreme edge. It has proposed with future self-aware and autonomic systems in mind [11]. The Mist server can exist with the Internet Service Provider (ISP) or separately in the network. Mist computing is proposed as the first computing node in the IoT-fog-cloud colony; it can be called as “IoT computing” or “things computing”. An IoT device may be portable like smart watch, a mobile device, or stationary like a smart AC.

Generally, the Load balancing seems to assume an imperative for scheduling the various types of the users’ tasks. Load balancing can be characterized [12] into different categories such as the applied state that maybe static or dynamic, the load balancer techniques which is hardware or software and the policies rules such as resource, information, selection, location and transfer. The workload in the static load balancing approach is based on the current performance of the processing nodes with careless about future changes. Moreover, in this approach the waiting tasks can't migrate from its processing nodes [13]. Also, the static load balancing methods treat the tasks in non-preemptive manner. Otherwise, the dynamic load balancing decides the tasks distribution during the run time based on the information of system status [14]. In this way, the task scheduling algorithm is employed to reserve the resources to the IoT devices on servers to satisfy the fair distribution. The satisfaction of the fairness will reduce the task waiting time. Furthermore, it will enhance the tasks execution speed using the free resources and optimum consumption of storage to minimize the turnaround time of the submitted tasks.

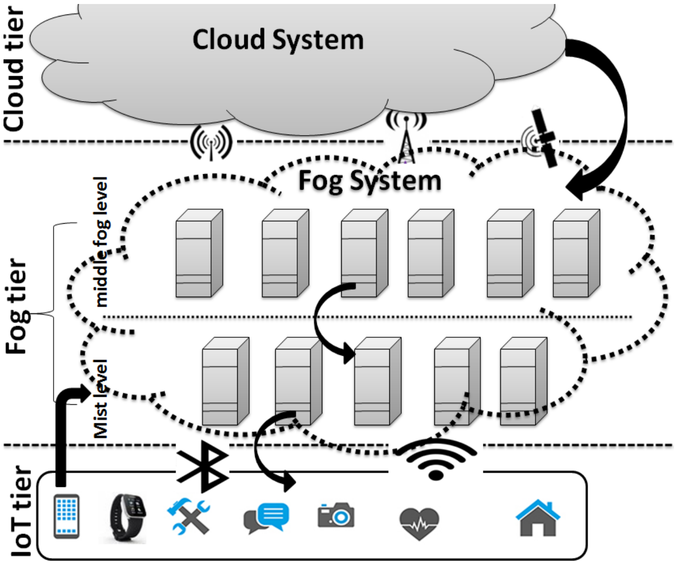


Fig. 1. IoT with Edge Computing and Cloud.

The proposed model in this paper is based on a dynamic load balancing algorithm. This model gives the real tasks the first priority to fit its deadline. Also, the real tasks can migrate from mist node to the others to avoid missing the deadline time. On other hand, the system preserves a specific quality of service (QoS) for each type of soft task requests. Moreover, the proposed load balancing algorithm is acting as a wind flow. It migrates the load of service requests from the high load nodes to the low load nodes. This strategy minimizes the communication overhead in spite of the user task migration. Hence, each node cooperates with the others nodes to maintain balanced load among them.

In the following, the rest of the paper is organized as follows. Section II; discuss the related work of the load balancing algorithms and techniques that are proposed for working with the cloud systems. In section III, the architecture of the proposed model is presented. In addition, the performance evaluation and the results of the simulations are introduced in sections IV and V. Section VI conclude the paper and provide the venues for the future work.

II. Related Work

In this segment, Several Load Balancing algorithms are presented for different authors. These algorithms are investigated dependent on the diverse parameters, for example, due date, execution time, data transmission, cost, need, dependability, adaptability, task length and throughput. Basically, the effective load balancing algorithms have been implemented in the cloud system.

Generally, the load balancing mechanisms in both of the cloud and Fog is same with just principle distinction. In the fog computing, the load balancing should maintain system more feasible and effective with in spite of resources limitation. It offers access to the assets of less transmission capacity and time. In this way, the mist computing has fulfilled the requirements for the closest IoT at a gigantic rate with no disarray like what may happen for the network traffic.

In this area, the first load balancing technique is introduced in [15]. This method is intended to achieve good services by increasing the resource utilization based on two parameters, which are the task priority and its length. The choice of the tasks for the scheduling might be gotten from both of the first and last indexed queue to accomplish an all the more relentless framework.

The tasks are scheduled dependent on the total credit system sponsored from grouping of credit length computed from task length and credit priority computed from the task priority. Finally, the priority of processing is given to the high credit task. However, this algorithm suffers from certain shortcomings when the absolute credits of several tasks became indistinguishable. For this situation, the FCFS has to be added without guarantee of tasks to be completed earlier or to its deadline.

Another algorithm depends on comparable to conduct of honey bee model (HBB-LB) is proposed by Dhinesh babu L.D et al. [16]. In this algorithm, the priority is taken as a fundamental QoS factor to Bar any procedure from hanging tight for quite a while in the line to diminish the execution time and augment the throughput. Similarly, the tasks can be acted as the Honey bees and the Virtual Machines can be acted as sustenance sources. Moreover, The VMs are classified according to three circumstances, balanced overload, high overload and low overload. When the VMs are overloaded, the tasks are evacuated and act as a honey bee. So, these tasks are migrated to the low load VMs. These duties are depending on how many high priority tasks are executed on those VMs. It should be noticed that the VM is chosen based on the low overload and the least number of the executed priority tasks. After proper tasks on VM, data is refreshed with the goal that the rest of the assignments can acquire their needs under load VM. This algorithm has presented certain advantages represented in the proper resource utilization; maximizing the throughput while keeping different QOS parameters which are built on the task priority. On the other hand, the disadvantages are introduced for the low need priority tasks which suffer from idle state or long time waiting in the queue. These tasks may be dismissed causing the unbalancing of the workload balancing.

For an enormous scale condition, e.g., cloud computing framework, there had been also various scheduling approaches proposed with the objective of accomplishing the better task execution time for cloud resources [17]. Independent task scheduling algorithms mainly include MCT algorithm [18], MET algorithm [15], MIN-MIN algorithm [15], MAX-MIN algorithm [19], PMM algorithm, and genetic algorithm. The MCT (Minimum Completion Time) algorithm assigns each task in any order to the processor core that causes the task to be completed at the earliest time. It prohibits some tasks to be allocated to the fastest processor core. The MET (Minimum Execution Time) algorithm allocates each task to a processor core in any order that minify the task execution time. As opposed to the MCT algorithm, the MET algorithm does not consider the processor core's ready time, which may prompt genuine burden unevenness crosswise over processor cores. The MIN-MIN algorithms calculates the minimum completion time of all unscheduled tasks firstly, and then chooses the task with the minimum turnaround time and allocate the task to the processor core that can minimize its turnaround time, repeating the process many times until all tasks are allocated. The same as the MCT algorithm, the MIN-MIN algorithm is also based on the minimum completion time. The MIN-MIN algorithm proposes all tasks that are not scheduled, but the MCT algorithm considers unique task at a time. The MAX-MIN algorithm is similar to the MIN-MIN algorithm, which also computes minimum completion time without scheduled tasks firstly and then selects the task with the largest minimum completion time and assigns the task to the processor core with the minimum completion time.

Mondala et at. use an optimized approach algorithm to have load balancing scheduling system [20]. This model is based on a centralized load balancing algorithm. In another words, the system is based on a central node that distributes the workload tasks. Hence, the main drawback is of this model is that if the central node fails, the whole working of the system will fail. This means that the central node is represent the system bottleneck. So here, using decentralized load balancing strategy solves this bottleneck. Resource utilization can be done effectively to enhance the throughput, accordingly decreasing the cost of an application running in a SAAS environment without break service level agreements [21].

Actually, the different scheduling algorithms based on QoS parameters have been introduced for different environments in [22]. The scheduling is performed to achieve the huge service requests and to enhance the efficiency of the workload. Subsequently, there are numerous modules that are implemented in each kind of the scheduling algorithms, for example, Min-Min, FCFS, Max-Min, Round-Robin algorithm.

Nevertheless, the one of the efficient methods among them is the heuristic method. Its allocating the tasks includes three stages in a cloud computing. At first, the VMs are located. Hence, the best target VM is chosen. At last, the task is assigned to the target VM. Lately, the Real Efficient Time Scheduling (RETS) is investigated in [23]. The main goal of RETS is to process the real-time tasks without delay. Therefore, it keeps one tenth of the available resources for the real-time tasks. Although, this ratio can be insufficient if the real-time tasks exceed this ratio. On the other hand, one tenth of the available resources will be idle if there are no real-time tasks.

Moreover, Anju et al. introduces multilevel of priority-based task scheduling algorithm (PBATS)[24, 25]. This algorithm has three levels of priorities, which prioritizes the tasks based on the length of the instructions. Also, to enhance performance of PBATS, it migrates the tasks under the minimum migration time policy. This policy can cause overload of node, which has low network overhead. Also, this policy doesn't distinguish between the real and soft tasks.

Also, Wang et al. proposed a task scheduling algorithm in the fog computing, which is called "hybrid heuristic (HH)" algorithm [26]. HH algorithm is mainly focus in solving high energy consumption in case of using limited computing resources. Unfortunately, HH method isn't distinguish between the mist and middle fog nods. Hence, this algorithm is not efficient method for real-time services.

III. Proposed Model

In a Fog computing environment, the load balancing is a pivot point for effective and efficient resource utilization, bandwidth and to achieves desired quality of service (QoS). Fog Computing system is divided virtually in two type of nodes, namely; mist and middle edge node. Actually, both types of fog nodes can have the same structure and resources. Nevertheless, the most closed node to IoT is called mist. Each Mist computing server is centered in the specific location mainly to receive the clients or/and IoT requests in a specific region. The fog colony is connected to a cloud system in the case of fog resources shortage to overcome the fulfillment of task requests.

In this paper, the new scheduling model (DLBS) is proposed in the cloud-fog-mist environment. The structure of this model is shown in Fig. 2. First of all, the Service Listener (SL) receives the user/IoT service request. Hence, SL creates a task for the service request and sends it to Load Balancing Allocator (LBA) module with required software from service container. Also, SL send task-metadata like, task type (real time or soft), expected execution time, etc. So, each Mist server is supplied by its own Load Balancing Allocator module (LBA). LBA is responsible for allocating the clients and/or IoT service requests into the fog resources. There two types of user/IoT request; real time and soft-tasks. The proposed model is designed to handle both types of tasks.

Mist node gives the real time task queue in resources allocation. The tasks in the real time queue will be allocated into one of idle local VMs in the node. If there is no idle, LBA preempt one of soft task VMs. In the worst case scenario, if there are no idle or soft VMs, ***Fog explorer*** module suggest the resources in the closest mist/middle edge node. Fog explorer detects the status of the other fog node by getting the status flags. The status flags are set by LBA module and broadcasted by the fog explorer. Each Mist node has four types of status flags, which determine the status of the node, namely, ***load lock***, ***real task lock***, ***receive status***, and ***send status***. Load lock flag, which is soft task waiting, is set by zero if the expected waiting time will not exceed QoS threshold (λ). In another word, λ grantees that the service of the soft tasks will be provided in a reasonable delay. If load lock flag is set by one, this fog node can't receive a soft task from other fog and its soft tasks will migrate outside the node. Also, real-time task lock is set by one if all VMs are allocated by real-time tasks. For any fog node if one of VMs is processing a soft task, the real-time task lock is set by zero. Finally, according to task migration the fog node blocks the receiving tasks from other nodes if its receive status or send status has value one. Obviously, the status flags are used to maintain the system balanced and available.

Mist node gives the real time task queue in resources allocation. The tasks in the real time queue will be allocated into one of idle local VMs in the node. If there is no idle, LBA preempt one of soft task VMs. In the worst case scenario, if there are no idle or soft VMs, ***Fog explorer*** module suggest the resources in the closest mist/middle edge node. Fog explorer detects the status of the other fog node by getting the status flags. The status flags are set by LBA module and broadcasted by the fog explorer. Each Mist node has four types of status flags, which determine the status of the node, namely, ***load lock***, ***real task lock***, ***receive status***, and ***send status***. Load lock flag, which is soft task waiting, is set by zero if the expected waiting time will not exceed QoS threshold (λ). In another word, λ grantees that the service of the soft tasks will be provided in a reasonable delay. If load lock flag is set by one, this fog node can't receive a soft task from other fog and its soft tasks will migrate outside the node. Also, real-time task lock is set by one if all VMs are allocated by real-time tasks. For any fog node if one of VMs is processing a soft task, the real-time task lock is set by zero. Finally, according to task migration the fog node blocks the receiving tasks from other nodes if its receive status or send status has value one. Obviously, the status flags are used to maintain the system balanced and available.

Example in Fig. 3 shows the closer fog region for Mist Y by dotted line, and the closer region for middle edge node C by the dashed line. In this example, Mist Y receives two service requests from IoT devices. The first request is real- time request, which come from Pacemaker device. This type of request is classified by the Fog Explorer as real-time request. Hence, this request must be handled in the local fog (nod Y). On the contrary, Mist Y is forwarding the soft request to middle edge server C. Also, for the node C the load is migrating to D. This strategy makes the load spread over all system nodes.

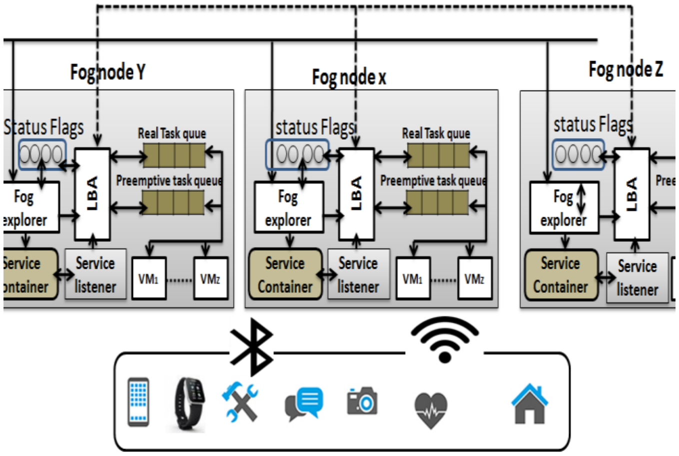


Fig. 2. DLBS Model.

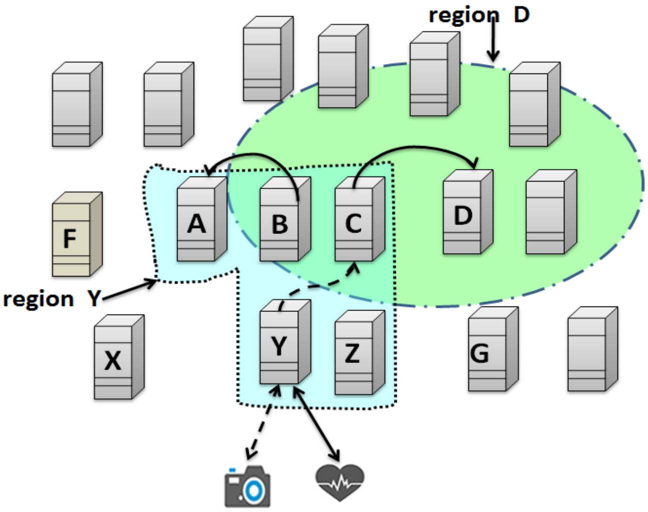


Fig. 3. DLBS Node Region.

A. *Load Balancing Allocator (LBA)*

The main objective of LBA is to allocate the task requests, which is received by Service Listener (SL). Also, LBA should allocate the real-time tasks to be executed before they met their deadline. Also, it guarantees an efficient response time for the soft tasks. LBA is maintaining to allocate the real-time task trivial waiting time. This accomplished by allocate the real-time task locally or to allocate the task in one of the closed server. In case of soft service is requested, the soft task should be exceeded waiting time threshold (λ). To maintain this condition, the following steps should be computed. First, the total expected execution time of the soft-waiting tasks can be computed as follows.

$$exe_{tot} = \sum_{\forall i} exetime(t_i)$$

Also, the total processing power of the mist node can be formulized as summation of MIPS (million instructions per second processor) for all VMs.

$$p_{tot} = \sum_{\forall j} MIPS(VM_j)$$

Hence, the total expected waiting time should not exceeding λ by achieving the follows equation.

$$w_x = \frac{exe_{tot}}{p_{tot}} + \delta < \lambda$$

Where, δ is the constant depending on the ratio of the real-time tasks μ and the average size of real-time task services ε .

$$\delta = \mu(\frac{\varepsilon}{p_{avg}})$$

Where, p_{avg} is the average of VMs processing power of the mist node. In the following the pseudo code of LBA function is introduced. The following algorithm represents the general steps of load balancing allocator procedure.

Load Balancing Allocator (LBA) Algorithm

```
Input:
  $t_k$ // receive task from the service listener or from other LBA

If ($t_k.type = real$)
  freeVM = findIdleVM() // find the idle VM
  If ($idel\_VM \neq \phi$)
    allocateTask(freeVM ,$t_k$)
  ElseIf(realTaskLock = 0) // all machines are busy in soft tasks
    freeVM = PreemptSoftTask()
    allocateTask(freeVM ,$t_k$)
    If all VM.tasks=real
      realTaskLock = 1
  Else // all VMs are busy in real tasks
    $VM_x = Min_{\forall i}(RemainTime(VM_i))$ /* find a VM with
                                the minimum remaining time */
    $t_k.turnaround = t_k.expectExeTime + VM_x.remainTime()$
    If ($t_k.turnaround \le t_k.deadline$)
      allocateTask($VM_x$ ,$t_k$)
    Else //find VM in the closest Mist node
      $F_R$ =FogExplorer.getFog(RealTask) //find closest unlock fog for real task
      SendStatusFlag=1
      SendRealTask($F_R, t_k$)
      SendStatusFlag = 0
    End if
  End if
Else //the second case; soft task type
  $exe_{tot} = \sum_{\forall i} exetime(t_i)$ /* Compute the total expected
                execution time of the soft-waiting tasks */
  $w_x = \frac{exe_{tot}}{p_{tot}} + \delta$ // Compute the total expected waiting time
  If($w_x < \lambda$)
    InSoftQueue($t_k$)
  Else
    loadLock=1
    $F_R$ =FogExplorer.getFog(SoftTask) /* find closest unlock
                                fog for soft task*/
    SendStatusFlag=1
    SendSoftTask($F_R, t_k$)
    SendStatusFlag = 0
  End if
```

```
3  End if
```

B. Fog Explorer, Service Container and Flags

Finally, fog explorer module is responsible for determine the closer fog region for each node. This region is defined as set of nodes which has minimum communication overhead. If any of the status flags is change in each node, the node broadcast this information to its closed region. Also, Fog explorer is responsible for broadcasting a copy of the Service Container to all fog and mist computing nodes. Moreover, it should send up-to-date a copy of additional changes in Service Container.

IV. Simulation Setup

As a mist landscape, we propose a fog-mist colony of 100 nodes. Half of the colony nodes are mist nodes, which receive the user requests. The fog-mist colony is connected to a cloud system in circumstance of shortage in the fog-mist sources to the fulfillment of soft task requests. Of these 100 fog-mist colony, 10 are concurrently issuing 1,000 task requests to the mist colony. Furthermore, IoT applications are characterized by two types (real and soft).

The proposed DLBS algorithm, have been implemented on simulator CloudSim [27, 28] 3.0.2 to execute tasks along with Window 7 OS, core i5 2.3 GHz processor and NetBeans IDE 7.2.1. CloudSim computes the execution time of a service request to fulfill a task requirement, hence computes the waiting time for soft task by aggregating the number of instructions necessary to execute the waiting soft tasks. In this experiment, the soft-task request and real-task requests required 0.05, and 0.03 million of instructions per second (mips) respectively. Both task types have 300 MB of incoming and 300 MB of outgoing data. Fog/Mist nodes able to able to handle 250 MIPS. Each fog node can create 10 VM's have the processing power 500 MIPS. The bandwidth between fog nodes is set to 100 Mbit/s, and between the cloud and fog nodes to 10 Mbit/s. All experiments are repeated for 10 times and the mean values are taken.

DLBS model is compared with four models. The first model is FCFS, which serve the tasks based the arrival time. Moreover, the others compared models was created for the cloud computing system, namely the Max-Min, the PBATS and the RETS. The Max-Min maintains a task status table to envision the real loads of the VMs and the evaluated finishing time of tasks, which can distribute the workload among nodes [29]. The Priority Based Autonomic Task Scheduling (PBATS) that schedule its tasks according to three different priorities levels [25, 30]. Furthermore, the Real Time Efficient Scheduling (RETS) depends on reserving a one tenth of the resources for the real-tasks [23]. All these scheduling techniques are matched by the proposed techniques to evaluate the load balancing in the proposed model.

V. Results and Discucion

The performance evaluations have been performed in three dimensions. The first dimension evaluates the performance of the system on the soft-tasks load. On another hand, the second dimension measures the system reliability for the real-time tasks. The performance evaluation based on three parameters, namely; turnaround time, the average waiting time and the throughput. Finally, the third diminution measures the suitability of the model for the real-time services by evaluating the number of failed tasks in the compared algorithms.

This section is organized into three subsections. Each subsection is concerned to evaluate a performance dimension. Hence, the following subsection evaluates the performance of the system using all types of tasks. Moreover, the second subsection evaluates the effect of the system on the real time tasks only. Finally, the failure in the real-service requests is measured in third Subsection.

A. System Performance using Real-Time and Soft Service Requests

In this section three tests are done. The first test measures the response time of variant number of tasks. The second test evaluate the waiting time of the system. Finally, the last test in this section measures the throughput.

1) Turnaround time performance test: The first experiment measure the system performance based on the Turnaround time parameter. DLBS is compared with previous mentioned four algorithms. The experiments are done using different number of workloads from 1000 to 10,000 tasks. The real time tasks will represent 20% from all of the inserted workload in each experiment. Obviously, we can notice that the FCFS curve is rapidly increased by increasing the number of service requests. The bad performance of FCFS is due to the non-preemptive property. Also, Max-Min curve is closed to the FCFS curve. Since, the Max-Min is allocating the longest tasks to VMs which has lest remaining execution time. In another word, in Max-Min scheduling algorithms the short tasks will wait a long time to get the resources, which increase the average of waiting time. In addition, the PBATS curve is keep a less in the average turnaround time results when compared to the FCFS and Max-Min. Indeed, the tasks in the PBATS algorithm are classified into three levels of priorities and underestimate the quality of services. Furthermore, the curve of the RETS refer to acceptable results with a light load up to 1,500 tasks, as shown in Fig. 4(A). Also, RETS gives an inefficient performance if compared by the proposed algorithm (DLBS). The performance of RETS is decreased as increasing the work load. The performance deterioration of RETS algorithm is due to static reservation for the real tasks. It assign one tenth of the resources for the real requests. Reserving a static ratio of the resources can cause problem if there are no proper real tasks. Actually, it is a dilemma if the real tasks exceed the reserved resources. Actually, the DLBS overcome these problems. It gives high priority to the real tasks for satisfy its deadline. Also, it maintains a specific response time for the soft tasks. Subsequently, the DLBS is the most efficient algorithm among all of the compared algorithms in the Mist-fog environment.

2) The waiting time performance test: This experiment measures the waiting time for the service request tasks. As shown in Fig. 4(B), the waiting time of the DLBS curve has the best performance. Moreover, for having a certain QoS the expected waiting time for the soft tasks parameter λ is set by 10 second. Hence, the DLBS curve values are very close to ten second after 5,000 tasks. It is worth noting that the FCFS curve has the worst performance. This bad performance is caused by the same reasons that increase the average turnaround times. Also, the Max-Min curve is the closest one to the FCFS curve. In the PBATS curve the tasks allocation is depending on three levels of priorities, which increase the waiting time for tasks according to their levels. Furthermore, the RETS algorithm has an acceptable performance until the workload less than or equal to 3,000 tasks. Unfortunately, as increasing the services requests, as the average of waiting time is rapidly increased for the RETS. All of these problems have been solved by DLBS algorithm as shown by the performance curve. DLBS maintains an upper bound of the waiting time for each soft task in mist node and send the exceeding load to the closest low load node or to the middle edge node.

3) The throughput performance test: This test measure the performance based on the average of system throughput. The throughput is defined as the total number of finished tasks per time. Additionally, the experiment is done based on the same workload of the past examination. The performance of the compared algorithms is shown in Fig. 4(C). We can notice that, the throughput of DLBS has the best throughputs enhancement compared by the other algorithms. The performance enhancement of DLBS is caused by the balanced distribution of the tasks that satisfy QoS. Also, the worst performance curve is the FCFS. Moreover, the RETS throughput curve is successor to DLBS curve. Since RETS gives the highest priority to the real tasks, which is the lightest processing tasks, then it increases the number of the finished tasks.

B. System Performance using Real-Time Service Requests

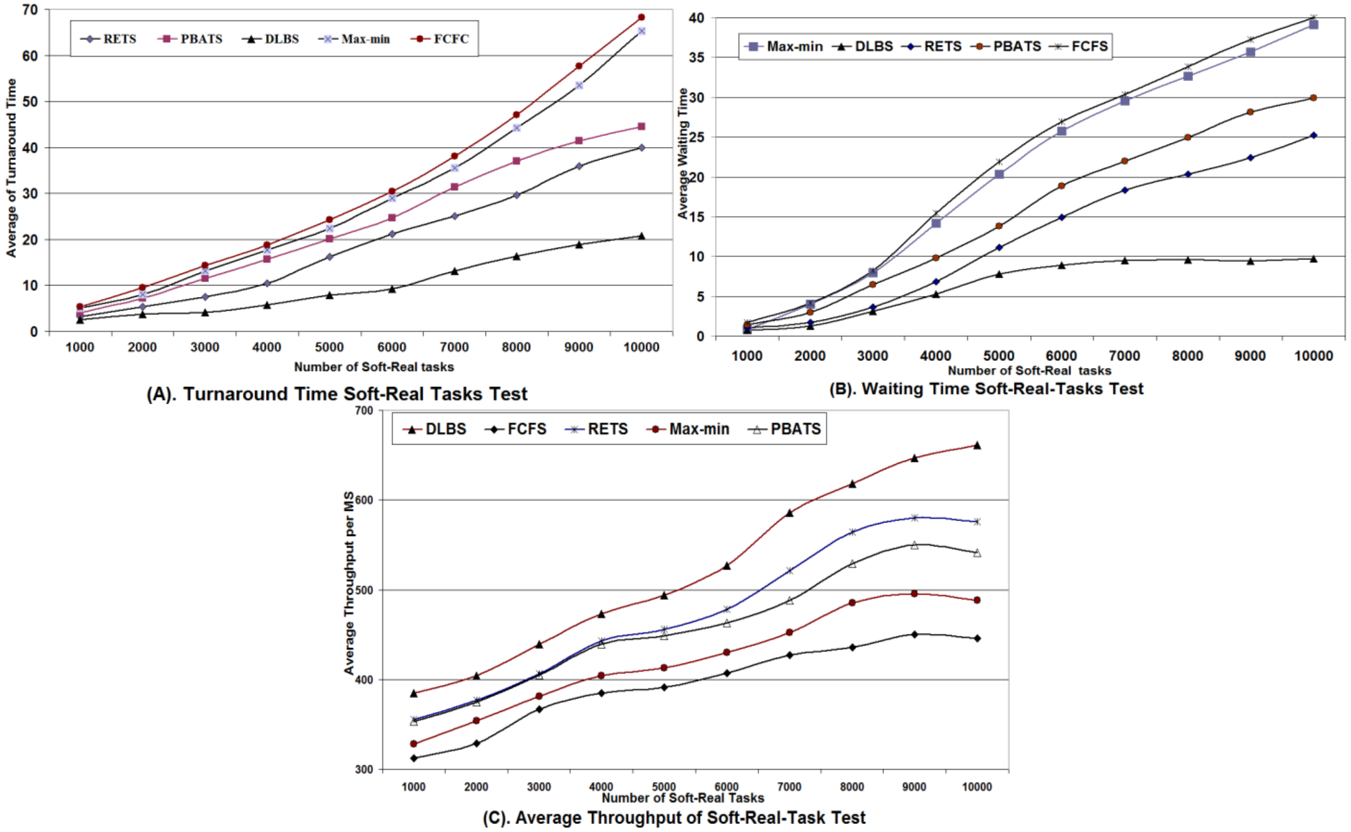
This experiment evaluates the effect of the proposed system on the real time tasks compared with other algorithms. Each experiment is completed on the real task ratio 25% of workload. Through the experiments, the workloads for all the tasks types are changed from 1,000 to 10,000 tasks. Hence, the real time tasks are changed from 250 to 2,500 tasks. However, all the experiments of the Real-Time tasks are performed in the existence of the soft tasks load. This section is organized as follows. The following subsection measures the turnaround time. Subsection (2) measures the waiting time and Subsection (3) measures the throughput. Finally, the Subsection (4) measures the suitability of the system for real service.

1) Turnaround Time performance Test: The turnaround time performance comparison of the compared algorithms is shown in Fig. 5(A). The worst performance is obtained by the curves that represent the FCFS, PBATS and the Max-Min algorithms respectively. The essential shortage of these algorithms is the disability to handle the requests of the real time service according to their deadline. Actually, these algorithms are not indeed to handle the real-time tasks. Hence, the real times tasks are treated as the soft tasks. On other hand, the RETS gives an acceptable performance when the number of the Real-Tasks are not exceed one tenth of the system resources. As obtained the figure, the performance of RETS result is acceptable until 1,000 real-tasks and is decay after this point. Furthermore, the RETS algorithm preserve the response time of the real time tasks to be less than their deadline times. In other words, the real-time tasks are not presented to any postpone, which limits the turnaround time. Moreover, the real-time tasks are migrated from fog node to another one to avoid waiting time.

2) The waiting time performance test: The averages of waiting time curves that expose the impact of the DLBS model on the duration time of the real tasks are shown in Fig. 5(B). In this figure, the lower mean waiting time is implied for DLBS. As mentioned before, the DLBS model is designed to give the first priority for the Real-Time tasks. Hence, the reserved resources for the soft tasks are released to allocate the real tasks. However, the RETS is the closest curve among all the compared algorithms to the DLBS. Unfortunately, as the real requests load in RETS is increased, as the average waiting time is increased. Hence, the deadline times of the real tasks will be exceeded in RETS model.

3) The throughput performance test: The throughputs curves, in Fig. 5(C), show the performance comparison between the competitive algorithms. Unmistakably, the highest throughput is accomplished by DLBS. The RETS throughput becomes consistent after satisfying the reserved ratio of the real tasks. Also, the DLBS throughput is increased as increasing the real time tasks. Since DLBS algorithm can assign the whole mist node resources and borrows additional resources to satisfy the real-time service requests. Moreover, FCFS has the worst performance because it isn't careless about the deadline of the real-time task.

4) Real-time task failure test: To judge about the suitability of the algorithm for real time services, the task failure should be concerned. To judge about the suitability of the algorithm for real time services, the task failure should be considered. Fig. 5(D) measures real task failure for the proposed and the compared algorithms. The number of task failure for the DLBS model is trivial if compared with the other models. RETS model gives a good performance in low load of the real time tasks. Unfortunately, RETS model doesn't supports flexibility in the reserved resources for the real time tasks. Also, it doesn't support task migration to provide the desired resources. The other algorithms failure values indicate inaptitude for real time services.

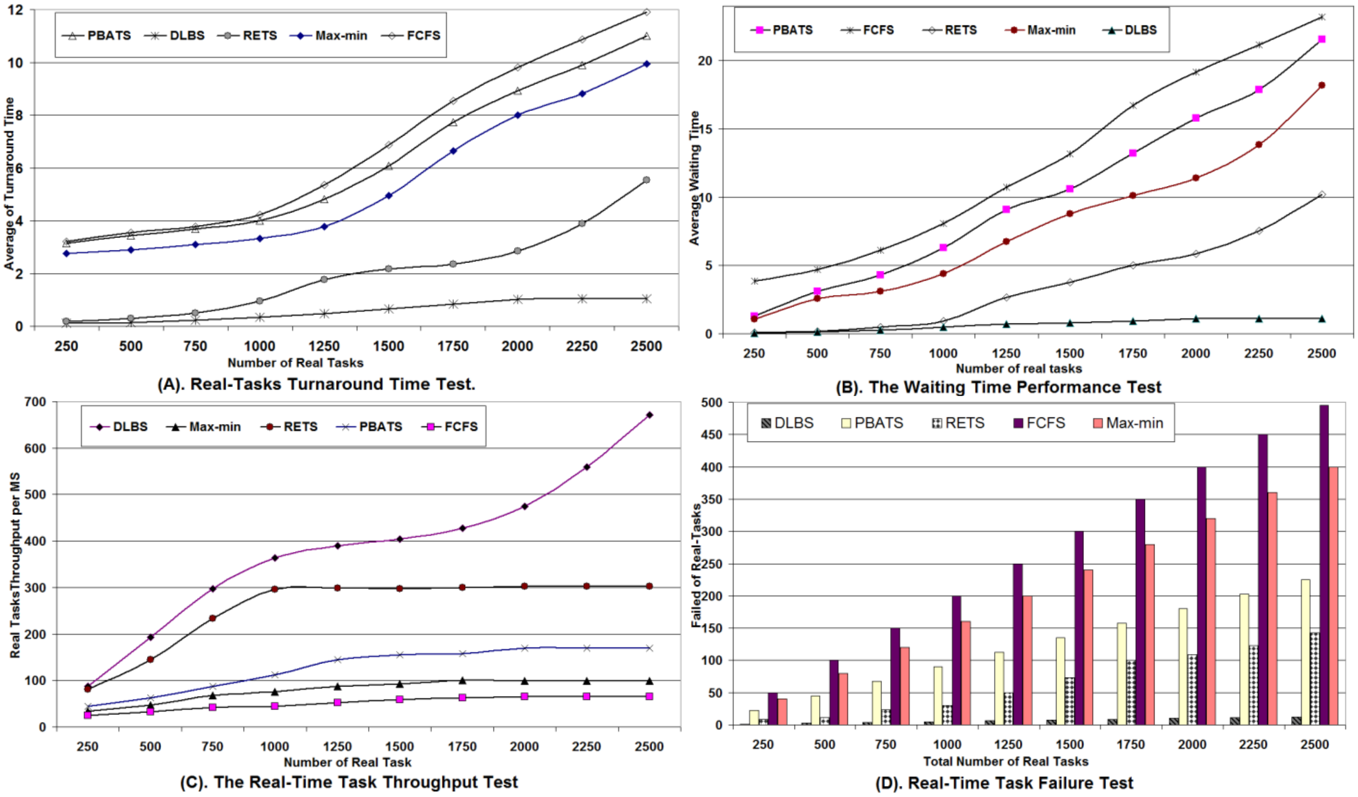


(A). Turnaround Time Soft-Real Tasks Test

(B). Waiting Time Soft-Real-Tasks Test

(C). Average Throughput of Soft-Real-Task Test

Fig. 4. Performance Comparision using Soft and Real-Time Service Requests.



(A). Real-Tasks Turnaround Time Test.

(B). The Waiting Time Performance Test

(C). The Real-Time Task Throughput Test

(D). Real-Time Task Failure Test

Fig. 5. Real-Tasks Turnaround Time Test.

VI. Conclusion and Future Work

In this paper, DLBS model is designed for managing soft and real time services in fog computing environments. The DLBS model introduces decentralize scheduling algorithm. Fog computing consists of two type of nodes, namely; mist and middle edge nodes. Mist nodes are the closer nodes to IoT devise, which receive its services requests. The DLBS model provides an efficient solution for having IoT service response time. This model is providing an efficient load balancing strategy for IoT service requests. Also, this model manages the IoT services requests load for each fog node in decentralize manner. The decentralize load management avoids the bottleneck problem, which exists in the majority of the other solution. Moreover, this model is designed to fit the real-time serves requests. The experiments show that our methods outperform the compared methods. In future work, this model will be developed to manage the heterogeneous Mist nodes.

References

[1] Chandrasekhar S. Pawar, Rajnikant B. Wagh, Priority Based Dynamic resource allocation in Cloud Computing with modified Waiting Queue, Proceeding of the IEEE 2013 International Conference on Intelligent System and Signal Processing(ISSP) Pages 311-316.

[2] Yusen Li, Xueyan Tang, Wentong Cai, Dynamic Bin packing for on demand cloud resource allocation, Proceedings of the IEEE Transactions on Parallel and Distributed Systems ,2015,Paged 1-14.

[3] Savani Nirav M, Prof. Amar Buchade,―Priority Based Allocation in Cloud Computing, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV3IS051140 Vol. 3 Issue 5, May - 2014.

[4] Brototi Mondala, Kousik Dasguptaa, Paramartha Duttab"Load Balancing in Cloud Computing using Stochastic Hill Climbing-A Soft Computing Approach", Elsevier, Procedia Technology 4(2012) pp. 783 –789.

[5] Ivan Stojmenovic, sheng Wen, "The Fog Computing Paradigm: Scenarios and security issues" Proceedings of the IEEE International Fedrerated Conference on Computer Science and Information Systems, 2014, pp.1-8.

[6] Mahmood A., Zen H. (2018) Toward Edge-based Caching in Software-defined Heterogeneous Vehicular Networks. In: Mahmood Z. (eds) Fog Computing. Springer, Cham. https://doi.org/10.1007/978-3-319-94890-4_13.

[7] Sari A. (2018). Context-Aware Intelligent Systems for Fog Computing Environments for Cyber-Threat Intelligence. In Fog Computing (pp. 205–225). Cham: Springer. 10.1007/978-3-319-94890-4_10.

[8] Nanxi Chen, Yang Yang, Tao Zhang, Ming-Tuo Zhou, Xiliang Luo, John K. Zao, "Fog as a Service Technology", Communications Magazine IEEE, vol. 56, no. 11, pp. 95-101, 2018.

[9] Luthra, M., Koldehofe, B. & Steinmetz, R. "Transitions for Increased Flexibility in Fog Computing: A Case Study on Complex Event Processing" Informatik Spektrum (2019). https://doi.org/10.1007/s00287-019-01191-0.

[10] [7] A. Davies, Cisco pushes IoT analytics to the extreme edge with mist computing. [Online]. Available: http://rethinkresearch.biz/articles/cisco-pushes-iotanalytics-extreme-edge-mist-computing-2, Blog, Rethink Research.

[11] J.S.Preden, K.Tammemäe, A.Jantsch, M.Leier, A.Riid, E.Calis, The benefits of self-awareness and attention in fog and mist computing, Comput. (Long Beach Calif) 48(7)(2015)37–45.

[12] Flavio Bonomi, Rodolfo Milito, Jiang Zhu, Sateesh Addepalli "Fog Computing and its Role in the internet of things", http://conferences.sigcomm.org/sigcomm/2012/pa per/mcc/p13.pdf.

[13] Manisha Verma, Neelam Bhardwaj Arun Kumar Yadav," An architecture for load balancing techniques for Fog computing environment", International Journal of Computer Science and Communication, Vol. 8 • Number 2 Jan - Jun 2015 pp. 43-49.

[14] S. F. El-Zoghdy and S. Ghoniemy, "A Survey of Load Balancing In High-Performance Distributed Computing Systems", International Journal of Advanced Computing Research, Volume 1, 2014.

[15] Mohsen and Hossein Delda, "Balancing Load in a Computational Grid Applying Adaptive, Intelligent Colonies of Ants", Informatica 32 (2008) 327–335.

[16] Brototi Mondala, Kousik Dasguptaa, Paramartha Duttab"Load Balancing in Cloud Computing using Stochastic Hill Climbing-A Soft Computing Approach", Elsevier, Procedia Technology 4(2012) pp. 783 –789.

[17] W. Lin, C. Zhu, J. Li, B. Liu, and H. Lian, "Novel algorithms and equivalence optimisation for resource allocation in cloud computing,"International Journal of Web and Grid Services,vol. 11, no. 2, pp. 69–78, 2015.

[18] M.Maheswaran,S.Ali,H.J.Siegel,D.Hensgen,andR.F. Freund, "Dynamic mapping of a class of independent tasks onto heterogeneous computing systems,"Journal of Parallel and Distributed Computing, vol. 59, no. 2, pp. 107–131, 1999.

[19] T.D.Brauny,H.Siegely,N.Beckyetal.,"AComparison Study of Static Mapping Heuristics for a Class of Meta-tasks on Heterogeneous Computing Systems,"parallel & distributed computing, vol.61, no.6, pp.810–837,2001.

[20] Brototi Mondala, Kousik Dasguptaa, Paramartha Duttab"Load Balancing in Cloud Computing using Stochastic Hill Climbing-A Soft Computing Approach", Elsevier, Procedia Technology 4(2012) pp. 783 –789.

[21] Atul Vikas Luthra and Dharmendra Kumar Yadav,"MultiObjective Tasks Scheduling Algorithm for Cloud Computing Throughput Optimization", International Conference on Intelligent, Communication & Convergence, Procedia Computer Science 48(2015) 107- 113.

[22] Mohamed A. Elsharkawey, Hosam E. Refaat,"CVSHR: Enchantment Cloud-based Video Streaming using the Heterogeneous Resource Allocation", International Journal of Computer Network and Information Security (IJCNIS), Vol.9, No.9, pp.1-11, 2017.DOI: 10.5815/ijcnis.2017.09.01.

[23] M.Verma, N. Bhardwaj and A. Kumar, "Real Time Efficient Scheduling Algorithm for Load Balancing in Fog Computing Environment",I.J. Information Technology and Computer Science, April, 2016, 4, 1-10.

[24] B.Anju and C.Inderveer (2016), "Multilevel Priority-Based Task Scheduling Algorithm for Workflows in Cloud Computing Environment". In Proceedings of International Conference on ICT for Sustainable Development: Volume.

[25] Swati Agarwal, Shashank Yadav, Arun Kumar Yadav,"An Efficient Architecture and Algorithm for Resource Provisioning in Fog Computing", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.8, No.1, pp.48-61, 2016. DOI: 10.5815/ijieeb.2016.01.06.

[26] Wang J, Li D. Task Scheduling Based on a Hybrid Heuristic Algorithm for Smart Production Line with Fog Computing. Sensors (Basel). 2019;19(5):1023. Published 2019 Feb 28. doi:10.3390/s19051023.

[27] W. Chen and E. Deelman,―Workflowsim: A toolkit for simulating scientific workflows in distributed environments, in 2012 IEEE 8th International Conference on E-Science, ser. eScience, 2012, pp. 1–8. [Online]. Available:https://github.com/WorkflowSim.

[28] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya,―CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms, Software: Practice and Experience, vol. 41, no. 1, 2011.

[29] Xiaofang Li, Yingchi Mao, Xianjian Xiao, "An improved Max-Min task-scheduling algorithm for elastic cloud", Computer, Consumer and Control (IS3C), 2014 International Symposium on.

[30] B.Anju and C.Inderveer (2016), "Multilevel Priority-Based Task Scheduling Algorithm for Workflows in Cloud Computing Environment". In Proceedings of International Conference on ICT for Sustainable Development: Volume.

Decision Making Systems for Managing Business Processes in Enterprises Groups

Ali F. Dalain
Human Resources Management Department
University of Jeddah, College of Business, KSA

***Abstract*—In the current economic realities, the forms of integration business entities through the creation of enterprise groups (EGs), reorganized from industry structures or created a new by acquiring existing companies, are becoming increasingly relevant. The economic activity of the enterprise is carried out in the conditions of economic instability and improvement of the system economic relations, which imposes fundamentally new requirements in the sphere of managing the interaction of enterprises. Under these conditions, the successful development of the enterprises and often their very existence depend both on the effective use of the management systems themselves and on the competence of the management decisions made. Consequently, for decision makers and managers of Group Policy (GP), the problem of evaluating the development of GP and promptly making sound management decisions in an unstable and rapidly changing economic environment is considered a particular relevance. One of the promising ways to solve this problem is the development of decision support systems (DSS), using scientifically based decision-making methods based on modern mathematical apparatus and computer equipment. At present, the approach to managing the development of the EGs is associated with the representation of the latter as a multi-agent system (MAS). The DSS does not replace, but complements the existing management systems in the EGs, interacting with them, and uses in its work information about the functioning of EGs units.**

Keywords—Management systems; decision support systems; multi-agent systems; group policy; enterprise groups

I. Introduction

At all stages of the economic development of EGs, the most important problem in the activities of enterprises is the problem of increasing the competitiveness of their products, which can be achieved both by modernizing production and by optimizing the interaction of enterprises in EGs [1],[2]. One of the priorities is the development and use of DSS in the management of business processes EGs.

Under the management of business processes we can smoothly understand the system of targeted impacts, in which, by means of decision-makers, decision makers implement measures to improve the efficiency of the EGs. Several types of impacts in this research are considered: the selection and implementation of investment projects (IP) [3], ensuring the production of competitive products; optimizations of the system parameters interaction with agents in MAS for the manufacture and sale of products; restructuring of EGs units [4],[5].

The survival of industrial enterprises in conditions of economic instability often directly depends on a successful management of business processes. At the same time, considerable difficulties appear in the EGs already at the initial stage of modernization of management [6]. This is due to the fact that most of the existing traditional software for managing the development of enterprises are built on the classical principles of budgeting - control and are not sufficiently effective at present to manage the business processes of EGs. In addition, the transition from the designated strategy for the development of EGs to specific actions of the performers is sometimes difficult to implement due to the lack of a regular unified mechanism that would set the necessary priorities, allow preparing and evaluating solutions, analyzing the distribution of IP across EGs [7], and controlling the consistency and efficiency of execution in IP and also supported the possibility of joint decision-making on a number of current issues [8].

However, it should be noted that the state of affairs in the field of DSS applications to ensure the effectiveness of business process management does not sufficiently meet the needs of the enterprise in modern conditions, and there are a number of scientific problems that require systemic solutions [9]. Among them, it can be noted: the need to develop a decision-making methodology in determining the optimal control regimes for the interaction of industrial enterprises in the EGs; the need to develop the principles of information and analytical support for DSS when optimizing the management systems of a EGs [10], the lack of common models for selecting and implementing IPs in several EGs under the direction of the managing company (MC). Therefore, the development and applications of DSS in the management of business processes and EGs is currently a pressing and timely task.

The aim of the work is a systematic analysis of information processes for managing the activities of EGs, the development of DSS to increase the efficiency of the implementation in IP, as well as the optimization of teamwork (interaction) of enterprises belonging to the EGs. To achieve these goals, the main research methods used were: system analysis, automatic control theory, decision theory, structural and dynamic analysis, system modeling, numerical methods, nonlinear programming [9].

Practical values of this research are considered the applications of the developed models in the management system of joint business processes in EGs allows taking into account the change in the performance indicators of each industrial enterprises over time of their joint work [7],[11]. This makes it possible to conduct an analysis of the stability in the functioning of EGs and to take control decisions at the optimal points in time to obtain maximum values of the functional quality of each enterprise [12].

II. Research Framework

Modern concepts of managing complex distributed systems in various industries are based on the man-machine organization of management processes, in which the role of decision-making is assigned to a person, and the machine provides information support for the stages of generating and generating alternative solutions [13]. It is noted that the use of this approach is a necessary measure to overcome a high level of uncertainty in the formulation and conditions for solving problems of managing complex, nonlinear and dynamic objects, such as EGs. At the same time, the effectiveness of the control systems of such EGs is largely determined by the subjective properties of the decision maker (DM) operating in the control loop, which in turn requires it to have a high level of competence both in managing business processes of the EGs and in the resulting problem situations [14],[15].

Currently known and widely used method of overcoming subjectivity and increasing the level of competence of decision makers is the use of DSS. DSS provides the correctness of solving problems by choosing rational options for managing business processes through the integrated use of a set of technologically interconnected services that implement traditional and advanced information technologies.

DSS can be represented as a set of management subsystems operating in the information environment in accordance with established information flow procedures, which determine the sequence of business process implementation steps, as well as methods of transmitting, storing and processing information in the management process [16],[17]. Consequently, one of the main conditions for the effective functioning of the DSS is a constant analysis and processing, establishment of links and ways of circulating information flows between information sources and receivers, which provides an integrated approach to the development and implementation of management decision options for decision makers.

For management of enterprises known and well-established in practice is an approach based on the use of multi-agent DSS, which combines various methods of analyzing situations and forming management decisions implemented by agents of MAS [9]. It is shown that multi-agent technologies can be considered as the basis for creating effective methods and tools for developing business process management systems. The issues of creation and application of DSS in situations related to the optimization of the interaction between agents in the network structures of the MAS and hierarchical structures are considered. The most important components of such systems, which include subsystems for modeling business processes, are highlighted [18].

An analytical review of the DSS structures and information processing methods in them is given. The most popular information processing systems, their functionality, architecture, software platforms, design principles, etc. is considered. The analysis of preliminary data preparation systems for making management decisions based on ERP, IDEF, MRP, ERPII standards was carried out. The advantages and disadvantages of replicable systems of this class are noted, including the complexity, duration and cost of implementation, and examples of implemented DSS for various fields of activity are given. It was concluded that there are no systems applicable for the purposes of managing the business processes of the EGs, taking into account the optimization of the modes of cooperation between the MAS agents [19],[20].

In the next stage of this research, business process management models are developed based on the selection and implementation of IP in DSS. The proposed management models are conceptual in nature and reflect the most significant aspects of the joint work of the EGs [6],[14]. In this case, the main focus is on determining the basic laws of management processes, as well as the trends and strategies for their development, depending on the parameters of the work of the EGs between themselves and the external environment.

The analysis of the MAS works, consisting of industrial enterprises is under the control of the Criminal Code. It is shown that each enterprise that is included in the EGs can be represented as an independent agent acting within the limits of the powers assigned to it. From the position of system analysis, such MAS can be represented by several classes: the central management company, territorial management companies, industrial enterprises (agents A, B, C) (Fig. 1).

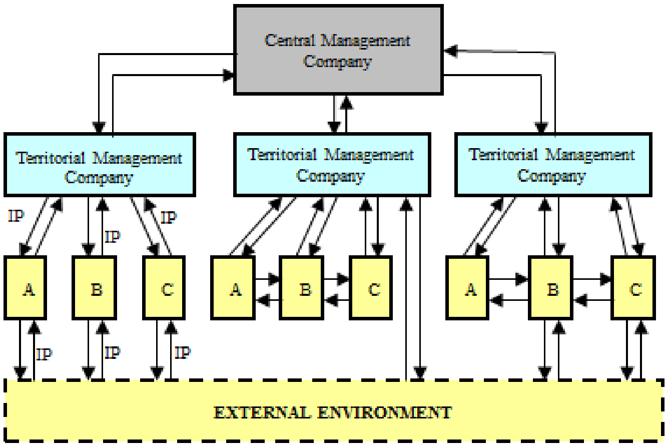

Central Management Company
Territorial Management Company
Territorial Management Company
Territorial Management Company
IP
IP
IP
A
B
C
A
B
C
A
B
C
IP
IP
IP
EXTERNAL ENVIRONMENT

Fig. 1. Scheme of Industrial Enterprises Group.

III. Proposed Methodology and Experimental Results

In solving the problems of managing business processes of the implementation of IP, the Solow single-product model is used, which can be applied both to a separate industrial enterprise and to the entire EGs. In addition, it is assumed that the MAS operates in the established external environment, whereas, there is no time lag between investments and their development, also the pace of enterprise development is characterized by the dynamics of production assets, which in turn is determined by the amount of investment resources (deductions from profits or external investments).

When optimizing the management of IP implementation at an industrial enterprise and the expense of external sources, it is assumed that at the initial moment of time the level of the fixed production assets of the enterprise in monetary terms is X0, and in order to further develop the enterprise, it is required to increase it to the required value of XT due to foreign investment of the agent investor over time [0, T]. The dynamics of changes in the current level of cost for production assets of the enterprise in this case can be written in the following form:

$$X'(t) = -\mu X(t) + D(t). \tag{1}$$

Here μ is the equipment depreciation factor, and *D(t)* is the investment rate (the investor's control parameter). An investor, when participating in an IP, tries to control the parameter *D(t)* in such a way as to obtain the minimum values of its own total investment costs *G(T)* (functional quality)

$$G(T) = \int_0^T D(t) e^{-\lambda t} dt \longrightarrow \min,$$

where λ is the discount rate of financial flows.

To determine the optimal investment process, it is necessary to integrate equation (1) taking into account the conditions for *G(T)*. If the investment is constant (*D(t)*=D_0=const), the solution to this problem can be obtained in the following form: $G_0=D_0(1-e^{-\lambda T})/\lambda$. From this expression it is clear that the total investment costs depend linearly on the rate of investment. Since the linear function has no extremes, it is not possible to optimize the investment process. To optimize the investment process, instead of linear, we will use a quadratic function, and then the investor quality functional will have the following form:

$$G_1 = \int_0^T D^2(t) e^{-\lambda t} dt \longrightarrow \min .$$

For this variant of investor's work, it is also necessary to integrate equation (1) taking into account the conditions for G1. Such a task belongs to the class of Lagrange problems of optimal control theory, and for its solution one can use the Pontryagin maximum principle.

To solve the problem in this way, the Hamilton function is compiled, and the expressions for *D*(t)* are found for which the Hamiltonian will have the maximum value. After integration (1), taking into account the requirements for G_1 and the boundary values (X_0 and X_T), the following trajectory of controlling the rate of investment was determined:

$$D^*(t) = \left(X_T - X_0 e^{-\mu T}\right)(\mu+\theta)e^{\theta t} / \left(e^{\theta T} - e^{-\mu T}\right), \tag{2}$$

Where $\theta = \mu + \lambda$, based on the obtained expressions for *D*(t)* and D_0, the total costs of the investor were calculated using the expression for *G(T).*

The calculations have shown that managing the investment process with $D^*(t)$ allows the investor to reduce his total costs while achieving the same values of X_T. Table I shows the decrease in total investor costs (as a percentage) when financing for $D^*(t)$ as compared to the rate of financing for D_0=const for a number of values of μ and λ for X_T / X_0=2 and T=2.

Thus, the management of the investment process according to the obtained expression *D*(t)* compared to D_0 = const can reduce the total investor costs when the same values of X_T are reached.

Also in this research discusses such IPs, when an industrial enterprise additionally uses its profits with selling an IP, which it invests in investments or in its own savings. The choice of one of these options is made by the company from the conditions for obtaining maximum profit. The investor participates in joint work by acquiring shares on which he receives dividends. The equation of agents working together in such MAS, by analogy with (1), will be:

$$X'(t) = -\mu X(t) + \nu\pi + \xi D(t). \tag{3}$$

Here π is the profit, and ν and ξ are the control parameters of the enterprise and the investor. It is assumed that when $v=1$, the profit of the enterprise goes to them for investment, and when $v=0$, it goes into its own savings. Similarly, when $\xi=1$, investments are made, and when $\xi=0$, there is no investment. Additionally, there is the following restriction*:* ν and ξ must vary in the range [0, 1]. At the same time, an instant investment is taken, the analysis uses a single-factor linear production function, and the change in the company's profit relative to fixed assets also has a linear relationship.

The objective function of an enterprise (G_{1m}) is both to increase its own savings and to increase the value of its fixed assets, and the objective function of an investor (G_2) is to get maximum profit:

$$G_{1m} = \int_0^T \pi(1-\nu)e^{-\lambda t}\, dt + X(T) \to \max,$$

$$G_2 = \int_0^T (\alpha\pi L - \xi D) e^{-\lambda t}\, dt \to \max \quad . \tag{4}$$

Here *α* is the share of the profit directed by the enterprise to pay dividends on its shares, and *L* is the share of shares owned by the investor from the total amount on which dividends are paid.

TABLE. I. COST REDUCTION (%) OF INVESTMENT AT D(T) COMPARED WITH D0

	λ=0.05	λ=0.10	λ=0.15	λ=0.20	λ=0.25	λ=0.30
μ=0.05	0.7	0.8	1.2	2.0	3.0	4.1
μ=0.10	1.1	1.6	2.3	3.2	4.4	5.9
μ=0.15	1.6	2.1	3.2	4.5	5.7	8.0
μ=0.20	2.3	3.3	4.5	5.7	7.3	9.6
μ=0.25	3.3	4.3	5.6	7.3	8.9	11.1

After analyzing the task, it was obtained that for integrating equation (3) with due regard for requirements (4), Pontryagin's maximum principle can be used, which allows one to obtain optimal trajectories of control parameters *v* and *ξ* in the presence of restrictions on them. As a result of solving the problem, it was obtained that over time of joint work [0, T], the control parameters *v* and *ξ* should change as follows:

$$v=\begin{cases}1 & \text{if} \quad t>t_0\,; \\ 0 & \text{if} \quad t<t_0\,;\end{cases} \qquad \xi=\begin{cases}1 & \text{if} \quad t<t_1\,; \\ 0 & \text{if} \quad t_1<t<t_0\,.\end{cases} \qquad \xi=\begin{cases}1 & \text{if} \quad t_0<t<t_2; \\ 0 & \text{if} \quad t>t_2\,.\end{cases}$$

As can be seen, changes in the control parameters are of a relay nature and equal to one of their two possible limit values. The moments of time at which changes in these parameters should occur are equal to:

$$t_0=T\frac{b-\mu}{b-\theta}; \qquad t_1=t_0+\frac{1}{\theta}\ln\frac{bL-\theta}{bL};$$

$$t_2=T+\frac{1}{\theta}\ln\frac{\alpha bL-\theta}{\alpha bL}$$

Where $\theta=\mu+\lambda$, $b=\partial\pi/\partial X$, wherein $t_1<t_0<t_2<T$.

In order to fulfill conditions (4), the optimal management of the use of its profits by an enterprise will be as follows. On the time interval [0, t_0], where *v (t)=0*, the profit of the enterprise is sent to its own savings, and on the time interval [t_0, T], where *v (t) = 1*, it is spent on the IP (Fig. 2).

For optimal investment management, the investor (Fig. 2) acquires shares at the time interval [0,t1], receiving additional dividends, at the time t1 stops buying shares (ξ=0) and only receives dividends at [t1, t0] on previously acquired shares. If the company in the time interval [t0,T] does not pay dividends (α=0), then from the moment of time t0 the investor stops his participation in the joint IP. If during the time period [t0,T] the company continues to pay dividends, the investor starts buying shares again before the time t2. At this point in time, the acquisition of shares is terminated, and in the period [t2,T], the investor will only receive dividends on previously acquired shares. Such management of joint work of the enterprise and the investor allows obtaining the maximum values of the functional quality (G1m and G2) to all participants of the implementation of the IP.

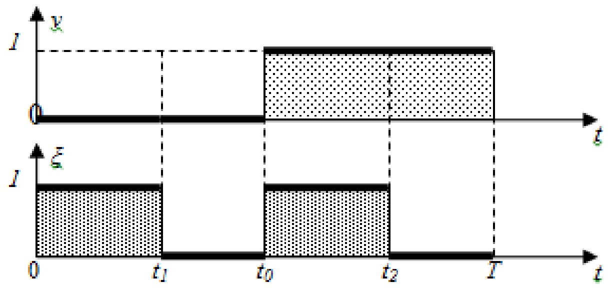


Fig. 2. Optimal Change of Control Parameters Enterprises (ν) and Investor (ξ).

In the next stage of the research, optimization of the management of the joint venture for the manufacture and sale of industrial products is carried out. It discusses the work of the MAS, consisting of several agents that are enterprises of EGs, for example, auxiliary and main production (Fig. 3). The auxiliary production enterprise (agent A) acquires the necessary materials and components, conducts their control, preprocessing and then transfers them to the main production for manufacture, testing, assembly and further implementation (agents B and C). The joint work of agents over time in such MAS can be represented by the following system of differential equations:

$$\begin{cases}x'=U(X1-x)-V(Z1-z)-R(Y1-y); \\ z'=V(Z1-z)-Wz\,; \\ y'=R(Y1-y)-Sy.\end{cases} \tag{5}$$

Here x, y, and z accordingly, the quantity of products in the warehouse of agents A, B and C, *U(t)* is the speed of production by agent A, *V(t),W(t),R(t),S(t)* the speed of the rate acquired and subsequently sold by agents B and C units of production, *X1, Z1* and *Y1* are the maximum production capacities of the warehouses of agents A, B and C.

During the joint work [0,T], each of the agents seeks to get the maximum profit for themselves (*J1* for agent A, *J2* and *J3* for agents B and C):

$$J_1=\int_0^T\left[c_1V(Z1-z)+c_3R(Y1-y)-c_0U(X1-x)-d_1x\right]dt\to\max \tag{6}$$

$$J_2=\int_0^T\left[c_2Wz-c_1V(Z1-z)-d_2z\right]dt\to\max \tag{7}$$

$$J_3=\int_0^T\left[c_4Sy-c_3R(Y1-y)-d_3y\right]dt\to\max \tag{8}$$

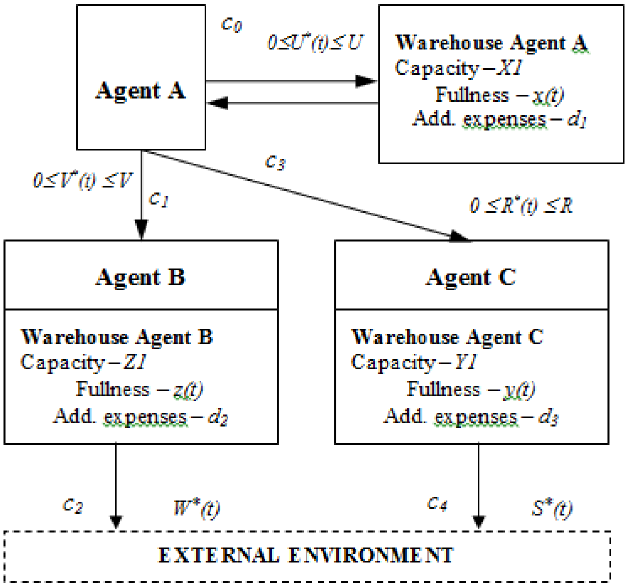


Fig. 3. Collaboration of EGs Agents.

where c_0, c_1, c_2, c_3, c_4 are the unit costs manufactured by agent A, the products purchased and sold by agents B and C, and d_1, d_2 and d_3 are the additional costs of transporting and storing them in the agents' warehouses.

Each agent can get the maximum value of its own functional quality by changing its control parameters *U (t)* (agent A), *V (t)*, *W (t)* (agent B) and *R (t)*, *S (t)* (agent C).

Consider first the work of agent B in such MAS. To do this, it is necessary to integrate (5) taking into account the fulfillment of the requirements (8). In solving this problem, we will use the Pontryagin maximum principle, for this we construct the Hamilton function:

$$H_3 = \psi_1 U\left(X1 - x\right) + h_1 R\left(Y1 - y\right) + h_2 Sy + h_3 V\left(Z1 - z\right) - d_3 y,$$

Where a ψ1, ψ2 and ψ3 are auxiliary variables defined by the expressions:

$$\psi_1^{'} = -\frac{\partial H_3}{\partial x} = U\psi_1; \psi_2^{'} = -\frac{\partial H_3}{\partial z} = Vh_3 + W\psi_2;\ \psi_3^{'} = -\frac{\partial H_3}{\partial y} = Rh_1 - Sh_2 + d_3$$

In accordance with the Pontryagin maximum principle, the optimal control of the work of agent C will be if the Hamilton function has the maximum value. This will be the case if the control parameters of agent C are changed as follows:

$$R^*(t) = \begin{cases} 0\ , & h_1 < 0; \\ R\ , & h_1 > 0; \end{cases} \qquad S^*(t) = \begin{cases} 0\ , & h_2 < 0; \\ S\ , & h_2 > 0. \end{cases}$$

It can be seen that the changes in the control parameters of the agent C are of a relay nature and are equal in magnitude to one of their two possible limiting values. Moreover, the whole process of changing these parameters consists of two intervals, in which these parameters have a constant value, and the duration of these intervals is determined by the auxiliary variables ψ1, ψ2 and ψ3, which can be found from the solution of the differential equations for them. The boundary values for these variables were defined in this part of the research, and they are respectively equal: $\psi_1(T) = \psi_2(T) = \psi_3(T) = 0$.

The expressions for auxiliary variables obtained in this chapter showed that the function $h_2>0$ on the entire interval [0, T], the function $h_1>0$ on the interval [0, t_2], and on the interval [t_2, T] the function $h_1<0$. The time t_2, at which the control parameter *R*(t)* becomes zero, and agent C stops purchasing products from agent A, is equal to:

$$t_2 = T + \frac{1}{S}\ln\frac{\left(c_4 - c_3\right)S - d_3}{c_4 S - d_3}; \quad t_1 = T + \frac{1}{W}\ln\frac{\left(c_2 - c_1\right)W - d_2}{c_2 W - d_2} \qquad (9)$$

The work of Agent B is considered in a similar way. It has been obtained that the optimal control of the process of his work will be the same as for Agent C, i.e. at the initial time interval [0, t_1], he acquires, and at the time interval [t_1, T] stops purchasing products from agent A. The value of time t_1, at which its control parameter *V*(t)* changes, is determined by 9).

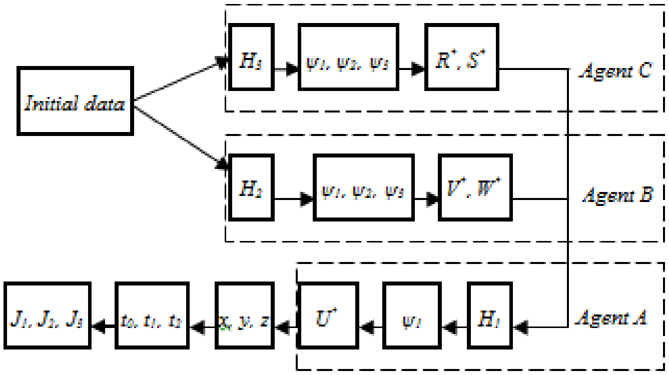


Fig. 4. Algorithm for Determining the Optimal Parameters for Managing Agents.

The determination of the optimal modes of operation for agent A was also carried out using the Pontryagin maximum principle, using the system of equations (5) and (6). As a result of the analysis, it was determined that in order to obtain maximum profit, Agent A should produce products in the initial period of time [0,t_0], and stop production in the period of time [t_0,T]. Fig. 4 shows the scheme for determining the control parameters of agents A, B and C when they work together in the MAS.

When determining the point in time t_0, it is necessary to take into account that Agent A usually tries to fully realize all products manufactured for them, i.e. receive $x(T) = 0$. In this stage of the research, an expression is obtained for determining the quantity of products in the warehouse of agent A at the end of the collaboration time:

$$x\left(T\right) = K_0 + K_1 x_0 + K_2\left(t_1 - t_0\right) + K_3\left(e^{-\left(V+W\right)t_1} - e^{-\left(V+W\right)t_0}\right), \qquad (10)$$

Where K_0,K_1,K_2 and K_3 are constant values depending on the initial indicators of the joint work of agents. Equating expression (10) to zero, we can find the unknown value t_0. Due to the fact that the resulting equation is non-linear, its solution is carried out by numerical methods. For this purpose, a program is compiled, the algorithm of which is shown in Fig. 5. Its input data are the left (*a*) and right (*b*) boundaries of the interval in which the desired one (t_0) is located, as well as the accuracy of calculations (ξ).

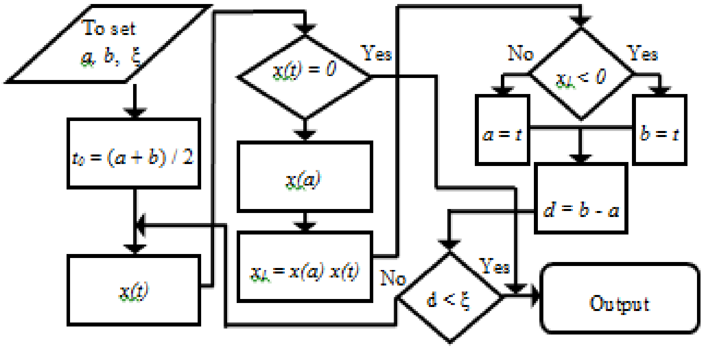


Fig. 5. Algorithm for Determining the Optimal Moment of Time t_0 Stopping Production of the Agent A.

The optimal process for managing the joint work of agents in such MAS will be as follows. On the time interval [0, t0], agent A makes at time t0, stops production and does not make production on the interval [t0, T]. Agent B in the time interval [0, t1], and agent C in the interval [0, t2] are purchasing products. At time t2 agent C and at time t1 agent B is stopping buying products from agent A. At the same time, agents B and C are engaged in selling products throughout the entire time interval [0, T].

The resulting collaboration scheme has a rationale - it is not profitable for any of the agents to produce (acquire) excess products, which subsequently cannot be fully realized. Such a scheme of parallel collaboration of agents does not always suit Agent A, since there are periods of time during which its production will stand idle. To eliminate this drawback, a sequential scheme of agents is proposed. For the case of the MAS of the three agents A, B and C, it is shown in Fig. 6.

Initially, agent A works with agent B, at time t0 stops the production of products for agent B, and instead of completely stopping his production, he begins to manufacture products for agent C during a period of time [t0,T]. From the moment of time T, work with agent B begins again and then the whole cycle of joint work of agents A, B and C is repeated again. With such a consistent scheme, all agents (A, B, and C) are excluded from work stoppages.

Fig. 7 shows the change in total profit of agent A (JA = JB + JC) over time t with the above sequential work pattern. Curves JB (agent B) and JC (agent C) show the change in the profit of agent A in the event that he works separately with each of these agents. The sequential scheme of work of agent A is that from the initial moment to t0 = 8.0 he works with agent C, and his profit is negative, i.e. the cost of manufacturing is still more than the cost of the products sold to them. From the moment t0 = 8.0, he ceases to manufacture products for agent C, and only sells to him previously manufactured products. Simultaneously, the production of products for agent B begins, which continues until time T = 20.0. From the time point T = 20.0, work with agent C begins again, and the whole cycle repeats again.

The moments of time t0, t1 and t2 depend on the initial parameters values for the joint work of agents. At the stage of discussing the conditions of joint work of agents, decision makers on the basis of the recommendations DSS agree on the values of these parameters that ensure the continuous operation of each agent.

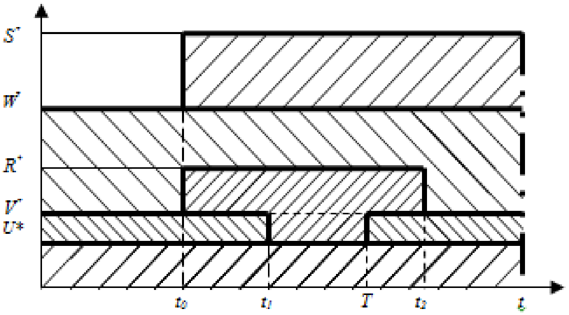


Fig. 6. Change of Control Parameters of Agents A, B and C During their Sequential Work.

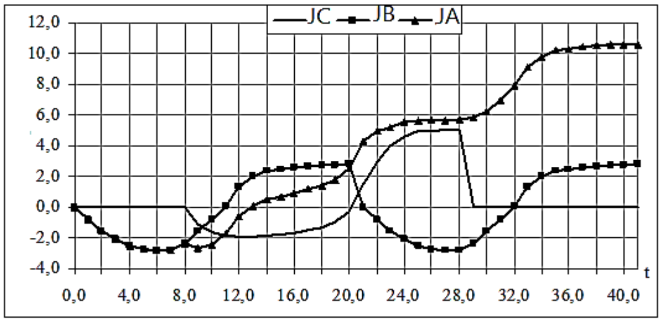


Fig. 7. Sequential Work of Agents A, B and C.

IV. Results and Practical Implementations

In practice, the joint work of manufacturers and consumers of products may differ from that discussed above. In the annex to the research, additional options for the joint work of agents are given, which differ from each other in the initial conditions that characterize the manufacture, acquisition and sale of products. So, for example, the joint work of the MAS of two agents - the manufacturer of the product (agent A) and his dealer (agent B) can be conditionally represented by the following system of equations:

$$\begin{cases} x' = U - V\,x; \\ z' = V\,x - W\,z. \end{cases}$$

The functional quality (J1 of agent A, J2 of agent B) for this variant of their joint work will be as follows:

$$J_1 = \int_0^T \left(c_1 V x - U c_0 - d_1 x\right) dt - c_1 x(T) \rightarrow \max$$

(11)

$$J_2 = \int_0^T \left(c_2 W z - c_1 V x - d_2 z\right) dt - c_2 z(T) \rightarrow \max \qquad (12)$$

The task for both agents is to search for the optimal trajectories of changes in the control parameters *U**, *V** and *W**, at which *J1* and *J2* will have maximum values. In this stage of this research, we analyze the operation of such a system and find the optimal modes for changing the control parameters of agents A and B, which ensure maximum efficiency for each of them.

Due to the unstable modern market situation, there are cases when one of the agents at some point in time decides to stop working together. In this case, he needs to fully realize all his products. This part of the research examines such possible cases and determines the minimum possible sales times, which are equal to (T1 and T2) for agents A and B:

$$T_1 = -\frac{1}{V} \ln \frac{V_1}{V_1 + V x_k}; \quad T_2 = -\frac{1}{W} \ln \frac{w}{w + W\,z_k}$$

Here x_k and z_k - the number of products in the warehouses of agents at the time of the decision to cease collaboration, *V1* and *w*, respectively, the minimum rate of acquisition and sale of products by agent B.

We provide examples of determining the optimal processes for managing the joint work of agents in MAS, and also discuss the practical implementation of the developed DSS in the management of EGs business processes.

One of the mandatory conditions for the joint work of agents is to obtain the maximum values of their functional quality by each of them. From the expressions for these functional quality (6–8, 11–12) it can be seen that their values depend on many initial indicators. Fig. 8 shows the typical nature of the change over time *t* of the joint operation of the J_2 functional of agent B depending on the cost c_1 of the product it purchases from agent A with the following initial data of joint work of agents: T=20.0; c_0=1.0; c_2=2.5. It can be seen that the change in J_2 from c_1 and *t* has a significant non-linear dependence, with the maximum value of $J_2(T)$ being c_1≈1.3-1.5.

The same character affects c_1 and *t* on the change in the functional quality of agent A, while its maximum value will be at a different value of c_1. The choice of a specific value c_1 should be made by the agents on the basis of the recommendations of DSS when coordinating the source data and modes of their joint work.

Also depending on the practical implementation, the graphs and tables show the effect of the initial indicators on the change in the control parameters of the agents, on the basis of which solutions for managing MAS agents are developed in DSS. Fig. 8 shows the dependence of the functional J_2 on t and c_1.

The maximum values of the functional quality of each agent are largely determined by the initial values of the collaboration indicators. Each of the agents usually seeks to select them such that would provide him the greatest profit. To determine the values of such indicators, it is necessary to optimize the profit margin relative to possible changes in the values of the initial parameters. To this end, it is proposed to use numerical methods for finding the maximum function of several variables.

This step analyzes the existing search methods and shows that in the DSS one can use the deformable polyhedron method. A modification of this method has been carried out, which allows for a given accuracy to significantly reduce the time for finding the maximum value. A scheme is given and the operation of this algorithm for optimizing the functional quality of agents' work is considered.

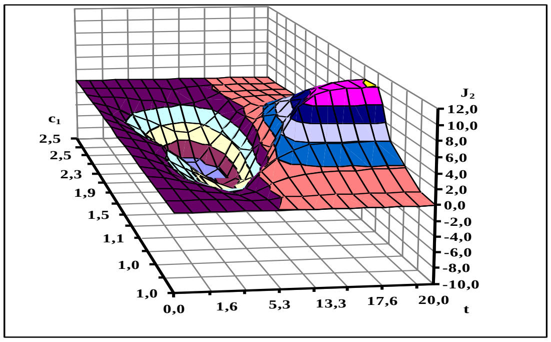


Fig. 8. The Dependence of the Functional J2 on t and c1.

V. Conclusions

It was conducted a system analysis of DSS in the management of business processes EGs. This made it possible to determine directions and methods for optimizing their management systems for use in implementing joint business processes. Also by using the Pontryagin maximum principle, the problem of managing the distribution of investments over the step of the implementation of a joint IP for the MAS of the enterprise and agent-investor has been solved. This made it possible to determine the optimal trajectories of changes in the management process of IP, which, unlike the existing ones, takes into account both the initial and final values of the required parameters of the investment process. It is shown that the application of the developed methodology allows the investor to reduce their costs while achieving the same results of the investment process.

Expressions are obtained for the functional quality (objective functions) of the MAS, taking into account both static and dynamic indicators of their joint work. This made it possible to determine the points in time at which management decisions should be made in the DSS to change the parameters characterizing the work of the MAS, depending on the tasks assigned to each agent. A consideration was given to managing dynamic MAS of agent-enterprise and agent-investor collaboration, taking into account the possible options for the agent-enterprise to use the profit and investment of the agent-investor. It is shown that in order to obtain maximum values of the functional quality for each of the agents from teamwork, they need to change the modes of using their profits (for the enterprise) and foreign investments (for the investor) over time. Expressions are obtained to determine the points in time at which the control parameters of the agents should be changed to ensure the greatest efficiency of the MAS operation. It is also considered the analysis of dynamic systems, consisting of manufacturers and consumers of industrial products for an arbitrary period of time.

A scheme and management options for parallel and sequential work of agents in the DSS, based on a dynamic analysis of the manufacturing processes and sales of industrial products have been developed. Modes and times of changes in control parameters, the duration and nature of changes in the performance of agents in such a system are determined from the condition for each agent to receive the maximum values of the functional quality from their joint work. The initial data of the sequential work of the enterprise have been determined, which enable them to organize continuous work in a closed loop in the manufacture and sale of products. The structure and development of DSS in the management of EGs business processes and structural transformations of EGs has been determined. Approbation and implementation of the DSS was carried out in the state enterprise, confirming the effectiveness and efficiency of its use on real objects.

References

[1] Leach, L.P., Critical chain project management. 2014: Artech House.

[2] AlRababah, A.A., A. AlShahrani, and B. Al-Kasasbeh, Efficiency Model of Information Systems as an Implementation of Key Performance Indicators. International Journal of Computer Science and Network Security (IJCSNS), 2016. 16(12): p. 139.

[3] Othman, S.B., et al., An agent-based decision support system for resources' scheduling in emergency supply chains. Control Engineering Practice, 2017. 59: p. 27-43.

[4] Chan, S.H., et al., Decision support system (DSS) use and decision performance: DSS motivation and its antecedents. Information & Management, 2017. 54(7): p. 934-947.

[5] Goodwin, P. and G. Wright, Decision Analysis for Management Judgment 5th ed. 2014: John Wiley and sons.

[6] Noe, R.A., et al., Human resource management: Gaining a competitive advantage. 2017: McGraw-Hill Education New York, NY.

[7] Rose, D.C., et al., Involving stakeholders in agricultural decision support systems: Improving user-centred design. International Journal of Agricultural Management, 2018. 6(3-4): p. 80-89.

[8] Ruiz, P.A.P., B. Kamsu-Foguem, and D. Noyes, Knowledge reuse integrating the collaboration from experts in industrial maintenance management. Knowledge-Based Systems, 2013. 50: p. 171-186.

[9] de Souza Melaré, A.V., et al., Technologies and decision support systems to aid solid-waste management: a systematic review. Waste management, 2017. 59: p. 567-584.

[10] Garousi, V., et al. Industry-academia collaborations in software engineering: An empirical analysis of challenges, patterns and anti-patterns in research projects. in 21st International Conference on Evaluation and Assessment in Software Engineering (EASE 2017). 2017. ACM.

[11] Van Huben, G.A. and J.L. Mueller, Data management system for file and database management. 2000, Google Patents.

[12] Al-Rababah, A. and N. Hani. Component linked based system. in Modern Problems of Radio Engineering, Telecommunications and Computer Science, 2004. Proceedings of the International Conference. 2004. IEEE.

[13] Al-rababah, A.A. and M.A. Al-rababah, Module Management Tool in Software Development Organizations 1. 2007.

[14] 14. Sahir, S.H., R. Rosmawati, and R. Rahim, Fuzzy model tahani as a decision support system for selection computer tablet. Int. J. Eng. Technol, 2018. 7(2.9): p. 61-65.

[15] Al-Rababah, A.A., T. AlTamimi, and N. Shalash, A New Model for Software Engineering Systems Quality Improvement. Research Journal of Applied Sciences, Engineering and Technology, 2014. 7(13): p. 2724-2728.

[16] Caniëls, M.C. and R.J. Bakens, The effects of Project Management Information Systems on decision making in a multi project environment. International Journal of Project Management, 2012. 30(2): p. 162-175.

[17] Al-Rababah, A.A. and M.A. Al-Rababah, Functional Activity Based Comparison Study for Neural Network Application. IJCSNS, 2007. 7(1): p. 153.

[18] Al Ofeishat, H.A. and A.A. Al-Rababah, Real-time programming platforms in the mainstream environments. IJCSNS, 2009. 9(1): p. 197.

[19] Gupta, P. Accelerating datacenter workloads. in 26th International Conference on Field Programmable Logic and Applications (FPL). 2016.

[20] AlRababah, A., Digital Image Encryption Implementations Based on AES Algorithm. VAWKUM Transactions on Computer Sciences, 2017. 13(1): p. 1-9.

An Extended Consistent Fuzzy Preference Relation to Evaluating Website Usability

Tenia Wahyuningrum[1], Azhari Azhari*[2], Suprapto[3]
Department of Informatics, Institut Teknologi Telkom Purwokerto, Indonesia[1]
Department of Computer Science and Electronics, Universitas Gadjah Mada Yogyakarta, Indonesia[1, 2, 3]

***Abstract*—In the current era, website developers recognize usability evaluation as a significant factor in the quality and success of e-commerce websites. Fuzzy Analytical Hierarchy Process (FAHP) is one method to measure the usability of the website. Several researchers have applied Logarithmic Fuzzy Preference Programming (LFPP) approach to deriving crisp weight from fuzzy pairwise comparison matrix of FAHP approach. However, there is a lack of LFPP method in determining the consistency index of the decision-maker judgment. In some cases, LFPP method will produce a consistency value of 0 from consistent fuzzy comparison matrices. This value indicates there is a contradiction with what the previous researchers have said, that a constant matrix value should be more than 0. This research proposes the extended Consistent Fuzzy Preference Relation (ECFPR) to assist the regular judgment for specifying the weights in measuring e-commerce website usability. The CFPR method used to form a new pairwise comparison matrix. ECFPR was calculating the lower and upper values at the fuzzy triangular number from the only *n*-1 comparison, where *n* is the number of criteria. The numerical experiment showed that the consistency index obtained by extended CFPR method was more significantly better than LFPP method. It was revealed that the optimal value always more than 0. The consistency index of ECFPR method has a higher mean value than LFPP, so that the use of the ECFPR method can improve the amount of consistency comparison matrices. The ECFPR method was also successfully implemented with the experimental case on evaluating e-commerce website usability.**

***Keywords*—Usability; e-commerce; website quality; logarithmic fuzzy preference programming; consistent fuzzy preference relations**

I. Introduction

Current usability measurement methods do not yet have the right uniformity and agreement on standards in software [1], [2]. One measure for websites usability is the sum of products between the weights of the criteria and the value of each of the criteria [3]–[5]. Researchers often regard the weighting of standards as a multi-criteria decision-making problem, given its complex structure. They usually break down the complex issues into its elements in a hierarchy. Several researchers have conducted usability measurements using a combination of fuzzy numbers and Analytical Hierarchy Process (FAHP) [6]–[9]. Fuzzy numbers consider the uncertainty and doubt factors in the experts in determining the level of importance between criteria. Fuzzy logic and fuzzy decision making are part of the branch of fuzzy theory. In fuzzy decision making consider optimizing problems with certain limitations while fuzzy logic is the basis of knowledge in fuzzy systems and controls [10].

Logarithmic Fuzzy Preference Programming (LFPP) is one method to evaluate usability based on FAHP method [11]. LFPP is an approach using non-linear programming to derive the weight of criteria. LFPP uses the logarithm of natural numbers to repairing Fuzzy Preference Programming (FPP) approach that caused the negative of fuzzy membership degree [12], [13]. However, the LFPP method has its drawbacks; in some cases, a pairwise comparison matrix that consistently produces a value of 0 [14], [15]. This case is not in line with the definition that states that the more consistent fuzzy pairing comparison matrix, the optimal value (λ^*) is closer to 1 [16]. There is a presumption that the probability of a value of 0 in λ^* is due to a matrix that is not consistent. Therefore, before calculating the weights using the LFPP method, it is necessary to ensure the consistency of the model. The technique for guaranteeing a matrix to be consistent is called Consistent Fuzzy Preference Relation (CFPR). CFPR is an approach to reduce the number of comparisons that are often done by users in determining preferences between criteria or alternatives. The weakness of CFPR only considers the modal value at triangular fuzzy number so that it produces a comparison matrix with crisp numbers. The extended CFPR is reviewing the upper and lower bound and creating a comparison matrix in pairs with fuzzy triangular numbers. By applying the extended CFPR method, it is expected to increase fuzzy preference relationships consistency provided by experts to make it better. The extended CFPR based on user judgment is expected to be valid and consistent so that it can give good weight also from each usability criteria.

The organization of this paper is as follows. Section 2 briefly reviews the LFPP and illustrates its consistency equal to 0. Section 3 proposes the extended CFPR to create the fuzzy pairwise comparison matrix. Section 4 explained the numerical case of usability evaluation method using extended CFPR and LFPP. The paper concludes in Section 5.

II. Literature Review

A. Logarithmic Fuzzy Preference Programming

Wang and Chin (2011) reforming the Fuzzy Preference Programming (FPP) weight derivation. They modified the FPP method by adding natural logarithms function to improve the negative membership degree and arising in multiple optimal solutions. Negative value to makes the expected solution less

*Corresponding author: arisn@ugm.ac.id. This research was supported by the Ph.D. grant of the Indonesia endowment fund for education (LPDP) ministry of Finance Republik Indonesia for Tenia Wahyuningrum (No. PRJ-4811/LPDP.3/2016).

valid [16]–[19]. The fuzzy pairwise comparison matrix from expert judgment can be expressed as

$$A = (a_{ij})_{n\times n}$$

$$A = \begin{pmatrix} (1,1,1) & \dots & (l_{1n}, m_{1n}, u_{1n}) \\ (l_{21}, m_{21}, u_{21}) & (1,1,1) & (l_{2n}, m_{2n}, u_{2n}) \\ \vdots & \vdots & \vdots \\ (l_{n1}, m_{n1}, u_{n1}) & \dots & (1,1,1) \end{pmatrix} \qquad (1)$$

where n is the number of criteria, $l_{ij} = 1/u_{ji}$, $m_{ij} = 1/m_{ji}$, $u_{ij} = 1/l_{ji}$ and $0 < l_{ij} < m_{ij} < u_{ij}$ for all $i, j = 1, 2, \dots, n, j \neq i$. To find a crisp priority vector $W=(w_1, w_2, \dots, w_n)^T > 0$ with $\sum_{i=1}^{n} w_i = 1$ for the fuzzy pairwise comparison matrix [16]. The approximate equation uses natural logarithmic numbers for the improvement of fuzzy pairwise matrix (1).

The LFPP method formulated as Minimize

$$J = (1-\lambda)^2 + M \cdot \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\delta_{ij}^2 + \eta_{ij}^2\right) \qquad (2)$$

Subject to

$$\begin{cases} x_i - x_j - \lambda \ln\left(\dfrac{m_{ij}}{l_{ij}}\right) + \delta_{ij} \geq \ln l_{ij}, i = 1,2,\dots,n-1; j = i+1,\dots,n, \\ -x_i + x_j - \lambda \ln\left(\dfrac{u_{ij}}{m_{ij}}\right) + \eta_{ij} \geq -\ln u_{ij}, i = 1,2,\dots,n-1; j = i+1,\dots,n, \\ \lambda, x_i \geq 0, i = 1,2,\dots,n, \\ \delta_{ij}, \eta_{ij} \geq 0, i = 1,2,\dots,n-1; j = i+1,\dots,n, \end{cases}$$

where $x_i = \ln w_i$ for $i = 1, 2, \dots, n$ and M is a specified large number such as $M = 10^3$. Equation (3) can be used to calculate the weight of each criterion :

$$w_i^* = \frac{\exp(x_i^*)}{\sum_{j=1}^{n} \exp(x_j^*)}, i = 1,2,\dots,n, \qquad (3)$$

where exp() is the exponential function $\exp(x_i^*) = e^{x_i^*}$ for $i = 1, 2, \dots, n$.

B. Consistent Fuzzy Preference Relation

Preference relations are usually constructed as a matrix that represents the degree of interest for the first criteria over the second criteria. The relationship of this assessment can be multiplicative preference relations of fuzzy preference relations. Multiplicative preference relations can be formulated as

$$R \subseteq A \times A, R = (r_{ij}), \forall i, j \in \{1,2,\dots,n\}, \qquad (4)$$

where A is the set of criteria or alternatives, r_{ij} is the preference ratio of criteria or alternative a_i to a_j, $a_{ij} \cdot a_{ji} = 1$, $\forall\ i, j \in \{1, 2, \dots, n\}$. Furthermore, the relationship will be represented by a pairwise matrix comparison $P=p_{ij}$, where the size is $n\times n$, $p_{ij} = \mu_p(a_i, a_j)$, $\forall i, j \in \{1,2,\dots,n\}$ and the value of the membership function of fuzzy logic. Elements in the pairwise comparison matrix are calculated using several propositions [20].

Proposition 1. Consider set criteria or alternatives, $X=\{x_1, x_2, \dots, x_n\}$ associated with a reciprocal multiplicative preference relation $A=(a_{ij})$ for $a_{ij} \in [1/9, 9]$. Then, the corresponding reciprocal fuzzy preference relation, $P=p_{ij}$ with $p_{ij} \in [0,1]$ associated with A is given as

$$p_{ij} = g(a_{ij}) = \frac{1}{2}(1 + \log_9 a_{ij}) \qquad (5)$$

Proposition 2. For each P=g(A), where $P=(p_{ij})$, the booth of equations (6) and (7) are equivalent.

$$p_{ij} + p_{jk} + p_{ki} = \frac{3}{2}, \forall i, j, k \qquad (6)$$

$$p_{ij} + p_{jk} + p_{ki} = \frac{3}{2}, \forall i < j < k \qquad (7)$$

Proposition 3. For each $P=(p_{ij})$, the booth of equations (8) and (9) are equivalent.

$$p_{ij} + p_{jk} + p_{ki} = \frac{3}{2}, \forall i < j < k \qquad (9)$$

$$p_{i(i+1)} + p_{(i+1)(i+2)} + \dots + p_{(i+k-1)(i+k)} + p_{(i+k)i} = \frac{k+1}{2}, \forall i < j. \qquad (10)$$

Proposition 3 is used to construct a consistent fuzzy preference relation from the set of n-1 values $\{p_{12}, p_{23}, \dots, p_{n-1n}\}$. A decision matrix with entries that are not in the interval [0,1], but in an interval $[-k, 1+k]$, $k>0$, can be obtained by transforming the result values using a transformation function that preserves reciprocity and additive consistency. It is given by the function $f:[-k, 1+k]$ to [0,1], $f(x)=(x+k)/(1+2k)$.

III. Proposed Method

A. Extended Consistent Fuzzy Preference Relation

Wang and Chin (2011) argued that the consistency index of fuzzy pairwise comparison can be seen at the value of λ^* and δ^*[16]. The inconsistency in a fuzzy pairwise comparison matrix could be expressed as a Proposition 4.

Proposition 4. If $(\delta^* > 0) \wedge (\lambda^* = 0)$ then the matrix is very inconsistent.

Proposition 4 can be broken down into two new schemes p and q. $p = (\delta^* > 0) \wedge (\lambda^* = 0)$ and q = *the matrix is strong inconsistent.* The proposition can be explained as $p \rightarrow q$, then the equivalent of Proposition 4 is $\neg q \rightarrow \neg p$. It is also can be expressed as a Proposition 5.

Proposition 5.

$$\neg q \rightarrow \neg((\delta^{*} > 0) \wedge (\lambda^{*} = 0))$$

$$\neg q \rightarrow \neg(\delta^{*} > 0) \vee \neg(\lambda^{*} = 0) \qquad (10)$$

$$\neg q \rightarrow (\delta^{*} \leq 0) \vee (\lambda^{*} \neq 0) = true$$

If $r = (\delta^{*} \leq 0)$, $s = (\lambda^{*} \neq 0)$ then $\neg q \rightarrow r \vee s$. Because of the $\delta^{*} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\delta_{ij}^{*2} + \eta_{ij}^{*2}\right)$ is always more than equal to 0, then the proposition r is always false. For the statement to be true, the λ^{*} must be not equal to 0. It can be said, that the consistency of the fuzzy pairwise matrix must be not equal 0 ($\lambda^{*} \neq 0$). The condition of this proposition to comply with the truth table of OR operations.

Experiment 1. Economic factors sub-criteria on shipping registry problem based on Wang and Chin (2011) using LFPP and AHP method used to analyze the consistency index of the fuzzy pairwise comparison matrix [16], [21], [22].

$$X = \begin{pmatrix} (1,1,1) & \left(\frac{2}{3},1,\frac{3}{2}\right) & (1,1,1) & \left(\frac{2}{5},\frac{1}{2},\frac{2}{3}\right) \\ \left(\frac{2}{3},1,\frac{3}{2}\right) & (1,1,1) & \left(\frac{2}{5},\frac{1}{2},\frac{2}{3}\right) & \left(\frac{2}{3},1,\frac{3}{2}\right) \\ (1,1,1) & \left(\frac{3}{2},2,\frac{5}{2}\right) & (1,1,1) & \left(\frac{2}{5},\frac{1}{2},\frac{2}{3}\right) \\ \left(\frac{3}{2},2,\frac{5}{2}\right) & \left(\frac{2}{3},1,\frac{3}{2}\right) & \left(\frac{3}{2},2,\frac{5}{2}\right) & (1,1,1) \end{pmatrix}$$

In the AHP method, the consistency ratio value represents an appropriate pairwise comparison matrix. Suppose $X(l, m, u)$ is a triangular fuzzy number (TFN), then the defuzzified value is computed as (11) [23].

$$x_{ij} = (l+4m+u)/6 \qquad (11)$$

where l is the lower value, m is the modal value, and u is the upper value of the support of X respectively. Equation (11) can be used to estimate the crisp pairwise comparison matrix X.

$$X = \begin{pmatrix} 1 & 1.02 & 1 & 0.5 \\ 0.97 & 1 & 0.5 & 1.02 \\ 1 & 2 & 1 & 0.5 \\ 2 & 0.97 & 2 & 1 \end{pmatrix}$$

Using the AHP method, the result of the consistency index (CI) is 0.06, and the consistency ratio (CR) is 0.07, which is consistency. The consistency index also calculated using LFPP method for investigating the differences of conclusion between two approaches. For X is a fuzzy pairwise matrix, the calculation can be written as Minimize

$$J = (1-\lambda)^{2} + M \sum_{i=1}^{3} \sum_{j=i+1}^{4} \left(\delta_{ij}^{2} + \eta_{ij}^{2}\right)$$

Subject to

$$\begin{cases} x_1 - x_2 - \lambda \ln(3/2) + \delta_{12} \geq \ln(2/3), \\ -x_1 + x_2 - \lambda \ln(3/2) + \eta_{12} \geq -\ln(3/2), \\ x_1 - x_3 - \lambda \ln(1) + \delta_{13} \geq \ln(1), \\ -x_1 + x_3 - \lambda \ln(1) + \eta_{13} \geq -\ln(1), \\ x_1 - x_4 - \lambda \ln(5/4) + \delta_{14} \geq \ln(2/5), \\ -x_1 + x_4 - \lambda \ln(4/3) + \eta_{14} \geq -\ln(2/3), \\ x_2 - x_3 - \lambda \ln(5/4) + \delta_{23} \geq \ln(2/5), \\ -x_2 + x_3 - \lambda \ln(4/3) + \eta_{23} \geq -\ln(2/3), \\ x_2 - x_4 - \lambda \ln(3/2) + \delta_{24} \geq \ln(2/3), \\ -x_2 + x_4 - \lambda \ln(5/4) + \eta_{24} \geq -\ln(2/5), \\ x_3 - x_4 - \lambda \ln(5/4) + \delta_{34} \geq \ln(2/5), \\ -x_3 + x_4 - \lambda \ln(4/3) + \eta_{34} \geq -\ln(2/3), \\ \lambda, x_1, x_2, x_3, x_4, \delta_{12}, \delta_{13}, \delta_{14}, \delta_{23}, \delta_{24}, \delta_{34} \geq 0, \\ \eta_{12}, \eta_{13}, \eta_{14}, \eta_{23}, \eta_{24}, \eta_{34} \geq 0. \end{cases}$$

Define a sufficiently large number for M=10^{3}, then the result of optimal value (λ^{*}) is 0, that means strong inconsistent [16]. This conclusion is contrary to the definition that increases the fuzzy pairwise comparison matrix, the higher the value, so research needs to be done to calculate whether the resulting value reflects fuzzy inconsistency. There is the difference conclusion between AHP and LFPP method about the consistency. This difference makes us decide to check and ensure that the fuzzy pairwise matrix comparison is truly consistent.

The proposed model was developed by modifying the pairwise matrix comparison steps, using the CFPR method before weighting criteria. By applying the CFPR method, the consistency of the fuzzy preference relationships provided by decision-makers will be improved. The CFPR method answers the weakness of the Analytical Hierarchy Process (AHP) method, which causes the situation to be inconsistent because there are too many questions and comparisons, creating waste and time inefficiency. The CFPR method only requires n-1 comparison that must be answered by the evaluator, and the rest is derived using a predetermined proposition formula [24]. Combining the CFPR and LFPP methods are expected to increase the optimal value, not equal 0 (called extended CFPR).

The extended CFPR Model (ECFPR) is a modification of the step comparison of the fuzzy pair matrix using the CFPR method before weighting the criteria. The ECFPR model gives a new proposition, which is formed as Proposition 6.

Proposition 6. Consider set criteria or alternatives, $X=\{x_1, x_2, \ldots, x_n\}$ associated with a reciprocal multiplicative preference relation $A = (a_{ij})$, $a_{ij} = (l_{ij}, m_{ij}, u_{ij})$, where a_{ij} is a member of a triangular fuzzy number. The TFN used to construct fuzzy evaluation matrix on ECFPR method. If the strong importance of element j over element i holds, then the pairwise comparison scale can be represented by the fuzzy number where $0 < l_{ij} < m_{ij} < u_{ij}$, for all i, j = 1, 2, …, n, j≠i. For each p_{ij} where i=j+1 or i=j, the elements can be transformed as (12).

$$p_{ij} = g(l_{ij}, m_{ij}, u_{ij}) = \frac{1}{2}(1 + \log_9(l_{ij}, m_{ij}, u_{ij})) \tag{12}$$

Proposition 7. For each $P = (p_{ij}) = (l_{ij}, m_{ij}, u_{ij})$, generally can be written as (13)

$$\begin{aligned}&(u,m,l)_{i(i+1)} + (u,m,l)_{(j+1)(i+2)} + \dots \\ &+ (u,m,l)_{(i+k-1)(i+k)} + (u,m,l)_{(i+k)i} = \left(\frac{k+1}{2}, \frac{k+1}{2}, \frac{k+1}{2}\right), \\ &\forall i < j.\end{aligned} \tag{13}$$

Experiment 2. In experiment 2, the fuzzy paired comparison matrix of the case criteria for shipping registration selection on economic sub-criteria (*X*) is a 4x4 matrix [16], [25]. Therefore, using the CFPR method, the number of comparisons to be used is only (= 4-1) pairs. Then a new model is formed as a result of the matrix transformation using the propositions that have been determined in the CFPR method.

The first step, transform *X* into *P* to describe the differences between AHP and CFPR method in the determining judgment. The formula will calculate the value of p_{ij}.

$$P = \begin{pmatrix} (1,1,1) & \left(\frac{2}{3}, 1, \frac{3}{2}\right) & p_{13} & p_{14} \\ p_{21} & (1,1,1) & \left(\frac{2}{5}, \frac{1}{2}, \frac{2}{3}\right) & p_{24} \\ p_{31} & p_{23} & (1,1,1) & \left(\frac{2}{5}, \frac{1}{2}, \frac{2}{3}\right) \\ p_{41} & p_{42} & p_{43} & (1,1,1) \end{pmatrix}$$

Matrix *P* shows that p_{ij} is matrix element entries that are not filled in by experts. The ECFPR method fills in values using proposition 9 and proposition 10 for each and so that the matrix output remains a triangular fuzzy number. The whole calculation is as follows;

$p_{11} = p_{22} = p_{33}$= ½ (1+log$_9$ (1, 1, 1) = (0.5, 0.5, 0.5),

p_{12} = ½ ((1,1,1)+log$_9$ (2/3, 1, 3/2) = (0.41, 0.5, 0.59),

p_{23} = ½ ((1,1,1)+log$_9$ (2/5, 1/2, 2/3) = (0.3, 0.35, 0.41),

p_{34} = ½ ((1,1,1)+log$_9$ (2/5, 1/2, 2/3) = (0.3, 0.35, 0.41),

p_{21} = (1,1,1) – (0.59, 0.5, 0.41) = (0.41, 0.5, 0.59),

p_{32} = (1,1,1) – (0.41, 0.35, 0.3) = (0.59, 0.65, 0.7),

p_{43} = (1,1,1) – (0.41, 0.35, 0.3) = (0.59, 0.65, 0.7),

p_{31} = (1.5,1.5,1.5) – ((0.59 + 0.41), (0.5 + 0.35),
(0.41+ 0.3)) = (0.5, 0.65, 0.79),

p_{42} = (1.5,1.5,1.5) – ((0.41 + 0.41), (0.35 + 0.35),
(0.3 + 0.3)) = (0.9, 0.8, 0.68),

p_{41} = (2,2,2) – ((0.59 + 0.41 + 0.41), (0.5 + 0.35 + 0.35),
(0.41 + 0.3 + 0.3)) = (0.59, 0.8, 0.99),

p_{13} = (1,1,1) – (0.79, 0.65, 0.5) = (0.21, 0.35, 0.5),

p_{24} = (1,1,1) – (0.9, 0.8, 0.68) = (0.1, 0.2, 0.32),

p_{14} = (1,1,1) – (0.9, 0.8, 0.59) = (0.1, 0.2, 0.32).

Table I depicts a new matrix of criteria that consists of four criteria. Maple 2016 Software calculated the consistency index using LFPP method.

Consistency index using LFPP method for matrix on Table I can be written as Minimize

$$J = (1-\lambda)^2 + M.\sum_{i=1}^{3}\sum_{j=i+1}^{4}\left(\delta_{ij}^2 + \eta_{ij}^2\right)$$

Subject to

$$\begin{cases} x_1 - x_2 - \lambda \ln(0.5/0.41) + \delta_{12} \geq \ln(0.41), \\ -x_1 + x_2 - \lambda \ln(0.59/0.5) + \eta_{12} \geq -\ln(0.59), \\ x_1 - x_3 - \lambda \ln(0.35/0.3) + \delta_{13} \geq \ln(0.3), \\ -x_1 + x_3 - \lambda \ln(0.5/0.35) + \eta_{13} \geq -\ln(0.5), \\ x_1 - x_4 - \lambda \ln(0.2/0.01) + \delta_{14} \geq \ln(0.01), \\ -x_1 + x_4 - \lambda \ln(0.41/0.2) + \eta_{14} \geq -\ln(0.41), \\ x_2 - x_3 - \lambda \ln(0.35/0.3) + \delta_{23} \geq \ln(0.3), \\ -x_2 + x_3 - \lambda \ln(0.41/0.35) + \eta_{23} \geq -\ln(0.41), \\ x_2 - x_4 - \lambda \ln(0.2/0.1) + \delta_{24} \geq \ln(0.1), \\ -x_2 + x_4 - \lambda \ln(0.32/0.2) + \eta_{24} \geq -\ln(0.32), \\ x_3 - x_4 - \lambda \ln(0.35/0.3) + \delta_{34} \geq \ln(0.35), \\ -x_3 + x_4 - \lambda \ln(0.41/0.35) + \eta_{34} \geq -\ln(0.41), \\ \lambda, x_1, x_2, x_3, x_4, \delta_{12}, \delta_{13}, \delta_{14}, \delta_{23}, \delta_{24}, \delta_{34} \geq 0, \\ \eta_{12}, \eta_{13}, \eta_{14}, \eta_{23}, \eta_{24}, \eta_{34} \geq 0. \end{cases}$$

Define a sufficiently large number for M=10^3, then the result of optimal value (λ^*) is 0.309. The extended CFPR produce different value of (λ^*) from traditional LFPP. Based on Wang and Chin (2011), it was noticed that the λ^* is 0. Some of the optimal values produced by the LFPP method are close to 0, while the calculation using the ECFPR method shows a cost of more than 0. However, further research is needed to see the pattern of λ^* values from the proposed method. Given an M=10^1, Table II shows the results of the comparison of the calculation of the optimal solution value in several paired matrix experiments on the appropriate shipping registry problem [16].

In this study, the examination of the consistency ratio (CR) value was done by the AHP method. Table II shows that matrices 10 and matrix 13 were inconsistent, with values of 0.23 and 0.27, respectively (more than 0.10). Line 10 is a paired comparison matrix on the labor quality and availability sub-criteria, while on the 13th line at the level of bureaucracy sub-criteria. So, it can be seen that only used 12 patterns to observed based on [16]. Table II shows that the values for the LFPP method vary from negative 1 to 1.

TABLE. I. ECFPR MATRIX

C	C_1	C_2	C_3	C_4
C_1	(0.5, 0.5, 0.5)	(0.41, 0.5, 0.59)	(0.21, 0.35, 0.5)	(0.01, 0.2, 0.41)
C_2	(0.41, 0.5, 0.59)	(0.5, 0.5, 0.5)	(0.3, 0.35, 0.41)	(0.1, 0.2, 0.32)
C_3	(0.5, 0.65, 0.79)	(0.59, 0.65, 0.7)	(0.5, 0.5, 0.5)	(0.3, 0.35, 0.41)
C_4	(0.59, 0.8, 0.99)	(0.68, 0.8, 0.9)	(0.59, 0.65, 0.7)	(0.5, 0.5, 0.5)

TABLE. II. CONSISTENCY INDEX COMPARISON

Matrix Number	CR	$M = 10^1$		$M = 10^2$		$M = 10^3$	
		a	*b*	*a*	*b*	*a*	*b*
1	0.02	0.36	0.66	-0.56	0.52	-0.83	0.5
2	0.07	0.73	-0.11	0.4	-0.38	0.31	-0.41
3	0.02	0.3	0.65	-0.82	0.15	-1.17	0.02
4	0.00	0.42	1	0.19	1	0.16	1
5	0.09	0.82	0.02	0.72	-0.56	0.71	-0.69
6	0.02	0.86	0.41	0.78	0.17	0.76	0.14
7	0.00	0.67	1	0.27	1	0.16	1
8	0.07	0.69	0.04	0.65	-0.28	0.65	-0.33
9	0.05	0.79	0	0.63	-0.36	0.6	-0.41
10	0.23	0.93	-1.15	0.91	-2.12	0.9	-2.26
11	0.03	0.8	0.34	0.69	0.16	0.66	0.13
12	0.09	0.96	-0.26	0.92	-1.21	0.91	-1.39
13	0.27	1.07	-0.96	1.13	-5.05	1.14	-6.64
14	0.03	0.29	0.4	0.14	0.14	0.12	0.1
Mean	-	0.67	0.35	0.34	0.03	0.26	-0.03

(a) Consistency index (λ^*) *of ECFPR*

(b) Consistency index (λ^*) *of LFPP*

The results show that there is no guarantee that the optimal value is always positive, depending on the *M*-value used. It can be seen the consistency index when $M = 10^2$ and 10^3 are still <0 in both methods. Whereas the values $M = 10^1$ in the ECFPR method produce the consistency index between 0.29 to 1, all are positive and selected for weight calculation. Based on Table II, the ECFPR mean is 0.67 has a higher value than LFPP (0.35). Therefore $M = 10^1$ can be used to get a higher consistency value.

B. Usability Evaluation Model

Flowchart of an ECFPR to evaluate e-commerce usability that consists of eight steps (Fig. 1). Determining usability criteria is the first step of evaluation. Developers and usability experts define essential rules in the assessment. The literature study activity often used to collect several papers relating to usability e-commerce, then looks for the right criteria in measurement. After choosing the proper rules, the next is to build a model hierarchy based on the taxonomy specified. The next step is to determine the weight of each measure with the help of experts. Experts will provide the level of importance of each test, using the CFPR.

If the expert judgment is considered to be consistent, the weight of each criterion is calculated using the LFPP method. Each weight is then used to calculate the website usability score, and recommendations based on the calculation results. Each block diagram was explained accompanied by a sample case in the following sub-chapters.

Research conducted by Amerson, et al. [20], the website usability performance can be presented by webpage loading time, average server response time, and webpage size in byte. The quality standard of loading time is less than 30 second, response time is less than 0.5 second, and page size must be less than 64 Kbytes. Pingdom and Bitcatcha can be used to measure these three criteria of usability. Average response time is the estimated response time of user interface when the user sends a request to a server. Webpage loading time is the estimated time required to bring up the website page (updated every month). The total page size is one of the essential criteria to optimize the webpage. It is used to estimate the rendering time of a webpage. The larger the size of the page, the longer the rendering time in most cases.

The hierarchical structure that consists of three levels, where the top level represents the goal and the lowest level has the website under consideration (Fig. 2). The decision maker assesses the importance of the criteria described in level 2, namely loading time (C_1), response time (C_2), and page size (C_3). Five of popular e-commerce websites from Indonesia was selected to assess their usability performance. There are Lazada, Blibli, Shopee, JDid, and Mataharimall.

This research used membership function linguistic scale based on Table III. Table IV shows the fuzzy comparison matrix of expert judgment from Decision-Maker (DM). The DM only requires *n*-1 (=3-1) comparison rating from three criteria. Loading time is fairly strong important than response time, but equally important than the page size.

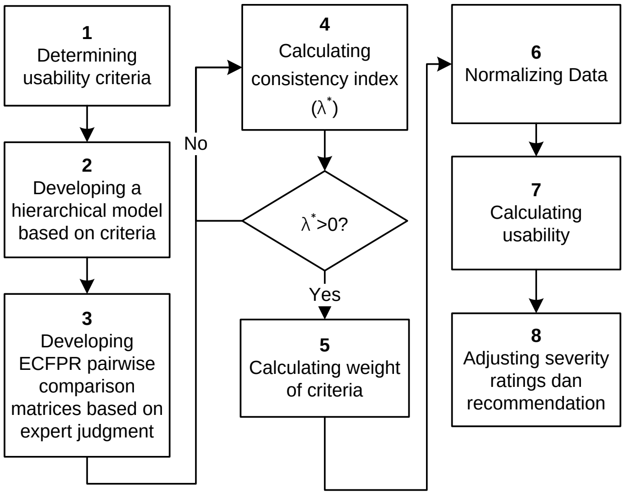

1
Determining usability criteria
2
Developing a hierarchical model based on criteria
3
Developing ECFPR pairwise comparison matrices based on expert judgment
4
Calculating consistency index (λ^*)
λ^*>0?
No
Yes
5
Calculating weight of criteria
6
Normalizing Data
7
Calculating usability
8
Adjusting severity ratings dan recommendation

Fig. 1. Usability Evaluation using ECFPR Method.

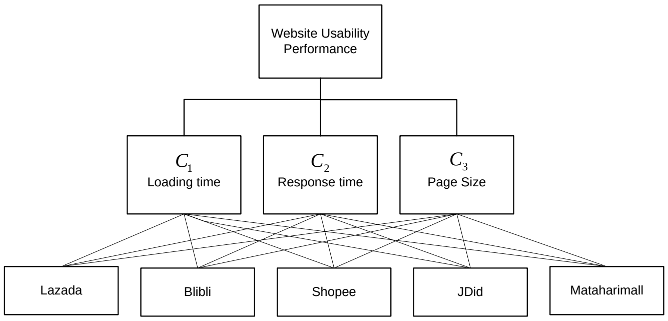


Fig. 2. The Hierarchical Model of Website usability Evaluation.

TABLE. III. MEMBERSHIP FUNCTION LINGUISTIC SCALE [6]

Convert from AHP scale to Fuzzy number	Linguistic expressions
1 = (1,1,1)	Equal
2 = (1,2,3)	Equal-moderate
3 = (2,3,4)	Moderate
4 = (3,4,5)	Moderate-fairly strong
5 = (4,5,6)	Fairly strong
6 = (5,6,7)	Fairly strong-very strong
7 = (6,7,8)	Very strong
8 = (7,8,9)	Very strong-absolute
9 = (8,9,9)	Absolute
2,4,6,8 Values between two adjacent assessments	

The next step is to fill the section of p_{ij} using ECFPR formula (12) and (13). Table IV represents the matrix transformation result.

TABLE. IV. EXPERT JUDGMENT COMPARISON MATRIX

Criteria	C_1	C_2	C_3
C_1	(1, 1, 1)	(4, 5, 6)	p_{13}
C_2	p_{21}	(1, 1, 1)	(1, 1, 1)
C_3	p_{31}	p_{32}	(1, 1, 1)

$p_{11} = p_{22} = p_{33}$= ½ (1+log$_9$ (1, 1, 1) = (0.5, 0.5, 0.5),

p_{12} = ½ ((1,1,1)+log$_9$ (4, 5, 6) = (0.82, 0.87, 0.91),

p_{23} = ½ ((1,1,1)+log$_9$ (1, 1, 1) = (0.5, 0.5, 0.5),

p_{21} = (1,1,1) – (0.91, 0.87, 0.82) = (0.09 ,0.13, 0.18),

p_{31} = (1.5,1.5,1.5) – ((0.91+0.5), (0.87 + 0.5), (0.82+ 0.5)) = (0.09, 0.13, 0.18),

Based on Table V, LFPP method (2) used to calculate the consistency index (λ^*), can be expressed as Minimize.

$$J = (1-\lambda)^2 + M\sum_{i=1}^{2}\sum_{j=i+1}^{3}\left(\delta_{ij}^2 + \eta_{ij}^2\right)$$

Subject to

$$\begin{cases} x_1 - x_2 - \lambda \ln\left(0.87/0.82\right) + \delta_{12} \geq \ln(0.82), \\ -x_1 + x_2 - \lambda \ln\left(0.91/0.87\right) + \eta_{12} \geq -\ln(0.91), \\ x_1 - x_3 - \lambda \ln\left(0.87/0.82\right) + \delta_{13} \geq \ln(0.82), \\ -x_1 + x_3 - \lambda \ln\left(0.91/0.87\right) + \eta_{13} \geq -\ln(0.91), \\ x_2 - x_3 - \lambda \ln\left(0.5/0.5\right) + \delta_{23} \geq \ln(0.5), \\ -x_2 + x_3 - \lambda \ln\left(0.5/0.5\right) + \eta_{23} \geq -\ln(0.5), \\ \lambda, x_1, x_2, x_3 \geq 0, \\ \delta_{12}, \eta_{12}, \delta_{13}, \eta_{13}, \delta_{23}, \eta_{23} \geq 0. \end{cases}$$

Define $M = 10^1$, the result value of λ^*, δ^*, η^*, and x can be represented as follows:

λ^* = 0.76, δ^*_{12}= 8.16 × 10^{-10}, η^*_{12}= 0.2229, x_1= 0.9060, δ^*_{13}= 0.2229, η^*_{13}= 2.622, x_2= 0.8118, δ^*_{23}= 8.94 × 10^{-10}, η^*_{23} = 0.2229, x_3 = 1.282. The result shows that the consistency index is 0.76 more than 0 or consistent. Thus the next step can be continued.

The weight derivation based on [16] of each criterion is calculated by (3). So, the weight can be shown as below.

$$w_1^* = \exp(0.906)/((\exp(0.906) + \exp(0.811) + \exp(1.282)) = 0.297,$$

$$w_2^* = \exp(0.811)/((\exp(0.906) + \exp(0.811) + \exp(1.282)) = 0.270,$$

$$w_3^* = \exp(1.282)/((\exp(0.906) + \exp(0.811) + \exp(1.282)) = 0.432.$$

Thus, it can be seen that the weight of response time is 0.297, webpage load time is 0.270, and webpage size is 0.432.

TABLE. V. ECFPR MATRIX COMPARISON

Criteria	C_1	C_2	C_3
C_1	(0.5, 0.5, 0.5)	(0.82, 0.87, 0.91)	(0.82, 0.87, 0.91)
C_2	(0.09, 0.13, 0.18)	(0.5, 0.5, 0.5)	(0.5, 0.5, 0.5)
C_3	(0.09 ,0.13, 0.18)	(0.5, 0.5, 0.5)	(0.5, 0.5, 0.5)

IV. Result and Discussion

These five sites were analyzed for one month (4th May to 4th June 2019). Pingdom used for collecting loading time (second) and page size (MB). Response time data collected using Bitcatcha (ms). Table VI shows the original data for five alternatives.

The original data is then normalized using the Linear Weightage Model [6], [26]. Load time, response time and page size are the maximum criteria because the smallest size is the best. Equation (14) used to normalize the original data.

$$r_{ij} = \frac{\max - x_{ij}}{\max - \min}, \text{for maximum threshold}, \tag{14}$$

where r_{ij} is the normalized criteria, max is the maximum value of the particular criteria among all websites, min is minimum value of the same criteria among the whole websites, and x_{ij} is the specific website that is considered at the time.

Table VII shows normalized data and the final result of extended CFPR method on accessing the usability of e-commerce websites. Equation (15) used to calculate usability score, where l is the number of alternatives, n is the number of criteria, r_{ij} is normalized value, and w_j is the weight of criteria.

$$\text{usability score} = \sum_{i=1}^{l}\sum_{j=1}^{n} w_j \times r_{ij} ; i = 1,2,...,l,\ j = 1,2,...,n \tag{15}$$

Mataharimall has the highest usability score (0.62), and Blibli has the lowest (0.27). The usability score can then be used as a reference recommendation to the developers. Blibli has a high loading time, but the response time is low. Therefore speed factor needs to be considered when it comes to website design improvements. The value of severity ratings also calculated so that the developers can know the extent of the level of seriousness of the website [27]. Severity rating assigned on a 4-point scale (1 = irritant, 2 = moderate, 3 = severe, 4 = unusable). Next, presentage of severity was assigned also on a 4-point scale (1= less than 10 percent ; 2 = 11 to 50 percent; 3 = 51 to 89 percent ; 4=more than 90 percent). Severity rating can be measured as (100%-(usability scores*100)). The best-ranking law in the ECFPR method is Mataharimall > Shopee > JDid > Lazada > Blibli. Mataharimall and Shopee have a moderate category. Lazada, Blibli, and Jdid have a severity value between 11-50% (severe type).

TABLE. VI. Original Data

Criteria	Lazada	Blibli	Shopee	JDid	Mataharimall
C_1	6.6	1.23	0.756	4.46	3.46
C_2	256.875	543	258.625	371.5	365
C_3	4.9	1.9	2.4	1.9	0.142578

TABLE. VII. Normalized Data

Criteria	Lazada	Blibli	Shopee	JDid	Matahari mall	Weight
C_1	0.00	0.92	1.00	0.37	0.54	0.297
C_2	1.00	0.00	0.99	0.60	0.62	0.27
C_3	0.00	0.63	0.53	0.63	1.00	0.432
Usability Score	0.30	0.27	0.52	0.45	0.62	-
Severity Rating	70.3	72.76	47.78	54.96	38.32	-
Category	severe	severe	moderate	severe	moderate	-

V. Conclusion

E-Commerce website usability can be evaluated from some criteria such as load time, response time, and page size. The usability evaluation framework was constructed using ECFPR to develop the fuzzy pairwise comparison matrix. This framework also used to determine the proper method of evaluating website usability performance. The numerical experiment showed that the consistency index obtained by ECFPR method was more significantly better than LFPP method. It was revealed that the optimal value always more than 0. The ECFPR method was also successfully implemented with the experimental case to evaluate the usability of five e-commerce website in Indonesia. For further study, criteria, and alternatives to testing whether the method works well can be added.

Acknowledgment

We thank specially to Mr. Handaru Jati, Ph.D from Universitas Negeri Yogyakarta Indonesia who gives the expert judgment about usability criteria and Mr. Ridwan Pandiya, our colleagues from Institut Teknologi Telkom Purwokerto Indonesia who provided insight and expertise that greatly assisted the research.

References

[1] H. R. Hartson, T. S. T. Andre, dan R. R. C. Williges, “Criteria for evaluating usability evaluation methods,” Int. J. Hum. Comput. Interact., vol. 13, no. 4, hal. 1–35, 2001.

[2] R. Y. Naswir, “Towards a Conceptual Model to Evaluate Usability of Digital Government Services in Malaysia,” Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 4, hal. 313–322, 2019.

[3] J. S. Challa, A. Paul, Y. Dada, V. Nerella, dan P. R. Srivastava, “Quantification of Software Quality Parameters Using Fuzzy Multi Criteria Approach,” in International Conference on Process Automation, Control and Computing, 2011, hal. 1–6.

[4] S. K. Dubey dan A. Mittal, “Measurement of Object Oriented Software Usability using Fuzzy AHP,” Int. J. Comput. Sci. Telecommun., vol. 3, no. 5, hal. 98–103, 2012.

[5] S. K. Dubey dan S. Pandey, “Measurement of Usability of Office Application Using a Fuzzy Multi-Criteria Technique,” Int. J. Inf. Technol. Comput. Sci., vol. 7, no. 4, hal. 64–72, 2015.

[6] P. D. D. Dominic dan H. Jati, “A comparison of Asian airlines websites quality : using a non-parametric test,” Int. J. Bus. Innov. Res., vol. 5, no. 5, 2011.

[7] I. Masudin dan T. E. Saputro, “Evaluation of B2C website based on the usability factors by using fuzzy AHP & hierarchical fuzzy TOPSIS,” in IOP Conference Series: Materials Science and Engineering, 2016, vol. 114, hal. 1–8.

[8] S. Aydin dan C. Kahraman, “Evaluation of E-commerce website quality using fuzzy multi-criteria decision making approach,” IAENG Int. J. Comput. Sci., vol. 39, no. 1, hal. 64–70, 2012.

[9] N. Sehra, Sumeet Kaur Brar, Yadwinder Singh Kaur, “Applications of Multi-criteria Decision Making in Software Engineering,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 7, 2016.

[10] L.-X. Wang, “A Course in Fuzzy Systems and Control.” Prentice Hall, Hongkong, hal. 1–441, 1997.

[11] R. Wardoyo dan T. Wahyuningrum, “University Website Quality Ranking Using Logarithmic Fuzzy Preference Programming,” Int. J. Electr. Comput. Eng., vol. 8, no. 5, hal. 3349–3358, 2018.

[12] L. Mikhailov, “A fuzzy programming method for deriving priorities in the analytic hierarchy process,” J. Oper. Res. Soc, vol. 51, hal. 341–349, 2000.

[13] L. Mikhailov, “A fuzzy approach to deriving priorities from interval pairwise comparison judgements,” Eur. J. Oper. Res., vol. 159, no. 3, hal. 687–704, 2004.

[14] E. Iryanti dan R. Pandiya, “Application of Logarithmic Fuzzy Preference Programming for Determining Priority as An Institutional Development Strategy,” in International Conference on Computer Applications and Information Processing Technology (CAIPT), 2017, hal. 1–5.

[15] Y. M. Wang dan K. S. Chin, “A linear goal programming priority method for fuzzy analytic hierarchy process and its applications in new product screening,” Int. J. Approx. Reason., vol. 49, no. 2, hal. 451–465, 2008.

[16] Y. M. Wang dan K. S. Chin, “Fuzzy analytic hierarchy process: A logarithmic fuzzy preference programming methodology,” Int. J. Approx. Reason., vol. 52, no. 4, hal. 541–553, 2011.

[17] S. Dožić, T. Lutovac, dan M. Kalić, “Fuzzy AHP approach to passenger aircraft type selection,” J. Air Transp. Manag., vol. 68, 2017.

[18] M. Balouchi dan E. Khanmohammadi, “Using logarithmic fuzzy preference programming to prioritization social media utilization based on tourists’ perspective,” Found. Manag., vol. 7, no. 1, hal. 7–18, 2015.

[19] M. Momeni, A. Sasani, M. R. Fathi, dan E. Khanmohammadi, “Applying Logarithmic Fuzzy Preference Programming for Ranking of Effective Organizational Factors on Creativity : A Case Study Mansour Momeni Faculty of Management University of Tehran,” Int. J. Bus. Soc. Sci., vol. 3, no. 14, hal. 83–95, 2012.

[20] E. Herrera-Viedma, F. Herrera, F. Chiclana, dan M. Luque, “Some issues on consistency of fuzzy preference relations,” Eur. J. Oper. Res., vol. 154, no. 1, hal. 98–109, 2004.

[21] T. L. Saaty, “How to make a decision : The Analytical Hierarchy Process,” Eur. J. Oper. Res., vol. 48, hal. 9–26, 1990.

[22] R. W. Saaty, “The analytic hierarchy process-what it is and how it is used,” Math. Model., vol. 9, no. 3–5, hal. 161–176, 1987.

[23] H. Khademi-Zare, M. Zarei, A. Sadeghieh, dan M. Saleh Owlia, “Ranking the strategic actions of Iran mobile cellular telecommunication using two models of fuzzy QFD,” Telecomm. Policy, vol. 34, no. 11, hal. 747–759, 2010.

[24] R. J. Chao dan Y. H. Chen, “Evaluation of the criteria and effectiveness of distance e-learning with consistent fuzzy preference relations,” Expert Syst. Appl., vol. 36, no. 7, hal. 10657–10662, 2009.

[25] M. Celik, I. D. Er, dan A. F. Ozok, “Application of fuzzy extended AHP methodology on shipping registry selection : The case of Turkish maritime industry,” Expert Syst. Appl., vol. 36, no. 1, hal. 190–198, 2009.

[26] A. A. Ali dan P. D. D. Dominic, “A Case Study of Linear Weightage Model for Supplier Selection Process,” 2008 Int. Symp. Inf. Technol., vol. 3, hal. 23–26, 2008.

[27] T. Tullis dan B. Albert, Measuring The User Experience. USA: Morgan Kaufmann, 2013.

Socialization of Information Technology Utilization and Knowledge of Information System Effectiveness at Hospital Nurses in Medan, North Sumatra

Roymond H. Simamora
Faculty of Nursing
Universitas Sumatera Utara
Jl. Prof T. Ma As No 3. Medan 20155

***Abstract*—Background of this research is the globalization and development of science, especially in the field of information and communication technology and communication that has influenced and has implications for changes and renewal of people's lives, including in the field of nursing. So that the role of information and communication in this aspect of life is very important, even the futurists, for the most part, have an agreement that one of the most important strengths as the source of future power is information. Purpose: identify the use of information technology in nursing to determine the effectiveness of the use of information systems in nursing, identify nurses 'knowledge about the effectiveness of nursing information systems, identify nurses' knowledge seen from the socialization of the effectiveness of nursing information systems. Method: Quantitative Research Type with a survey approach conducted on 220 nurses. Significant validity test is <0.05, Cronbach Alpha reliability test> 0.60. The data is then tested in a classic assumption test consisting of multicollinearity tests, autocorrelation tests, heteroscedasticity tests, normality tests, multiple linear regression, t-tests, F tests, coefficient of determination tests. Results: the use of information technology affects the effectiveness of nursing information systems. Nurse knowledge does not affect the effectiveness of nursing information systems. Nurse knowledge seen from socialization does not affect the effectiveness of nursing information systems. The use of information technology and nurse knowledge influences the effectiveness of nursing information systems. The results of the coefficient of determination that affect the use of information technology, knowledge of nurses, socialization as a control variable on the effectiveness of nursing information systems. Suggestion: Hospital managers must pay attention to the quality of nursing human resources, through training, certification, recognition of competencies, supervision, selection, and guidance aimed at improving safe, comfortable and satisfying services for patients, families, communities.**

***Keywords*—Information systems; knowledge; nursing; socialization**

I. Background

Information systems are computer systems that collect, store, process, retrieve, show, and communicate information needed in practice, education, administration and. Many benefits are obtained in the use of information systems. This benefit not only reduces errors and increases speed and accuracy in care, but also reduces health costs by coordinating and improving service quality. The rapid progress and development of information needs, especially technology in the globalization era, has had a significant influence on the application of information systems in the health sector. Health Information System is a set of arrangements that includes data, information, indicators, procedures, tools, technology, and human resources that are interrelated and managed in an integrated manner that provides information support for decision making processes, health program planning, implementation of monitoring and evaluation at every level of health administration Health information systems are a very important part of improving hospital efficiency and supporting competitiveness by providing health service information for management [1].

Nursing services in a hospital environment is one of the services in the health sector that has an important role in determining the success of services provided in hospitals. With the highest number of nurses in the hospital environment, the existence of nursing services must be managed properly to produce the quality of nursing services provided. Improving the quality of nursing information systems is one solution to improve the quality of nursing services. It is undeniable that so far the development of nursing information systems in this country has not been going well. Nursing information systems are a combination of computer science, information science and nursing science that are organized to facilitate management and the process of gathering information and knowledge used to support the implementation of nursing care. Meanwhile, according to America Nursing Asociation [2] nursing information systems relating to legality to obtain and use data, information and knowledge about documentation, communication standards, support the decision making process, develop and disseminate new knowledge, improve quality, effectiveness and efficiency of care and empower patients to choose the desired health care. The reliability of an information system in an organization lies in the interrelationship between existing components so that it can be produced and flowed into useful, accurate, reliable, detailed, fast, relevant information for an organization. This information system is expected to improve service quality in achieving service quality standards. Clinical indicators of service quality include: measurement of decreased patient rates, pressure sores, nosocomial pneumonia, nosocomial infections, and incidences of medical errors. This computer-based information system will identify various types of patient needs, ranging from care documentation, medication documentation, to financial calculations that must be paid by patients for the care they have received [3]. Many people complain about the health services they receive from nurses.

For this reason, nurse performance needs to be improved so that the quality of care services can be provided properly. One measure of whether quality health services are provided to the community is the level of satisfaction for the people who receive the services themselves. One of the benefits of implementing a nursing information system in a hospital setting is to help nurses document nursing care. Nursing care in meeting the basic needs of patients is provided by nurses in various health care settings using the nursing process. Nurses use nursing information systems with the aim to clearly assess patients, prepare nursing plans, document nursing care, and to control the quality of nursing care. Nurses can have an integrated view of data (for example integration between nurses and doctors in patient care plans) [4]. By utilizing nursing information systems nurses can save time to do the recording compared to when done manually recording. In addition, data recorded using a nursing information system will be more secure. The risk of recorded data will be lost is very small. In contrast to paper-based records, where the possibility of data loss is very likely to occur. In addition, the existence of a nursing information system will also increase the effectiveness and efficiency of nursing staff work. Implementation of nursing information systems in hospitals, which combines computer science, information science, and nursing science that is designed to facilitate management and the process of collecting data, information, and knowledge to support the implementation of nursing care [5]. Nursing information systems are being developed on an ongoing basis in the future nursing knowledge will depend on the ability of information systems to facilitate the results of diagnosis, management, research, education, information exchange, and collaboration / collaboration, that the integration of nursing, computer science can be used to identify, collect, process, organize data and information to support nursing practice, administration, education, research, and development of nursing science [5]. The need for a management information system supports nurses in assisting decision making. Advances in technology in hospitals enable nurses to use management information systems to support the delivery of nursing care, so that better quality nursing care is achieved [6].

According to [7] the research focused on exploring Computerized Provider Order Entry (CPOE) and its impact on the work done by nurses. The result is that CPOE is a technology designed to replace paper entry, communication and coordination with automated methods, one of which is in collaborative collaboration to prescribe drugs in acute care. CPOE has been proven to improve communication efficiency and reduce drug transcription errors and reduce patient care time, so the patient's morbidity and mortality are reduced. Actually to implement a nursing information system in a hospital environment is not too difficult to implement, only a commitment to implement it is needed. In times of technology such as now, maybe almost all nurses can operate a computer as a device in the application of nursing information systems. This is a very large capital that strongly supports the implementation of nursing information systems.

Now the only problem is how we are committed, from top management to the lowest management to fight for the implementation of nursing information systems in each nursing service unit. The reason for the lack of availability of funds to develop nursing information systems is a classic reason that should no longer exist. Especially seeing the importance of nursing information systems to improve the quality of nursing services in particular and health services in general. Based on the above problem, this research was conducted with the aim of 1) identifying the use of nursing information technology on the effectiveness of nursing information systems, 2) identifying nurses 'knowledge about the effectiveness of nursing information systems, 3) identifying nurses' knowledge as seen from the socialization of the effectiveness of nursing information systems. The use of information technology is the use of technology in the Nursing Nursing Service System Nursing is the understanding of nurses in the operation of Nursing Information System software to provide accurate and timely results in care reports so that they have an increasing impact on the Hospital. Socialization is the dissemination of information about Nursing Information Systems conducted by hospitals to nurses in operating information systems. The effectiveness of the Nursing Information System is the operation of the Nursing information system in the hospital with ease of use, accurate results, and timeliness.

II. Method

This research is quantitative, use a survey research, with a sample of 220 nurses from several hospitals in Medan. The instrument used was tested for validity and reliability, the results of the analysis showed all items were valid because each indicator showed a significant result that was <0.05. The reliability test results showed that all research instruments were reliable because each instrument produced a Cronbach Alpha >0.60. The data is then tested in a classic assumption test consisting of multicollinearity tests, autocorrelation tests, heteroscedasticity tests, normality tests, multiple linear regression, t tests, F tests, coefficient of determination tests.

The hypothesis in this study was formulated as follows:

H1: The use of information technology has a significant positive effect on the effectiveness of nursing information systems.

H2: Nurse knowledge has no significant effect on the effectiveness of nursing information systems

H3: Nurse knowledge seen from socialization has no significant effect on the effectiveness of nursing information systems.

III. Result and Discussion

A. Descriptive Analysis

Descriptive analysis shown in Table I, results about the use of information technology on average showed 4.39 where the average respondent answered agree, the nurse's knowledge showed an average value of 4.55 where the respondent responded most agreed, socialization showed an average value of 1.9 where the average respondent disagrees. participate in the socialization of Nursing Information Systems. The effectiveness of the nursing information system shows an average value of 4.18 where the nurse answers the average in the agreement.

B. *The Classic Assumption Test*

In Table II, test results show that there is no classic consideration problem. In the normality test shows the value of sig. equal to 0.715 > 0.05 which means that the data are declared normally distributed. Multicollinearity values indicate if tolerance values >0.1 and VIF <10, which means there are no multicollinearity problems. The autocorrelation test results showed a probability value of 0.818 > 0.05. Heteroscedasticity test results >0.05 which shows no heteroscedasticity problems.

C. *Multiple Linear Regression Analysis*

This analysis is used to examine the effect of utilizing information technology, nurse knowledge, and socialization. Based on data processing using the SPSS program, the following results are obtained as shown in Table III.

TABLE. I. DESCRIPTIVE ANALYSIS

Variabel	Min	Max	Mean	Std. Deviation
Utilization of Information Technology	4	5	4,39	0,32
Knowledge of Nurses	1	5	5	0,49
Socialization	0	12	1,9	2,82
Nursing Information Systems	3	5	4,18	0,37

TABLE. II. CLASSICAL ASSUMPTION TEST

Variable	Multicollinearity		Heteroske-dasticity	Normality
	Tolerance	VIF		
Utilization of Information Technology	0,569	1,757	0,757	0,715
Knowledge of Nurses	0,491	2,035	0,373	
Socialization	0,822	1,217	0,527	

TABLE. III. MULTIPLE LINEAR REGRESSION TEST

Model	β	t	Sig
a costant	11,852	0,682	0,505
Utilization of Information Technology,	1,319	0,504	0,19
Knowledge of Nurses,	0,038	0,64	0,950
Socialization	0,134	0,282	0,782
Adjusted R^2	0,324		
F count	4,042		
Sig. F	0,026		

Based on these results it can be estimated that multiple linear regression is as follows: Y = 11,852 + 1,319 X1 + 0.038 X2 + 0.134X3

Interpretations of the regression equation are:

a: 11,852 means that the use of information technology (X1), knowledge of nurses (X2) and socialization (X3), is equal to zero, so the utilization of information systems (Y) is equal to 11,852.

b1: 1,319 which means that the influence of nurse knowledge (X2) and socialization (X3) variables on the information system capability (Y) is positive, which supports the ease of information technology increased by one (unit), can improve nursing information systems (Y) by 1,319 with nurses' knowledge assumptions (X2), and socialization (X3) are considered permanent.

b2: 0.038 means that the influence of the use of information technology (X1), socialization (X3) on the ability of a positive accounting information system (Y), which supports nurses' knowledge to increase one (unit), can improve nursing information systems (Y) by 0.038 with the assumption information technology (X1) and socialization (X3) are considered permanent.

b3: 0.134 means that the use of information technology (X1), nurse knowledge (X2) for accounting information systems (Y) is positive, which supports the socialization of an increase in one (unit), so as to improve nursing information systems (Y) by 0.134 assuming benefits information technology (X1) and nurse knowledge (X2) are considered permanent.

D. *Coefficient of Determination (Adjusted R Square)*

Based on the SPSS calculation results, the adjusted R square value is 0.324. This shows 32.40 percent, the variable utilization of information technology, knowledge of nurses, and socialization of the effectiveness of nursing information systems. While the rest (100% -32.4%) = 67.6% is explained by other variables not included in this regression model or not examined in this study.

E. *F Test*

Based on the calculation of multiple linear regression shows that Sig. F = 0.026 <α 0.05. This reflects if the model used in this study is suitable. The use of information technology, nurse knowledge and socialization simultaneously has a significant influence on the effectiveness of nursing information systems.

F. *Utilization of Information Technology about the Effectiveness of Nursing Information Systems*

The results of the probability value of 0.019 < 0.05 then H1 is rejected as a positive information technology that supports the use of nursing information systems. The results of this study are supported by research conducted by [8] in which the results obtained are related to positive technology for the use of information systems. Thus, the higher the level of information technology usage, the higher the level of nursing information system requirements. Therefore, the results of this study support the results of research conducted by [8] and [9].

Utilization of technology in general to process data, process, store, obtain, display, and send in various forms and ways used to produce benefits that can benefit the user. The information obtained can assist nurses in solving problems, solving problems, and evaluating them, making the information obtained must be of high quality. Quality information must be accurate, relevant, timely. Accurate means error free, not biased because it denies that biased information can mislead the recipient or user of that information [10]. Utilization of technology produces a number of technologies that are considered temporary in the sophistication of information seen from the nature of the application portfolio. The presence of technology is a source of strength that makes a company have a competitive advantage, and is identified as a factor that contributes to the company's success. Therefore, information technology has a high influence on the company's success in managing the company. In order for more sophisticated information technology to be applied, the effectiveness of the information system produced will be higher and the influence of information users [11].

G. Nurse Knowledge about the Effectiveness of Nursing Information Systems

The results of the probabily value 0.950 > 0.05 then H2 is accepted so that the nurse's knowledge has no significant positive effect. The results of this study are not supported in research [8] and [12] because the results of this study have a significant effect and reading employee knowledge influences the effectiveness of the information system. Nurses in their knowledge are still weak and do not understand nursing information systems techniques so that the results of the information are still not timely and inaccurate so that the goals for the hospital have not been reached to the fullest. Can be seen if the higher the level of knowledge of nurses, the higher the level of effectiveness of nursing information systems.

H. H.Nurses' Knowledge can be Seen from the Socialization of the Effectiveness of Nursing Information Systems

The probability value of 0.782 > 0.05 then H3 is accepted so that the nurse's knowledge does not have a significant positive effect. the results of this study are not appropriat [14] a positive relationship between user training, user attitude and success. Nurse knowledge seen from socialization does not affect the effectiveness of nursing information systems because nurses to socialize about nursing information systems are still many who do not follow the socialization so that the nursing information system is generated from nursing information to various users both internal and external parties. still not enough. According to [13] that socialization is usually done when workers lack expertise or when an organization changes the system and needs to learn about new skills. Nurses, when participating in nursing information system socialization, will produce effective nursing information systems for hospitals that can provide added value to users in various nursing information for planning, control and decision making activities, which in turn has an impact on improving hospital performance by whole [15].

IV. Conclusion

The results of the partial hypothesis that: the use of information technology has an effect on the effectiveness of nursing information systems, in this study the results have a significant effect that is reading the use of information technology has an effect on the effectiveness of nursing information systems. Nurse knowledge influences the effectiveness of nursing information systems, in this study the results did not have a significant effect so that the knowledge of nurses reading did not affect the effectiveness of nursing information systems. Nurse knowledge seen from the effect of socialization on the effectiveness of information systems, in this study did not have a significant effect so that the knowledge of nurses reading from socialization does not affect the effectiveness of nursing information systems. Simultaneous hypothesis that: the use of information technology and knowledge of nurses affect the effectiveness of nursing information systems, in this study the results have a significant effect that is reading the use of information technology and knowledge of nurses affect the effectiveness of nursing information systems. The results of the coefficient of determination that affect the use of information technology, knowledge of nurses, socialization as a control variable on the effectiveness of nursing information systems. Suggestion: Hospital managers must pay attention to the quality of nursing human resources, through: Ongoing training, certification, competency recognition, supervision, selection, and guidance aimed at improving safe, comfortable and satisfying services for patients, families, communities.

References

[1] Alsarayreh M.N., Replyreh O.A., Jaradan M.F., and Alamro S.A, 2011, Impact of Technology on the Effectiveness of Accounting Information Systems (AIS) Applied by Aqaba Tourist hotels. European Journal of Scientific Research, p. 361-369.

[2] Jumaili Salman, 2005, Trust in New Information Systems Technology in Evaluating Individual Performance, Collection of Materials for the VIII National Symposium, Solo, 15-16 September 2005.

[3] Baig, A.H. and Gururajan, R, 2011, Preliminary Study to Investigate Determinant Factors that Affect SI / IT Outsourcing, Journal of Information and Communication Technology Research, 1 (2), p. 48-54.

[4] Lestari, Endah Sri, et al. 2016. Evaluation of Health Information Systems in Central Java Province in the context of Strengthening the National Health Information System. Semarang: Journal Indonesian Health Management, Volume 4 No. 3

[5] Indari 2015. Effect of Application of Technology-Based Management Information Systems (SIM) for Child Care on Knowledge of Standard Operating Procedures (SOP) in the Treatment Room at Saiful Anwar Hospital Malang. Malang: Journal of Health Hesti Wira Sakti, Volume 3, Number 3

[6] Widjajanto Nugroho, 2001, Accounting Information Systems, Erlangga: STIE Trisakti.

[7] Zubaidah 2011. The Role of Nursing Management Information Systems on Patient Safety in Child Nursing Jakarta.

[8] Sukma Putra, 2014, Employee Knowledge and Information Technology Utilization in Information Systems Effectiveness, E Journal of Accounting, Ganesha Educational University, Vol: 2 No. 1 of 2014.

[9] Ratnaningsih and Agung, 2014, The Effect of Information Technology Sophistication, Management Participation, and Knowledge of Accounting Managers on the Effectiveness of Information Systems, E Journal of Accounting, Udayana University, ISSN 2302-8550.

[10] Hamza. 2016. Design and Build a Nursing Care Information System for Patients with Pneumonia. Yogyakarta: Journal of Information Systems (JSI), VOL. 8, NO. 1

[11] Gomez-Mejia L.R, Balkin, D.B. & Cardy, R.L, 2011, Managing Human Resources, International Edition. Prentice Hall International, Inc.

[12] Ratnaningsih, 2013, The Influence of Information Technology Sophistication, Management Participation, Management Participation Knowledge, and Accounting Manager Knowledge on the Effectiveness of Information Systems in Starred Hotels in Badung Regency, Faculty of Economics, Udayana University.

[13] Ningsih, Ratna. 2010. Application of Nursing Information Systems in Complete Nursing Documentation at Hospitals. Jakarta

[14] Komala, 2012, The Influence of Knowledge of Accounting Managers and Top Management Support on Accounting Information Systems and Their Effects on Information Quality: Survey at the Zakat Management Institute in Bandung. Third International Conference on Business and Economic Research (Third Celebration 2012).

[15] Soudani S N, 2012, The Use of Information Systems for Effective Organizational Performance. International Journal of Economics and Finance, 4 (5), Pp: 136-145.

Line Area Monitoring using Structural Similarity Index

Supervising Car Reverse Test in Driving License Exam

Abderrahman AZI[1], Abderrahim SALHI[2], Mostafa JOURHMANE[3]
Information Processing and Decision Laboratory
Sultan Moulay Slimane University
Beni-Mellal, Morocco

***Abstract*—Real-time motion detection in specific area is considered the most important task in every video surveillance system. In this paper, a novel real time motion detection algorithm introduced to process Line zones called Line Monitoring Algorithm (LMA). This algorithm integrates Bresenham's Algorithm and Structural Similarity Index (SSI) in order to achieve the best performance. Bresenham's Algorithm is used to collect line pixels from two given points. Then, the SSI is used for real-time calculation of similarity for line motion detection. The most attractive side of using the LMA is that the algorithm does not need to compare all pixels of the whole images or regions for line areas monitoring. This algorithm has high capability, treatment speed and efficiency for motion detection and also demands less compilation time for the hardware performance. The main objective of this paper is to use a video surveillance system implementing LMA to supervise the Car Reverse Test (CRT) for driving license exam in Morocco. The evaluation of the experiment results in implementing the proposed algorithm is reported in this paper.**

Keywords*—*Bresenham's Algorithm; Structural Similarity Index; SSI; motion detection; Line Monitoring Algorithm; LMA; OpenCV; surveillance; camera; video surveillance system

I. Introduction

According to the World Health Organization's (WHO) 2015 [1] Road Safety Situation Report, nearly 1.25 million people die every year on the roads. In Morocco although the total number of registered vehicles (per capita) is low, the rate of road traffic deaths is high with approximately 10 people killed every day. 36% of the fatalities are car occupants and a further 21% are motorcycle riders.

Road Safety has become a matter of national security in Morocco. Many arrangements have been made by the government to decrease the number of car accidents such as installing Speed Radars, Road Signs and Speed Measuring Panels without neglecting the idea of improving driving license exam.

The driving license exam in Morocco is composed of two sections:

- Theoretical test: a sort of quiz which, according to the category desired, the candidate must answer correctly to a group questions (Ex: **32/40** for Cars category "**B**").
- Practical test: shows the ability of the candidate to drive, park and control the vehicle.

Before 2008, the theoretical test was in form of Asking/answering test between the supervisor and the candidate by using a pre-defined list of questions related to road situations, prerequisites and notions in order to evaluate the candidate's knowledge by calculating the number of correct answers, which must exceed the pass- average related to the chosen category. The pass average is specified by the law for each category (Ex 32/40 for Category "B": Automobile) when a candidate fails in the exam, he/she will be re-called for the second session after fifteen days for his/her last chance.

In 2008, according to the law, the Moroccan government has decided to employ technology in the theoretical test due to various problems. These problems are caused by kipping completely the judgment on supervisors hand besides the time and resources wasting with not much reliable results. Therefore, the proposed solution was to regroup candidates into sessions, each one passes the exam in front of a monitor related to remotes for each candidate. In so doing, it allowed them to answer separately to questions shown in the monitor and get the results by the end of the session. This arrangement made by the government gives more reliable results and solves a lot of problems. However, there were other problems caused by the system such as remotes crashes problems and unsaved results. Therefore, in 2010, computers having a screen-touch and cameras are used for the first time instead of the old system. Thus, each candidate can take anytime his exam separately with random questions. This would give credibility, transparency and efficiency to the test and evaluate objectively the candidate's abilities and prerequisites and moreover, to facilitate the job for the supervisor to become just an observer instead of being a judge with no pressure.

On the other hand, practice test did not change. That is, according to Moroccan law [2] the practice test divided on four stages for category "B" correspond to car driving :

Parking Car Test: The candidate must successfully park the vehicle in and out of a specific zone (Figure 1) without touching any bars surrounding that zone.

Entering Car to the garage Test: The candidate must successfully enter the vehicle in and out of a dedicated vertical zone (Figure 2) without touching any bars surrounding the zone which simulates the garage.

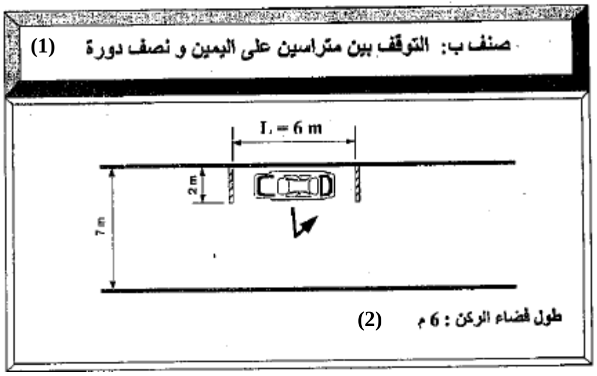


Fig. 1. Parking Car Test According to Moroccan Law N° A2709.10 of 30/09/2010 [2].

1) Title: category “B”: make a stop between two barriers on the right and do a half cycle.

2) Width of the parking space “L”.

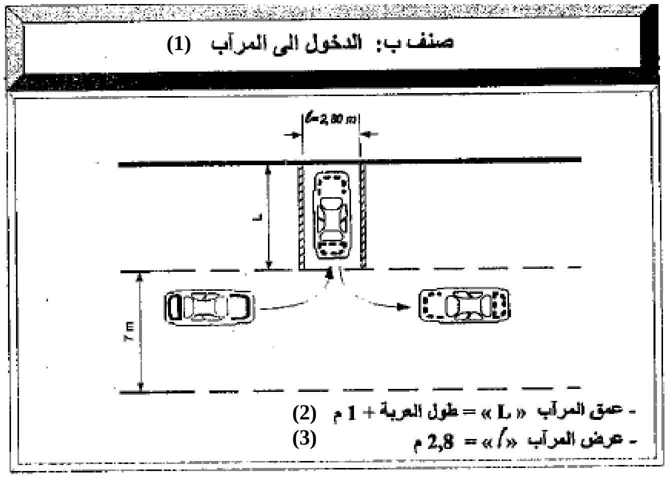

(1) صنف ب: الدخول الى المرآب
ℓ=2,80 m
L
7 m
(2) - عمق المرآب « L » = طول العربة + 1 م
(3) - عرض المرآب « ℓ » = 2,8 م

Fig. 2. Figure 1: Entering Car to the garage Test according to Moroccan law N° A2709.10 of 30/09/2010 [2].

1) Title: category “B”: entering to the garage.

2) Garage area height “L”.

3) Garage area width “l”

Car Reverse Test: The candidate should drive throw a dedicated zone in reverse without touching the surroundings of that zone or stopping the vehicle (Figure 3).

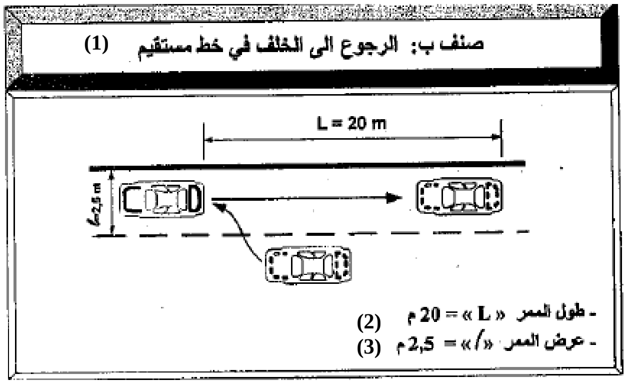

(1) صنف ب: الرجوع الى الخلف في خط مستقيم
L = 20 m
ℓ=2,5 m
(2) - طول الممر « L » = 20 م
(3) - عرض الممر « ℓ » = 2,5 م

Fig. 3. Car Reverse Test According to Moroccan law N° A2709.10 of 30/09/2010 [2].

1) Title: category “B”: driving backward in a straight line

2) Passage width “L” = 20m.

3) Passage height “ℓ” = 2.5m

Car Ride Test: It tests the candidate’s ability to drive the car and control it in order to evaluate his reactions and knowledge on the field.

As it is mentioned before, unlike the theoretical test, the practice one did not changed and still depends completely on supervisor eyes judgment with total absence of tools that would facilitate their job. Today, the practice test should follow the same path using technology based on motion detection. In this sense, this new vision integrates motion sensors and video surveillance. Each one of them comes with advantages and requirements. Therefore, this research will focus on the concept of video surveillance system and propose methods and algorithms to be implemented in order to supervise and evaluate objectively in real-time the Car Reverse Test (CRT) in driving license exam.

Technically, this paper is going to proceed as the following: In Section 3, an overview of the method and system architecture is presented to show the whole procedure which will be divided in three subsections: subsection 3.1 is about Surveillance System Architecture that would represent the architecture of the system that is used for the proposed solution. The subsection 3.2 is dedicated to the method adapted on the solution and also a literature review to all the steps. The subsection 3.3 describes the application of the proposed method in Car Reverse Test (CRT). Section 4 represents the experimental results. Finally, Section 5 concludes the achievement of the paper.

II. Related Work

The implementation of video surveillance system to manage CRT requires using motion detection techniques to process footages that are captured and tries to analyze them so as to come with a judgment, the common idea in most video motion detection algorithms is to begin with a defined frame $\mathbf{f_0}$ called reference frame and a successive frame $\mathbf{f_i}$. This algorithm compares the two frames to detect any possible changes then moves to the next frame $\mathbf{f_{i+1}}$ so the reference frame $\mathbf{f_0}$ becomes $\mathbf{f_i}$. This process is successively repeated for all frames.

Many methods for motion detection exist in the literature. For instance, the authors of [3] discuss three methods to detect motion in given images: background subtraction, optical flow and temporal difference which are commonly used alone or side by side with other techniques in many researches on subjects related to motion detection.

The background subtraction was used by authors of [4] and [5] and it consists on comparing an image with a static reference image to distinguish the dynamic foreground of the static background of those images. The optical flow method as cited by the authors of [6] is focused on how much each pixel of the current frame moves between adjacent frames. The temporal difference method used by authors of [7] compares successive frames by analyzing all frame pixels in order to detect any moving regions.

To summarize, there are many methods for motion detection which could be used alone or side by side to achieve significant results.

III. System and Method Review

A. Surveillance System Architecture

Within the same realm, the idea here is to encompass the use of the video surveillance system in CRT. This system consists of video cameras, video processing unit (VPU), network, visualization centre and video database and retrieval tools. The first part of the system is the camera which has the role of capturing videos. Then, in the second part, the video will be transmitted to the VPU through the network which contains the software to process received footages using appropriate algorithms. The system is equipped with video storage and retrieval tools in order to help the VPU to Store and retrieve the related content of a video for processing or archiving. Then, the useful information will be transmitted to the next part, which is Visualization Centre whereby the supervisor (in our case) will get the score and validate the final result. In Figure 4 the architecture of the video-surveillance system is illustrated according to the author of [8].

B. Method Overview

The Method represented in the flow chart of Figure 5 is composed of three steps to be executed in the VPU which will be called Line Monitoring Algorithm (LMA).

The first step is using a noise removing algorithm in order to remove any noise caused by the image capture or the environment which gives clearer image as a result. The second step is to get line pixels using Bresenham's Algorithm which serves to get minimal calculated pixels between two pre-selected points in a straight line relaying those points. That algorithm is applied in the reference frame. In this stage, the coordinates of generated pixels will be stored in order to extract the same line location in next frames. The last step is to apply the Structural Similarity Index Method (SSI) in which changes in the generated Line between the reference frame and the current frame is calculated, and the result will be compared to the empirical threshold $\mathbf{T_h}$.

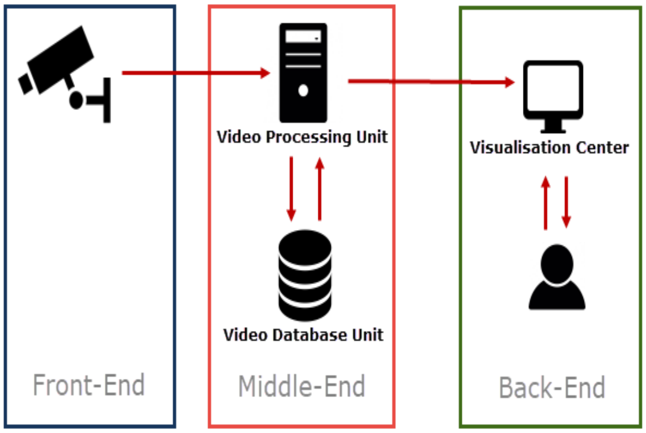


Fig. 4. Architecture of Video Surveillance System by [7].

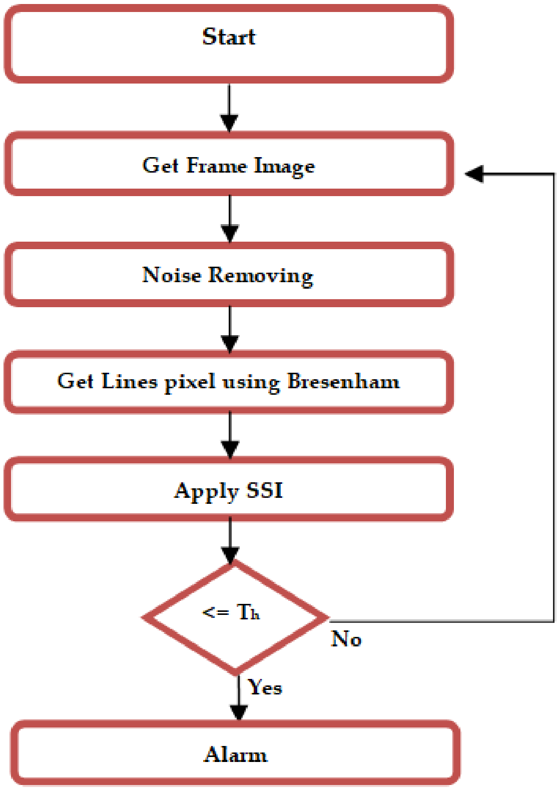


Fig. 5. Flow Chart of Line Monitoring Algorithm.

The experimental result in Section 4 presents a set of images to help in understanding the processes achieved in the present method.

1) Structural Similarity Index: In 2002, authors of [9] proposed a new universal image quality measurement method named Structural Similarity Index (SSI) was invented. This method is based on calculating three factors: the luminance, the contrast and the structure. SSI is easy to calculate and apply on various image processing applications.

Mathematically the SSI is represented also by authors of [10] as:

Suppose **S1** = {v$_i$|i=1,2,…,N} and **S2** = {v$_i$|i=1,2,…,N} are two non-negative image signals, which will represent, in this work, the signals corresponding to the lines extracted from reference frame and the current frame successively where **i** is the index of pixel in the line, and $\mathbf{v_i}$ is the intensity of that pixel.

According to [10] by the diagram shown in Figure 6, the comparison of luminance, contrast and structure measures is given as follows:

$$l(S1,S2) = \frac{2\mu_1\mu_2 + C_1}{\mu_1^2+\mu_2^2+C_1} \tag{1}$$

$$c(S1,S2) = \frac{2\sigma_1\sigma_2 + C_2}{\sigma_1^2+\sigma_2^2+C_2} \tag{2}$$

$$s(S1,S2) = \frac{\sigma_{1,2} + C_3}{\sigma_1\sigma_2 + C_3} \tag{3}$$

Where $\mathbf{C_1}$, $\mathbf{C_2}$ and $\mathbf{C_3}$ are very small constants, which are linked to the range of pixels values.

The combination of equations (1), (2) and (3) gives the general form of structural similarity between two signals **S1** and **S2** which is defined as:

$$\mathrm{SSIM}(S1,S2) = [\mathrm{l}(S1,S2)]^{\propto} . [\mathrm{c}(S1,S2)]^{\beta} . [\mathrm{s}(S1,S2)]^{\gamma} \quad (4)$$

Where $\boldsymbol{\alpha} > 0$, $\boldsymbol{\beta} > 0$ and $\boldsymbol{\gamma} > 0$ are parameters defining the importance of the three components. In order to simplify the expression (4), the authors of [10] set $\boldsymbol{\alpha} = \boldsymbol{\beta} = \boldsymbol{\gamma} = 1$ and $\mathbf{C_3} = \mathbf{C_2} / 2$ so the resulting SSIM form:

$$SSIM(S1,S2) = \frac{(2\mu_1\mu_2 + C_1)(2\sigma_{1,2} + C_2)}{(\mu_1^2+\mu_2^2 + C_1)(\sigma_1^2+\sigma_2^2 + C_2)} \quad (5)$$

Three conditions are been satisfied by (5):

- Symmetry: SSIM(S1,S2) = SSIM(S2,S1);
- Boundedness: SSIM(S1,S2) <= 1 ;
- Unique maximum: SSIM(S1,S2) = 1 if and only if S1=S2.

2) Bresenham's line algorithm: Bresenham's Line Algorithm (BLA) [11] is named after Jack Elton Bresenham who developed it in 1962 at IBM. This algorithm uses integrated/embedded arithmetic only, and is faster and more accurate than other obvious algorithms that use floating-point arithmetic.

The main idea of BLA is done for the purpose of connecting two given points in a square grid. The result of the BLA for two points **P1**(x1,y1) and **P2**(x2,y2) is represented in Figure 7.

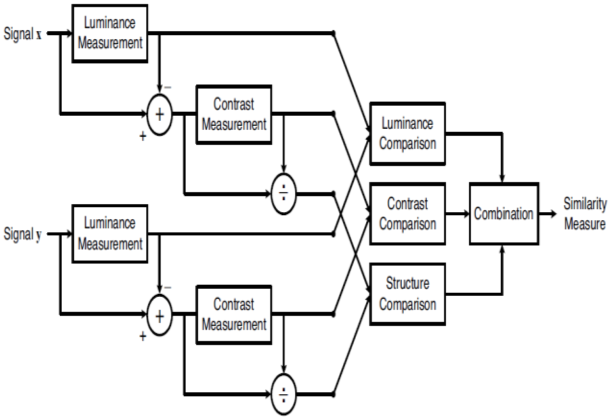


Fig. 6. Diagram of the SSI According to Authors of [10].

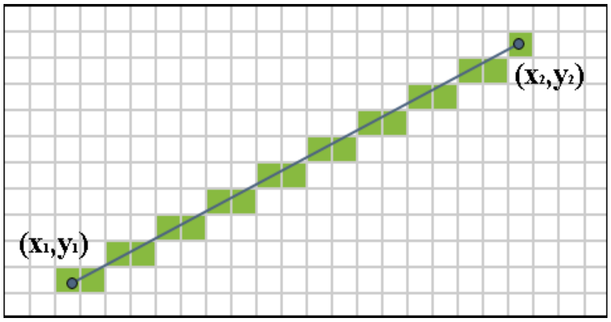


Fig. 7. Testing Bresenham's Algorithm Result Example.

C. Proposed LMA applied on CRT

The main objective is to automatically supervise CRT using the Artificial Intelligence AI. The decisions made by the system, without human intervention, are based on computer vision techniques using only installed cameras. This proposed technique will offer objectivity to the test.

According to the Moroccan law, there are four parameters that should be specified concerning CRT shown in Figure 3 and also explained in Figure 8.

Start Line (SL): crossing the start line indicates that the candidate is about to start the exam, the system monitors the tracking area and limits lines to track the vehicle.

End Line (EL): crossing the End Line means that the test is finished. The system stops monitoring the Tracking area and the score with recorded video are saved.

Limit Lines (LL): Two straight Lines indicating Limits that should not be crossed by the vehicle.

Tracking Area (TA): is created between Limits Lines, and it allows system to monitor the car through to the End Line.

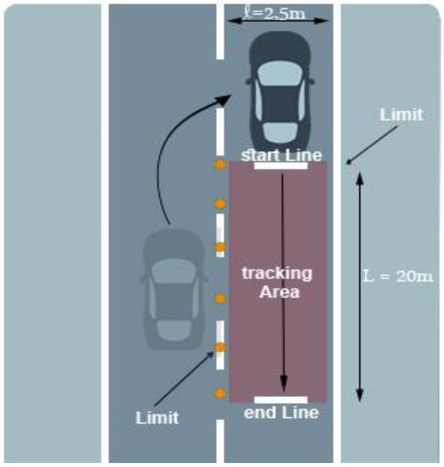


Fig. 8. Car Reverse Test Scene Review (Bird-Eye View).

The proposed algorithm to detect motion on specific line is labeled Line of Interest (LOI). The latter is represented as it is shown in Figure 10. The algorithm can be summarized in the following:

1- The algorithm starts by loading the video from video capturing device through the camera.
2- Display the video feed in Visualization Center to let the user initialize the system.
3- Initializing the system by choosing two points to draw the first line which indicates the Start Line.
Afterwards, the second two points bespeak End Line. And finally, the last four points refer to the Tracking Area.
4- Draw the selected areas to be pictured in visualization center
5- Capture the first frame $\mathbf{f_0}$
6- Extract pixels using Bresenham's Algorithm for reference lines: Start Line $\mathbf{SL_0}$, End Line $\mathbf{EL_0}$, Tracking Area $\mathbf{TA_0}$, and Limit Lines $\mathbf{L1_0}$ and $\mathbf{L2_0}$
7- Start Monitoring changes at Start Line
    a- Capture the first frame $\mathbf{f_t}$ at the time instant **t**
    b- Extract the line pixels of $\mathbf{SL_t}$ using Bresenham's Algorithm
    c- Convert $\mathbf{SL_t}$ to gray level format
    d- Compute the $\mathbf{SSI(SL_t, SL_0})$
    e- Go to "step a" until the similarity between the $\mathbf{SL_0}$ and $\mathbf{SL_t}$ is below the predetermined threshold $\mathbf{T_h}$ and this means a motion activity is detected.
8- Start monitoring changes in Tracking Area and Limit Lines and End Line.
9- Re-execute "step 5" for $\mathbf{EL_0}$ while no activity is detected.
10- Save recorded video and result.

To define the threshold value $\mathbf{T_h}$, values of coefficients in combination of SSI score (luminance, contrast and structure), scene environment and tolerated intensity of changes (ex: changes accrued on 50% of line) should be considered.

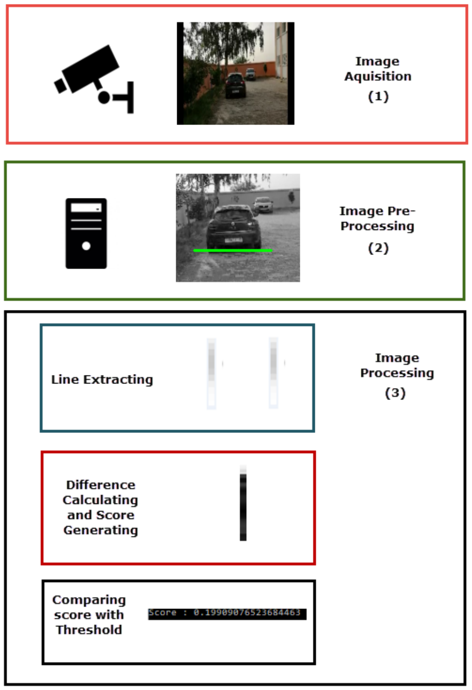


Fig. 9. CRT System Review.

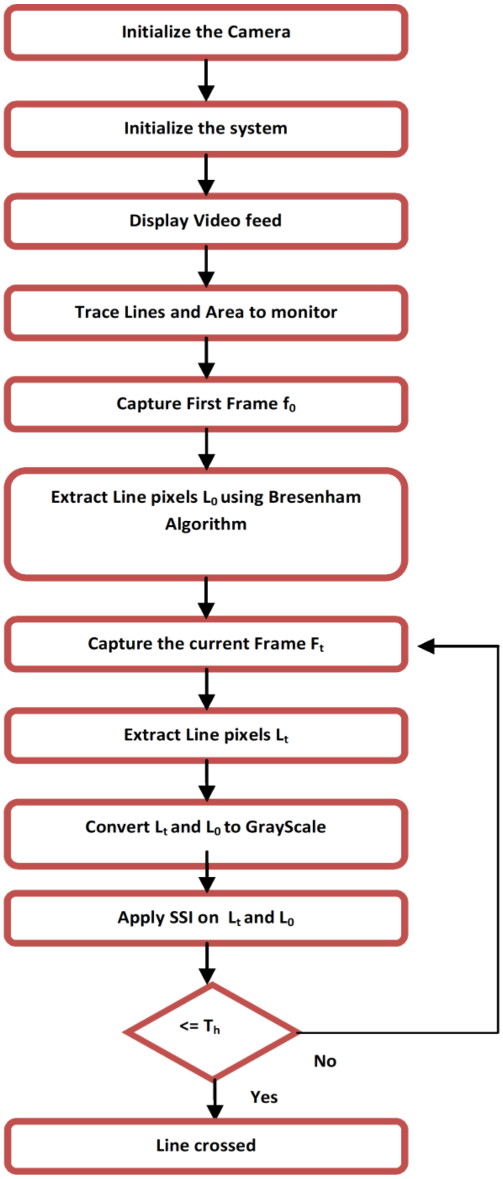

Initialize the Camera
Initialize the system
Display Video feed
Trace Lines and Area to monitor
Capture First Frame f_0
Extract Line pixels L_0 using Bresenham Algorithm
Capture the current Frame F_t
Extract Line pixels L_t
Convert L_t and L_0 to GrayScale
Apply SSI on L_t and L_0
<= T_h
No
Yes
Line crossed

Fig. 10. Flow Chart of CRT Supervisor System.

For each line from parameter lines specified in subsection 3.3 above, the first step in CRT system shown in Figure 9, and also in Figure 10 is extracting the Line Of Interest ($\mathbf{L_t}$) pixels from the frame $\mathbf{F_t}$, then, the next step is comparing the line resulted from previous step with the Line of reference ($\mathbf{L_0}$).

IV. Result and Discussion

A. *Comparaison between LMA and SSI based ROI*

SSI based Region Of Interest (ROI) was used in a various fields of research with similar approaches to this work as [14] and [15]. In this respect, the comparison with SSI based ROI is jugged suitable to appreciate the proposed technique LMA. The main idea in LMA is to monitor limit lines of the whole region instead of what SSI based ROI does.

For the experimental test, flow charts shown in Figure 11 and Figure 12 represent a qualitative and quantitative comparison of the LMA and SSI based ROI method. The Flow chart in Figure 11 represents the SSI based ROI test program. After the initialization, the program starts with extracting ROI from the current frame. Then the denoising technique is applied using Non-local Means Denoising Algorithm [13], the SSI score is calculated to detect change accrued in the ROI. In this case, the score and elapsed execution time are reported. The flow chart shown in Figure 12 describes the same processing, in which, LMA uses the BLA to extract LOI instead of ROI in the previous program. Same outputs (score and execution time) are reported for the comparison.

All tests are run under Virtual Machine CPU Core Duo 1GHz and 1 GB of memory with Linux Ubuntu 16.4 Debian Distribution 64 bits. As for the programming language, Python is used with OpenCV library for graphic manipulation.

The test program uses as input a recorded video [12] of 1 min captured by a CCTV camera of a parking space. In order to manage the extracting and comparing images, the Multithreading technique is implemented to provide simultaneous execution in order to enhance the performance of test programs and to save more processing time.

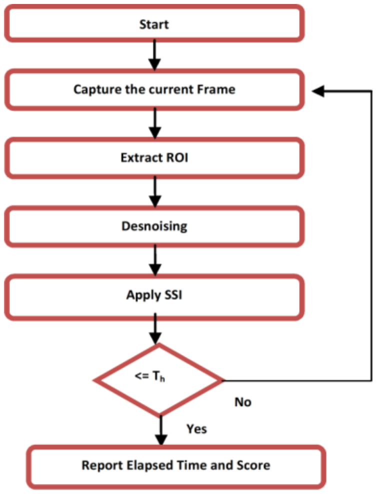


Fig. 11. Chart for SSI based ROI Test Program.

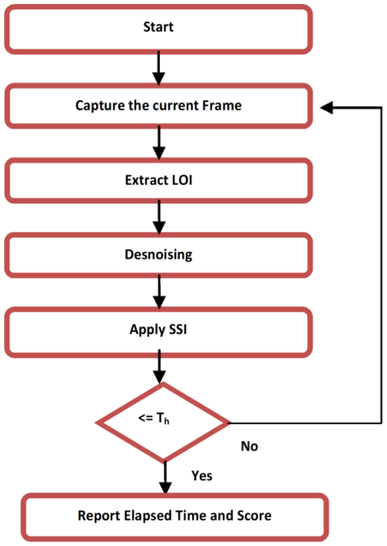
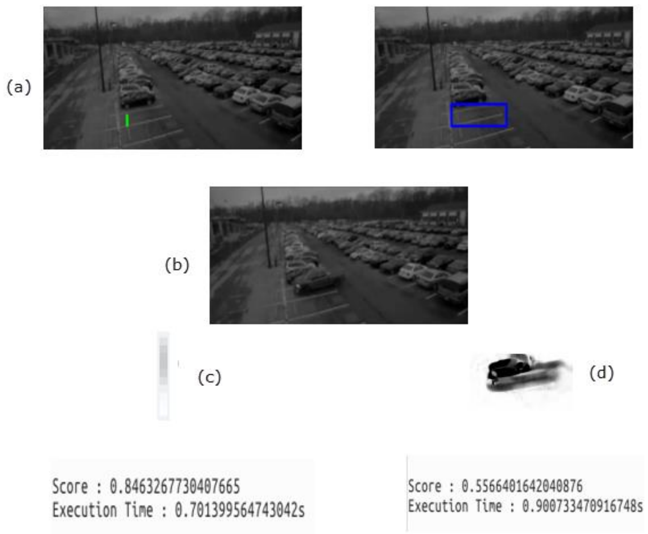


Fig. 12. Flow Chart for LMA Test Program.



```
Score : 0.8463267730407665
Execution Time : 0.701399564743042s
```

```
Score : 0.5566401642040876
Execution Time : 0.900733470916748s
```

Fig. 13. Exp.Result of using SSI based ROI and LMA.

a) Defining Areas

b) Car parking

c) LMA SSI Difference Map (zoomed)

d) ROI SSI Difference Map

The experimental result shown in Figure 13 shows that LMA and SSI based ROI successfully detects the movement inside the chosen area. However, in this comparison, LMA takes less time to detect changes, and this is justified by the small amount of data processed. Additionally, the SSI score is near 1 before changes accrued, which means that proposed algorithm is more insensitive to noise than SSI based ROI.

The advantage of using LMA is shown also in multiline monitoring specially in real-time implementation. The result shown in Table I denotes the FPS variation by the number of drawn lines. As a result, for each 10 processed lines, the video speed is dilated by only ~ 3 FPS which is less than SSI based ROI.

In this section, the comparison between LMA and SSI based ROI shows that the combination of the two algorithms would give a remarkable solution especially where LMA is been used as a trigger of SSI based ROI which can not affect the video speed.

TABLE. I. FRAME PER SECOND VARIATION USING LMA

Number of Lines	**FPS**
1	~ 24.7
10	~ 21.5
20	~ 18.5
30	~ 16.1

B. Performance of CRT Supervisor System

To determine objectively the accurate threshold value $\mathbf{T_h}$ and also to evaluate the performance of the proposed CRT Supervisor System, a primary test has been made using recorded videos. Therefore, in this work, the performance measures discussed by the authors of [16] will be adopted. These measures are introduced in this work as the following:

-True Line Crossed (TLC): It is the number of every couple of successive frames where the vehicle crosses the line meanwhile the CRT system declares the change.

-False Line Crossed (FLC): It is the number of every couple of successive frames where the vehicle crosses the line but no change is declared by CRT system.

-True Line Not Crossed (TLN): It is the number of every couple of successive frames where the line is not crossed while the CRT system declares the change.

-False Line Not Crossed (FLN): It is the number of every couple of successive frames where the line is not been crossed but no change is declared by CRT system.

The extracted values are classified by the confusion matrix shown in Table II.

Using these measures, the following Performance indicators are deduced for the CRT system:

$$\text{precision} = \frac{\text{TLC}}{\text{TLC} + \text{FLN}}$$

$$sensitivity = \frac{\text{TLC}}{\text{TLC} + \text{FLC}}$$

$$specificity = \frac{\text{TLN}}{\text{FLN} + \text{TLN}}$$

$$accuracy = \frac{\text{TLC} + \text{TLN}}{\text{TLC} + \text{FLN} + \text{FLC} + \text{TLNs}}$$

To check the quality of the proposed CRT system, different Threshold values are chosen to measure their effect in program results and correctness of the algorithm.

For the experimental purpose, a recorded video of the real scene was used with pre-determined motions. This choice is made in order to simplify the visual motion detection so that the detected motions would be easily compared with those of the proposed algorithm. The main objective is to run this test using a different threshold values to choose finally the most accurate value. The analyzed video specifications are:

- Length of video: 26s
- Number of Frames: 666 (57 where line is crossed “LC”)
- Frame rate: 25 f/s

The application of LMA on this video under different threshold values generates results that are tabulated in Table III and also represented in the graph in Figure 14. The idea is to choose the **T_h** value which gives a minimal false detection (FLC) and nearer value of true detection TLC to the predetermined one. In Table III, the **T_h** value which fills those conditions is **T_h=0.3**.The results of calculating the performance indicators displayed in Table IV and in Figure 15 show that **T_h=0.3** gives the best accuracy percentage, and this means that the program detects perfectly most the frames where the car is crossing the line.

TABLE. II. A CONFUSION MATRIX FOR SYSTEM RESULT CLASSIFICATION

System result/Test	Line crossed	Line not crossed
Change declared	TLC	FLC
Change not declared	FLN	TLN

TABLE. III. RESULT OF APPLYING LMA ON TEST VIDEO

T_h	0.95	0.85	0.75	0.45	0.3
TLN	351	401	422	472	586
TLC	145	95	74	64	57
FLN	0	0	0	0	0
FLC	170	170	170	130	23

TABLE. IV. PERFORMANCE RESULT OF APPLYING LMA ON TEST VIDEO

T_h	0.95	0.85	0.75	0.45	0.3
Precision	1	1	1	1	1
Sensitivity	0.46	0.36	0.3	0.33	0.71
Specificity	1	1	1	1	1
Accuracy	0.74	0.74	0.74	0.8	0.97

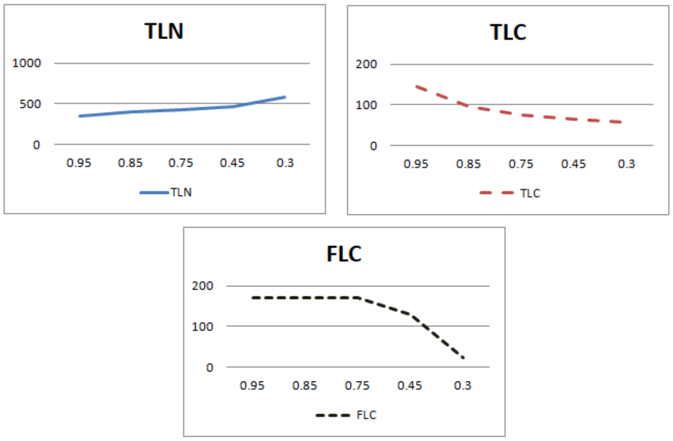


Fig. 14. Performance Measures Evolution by Threshold.

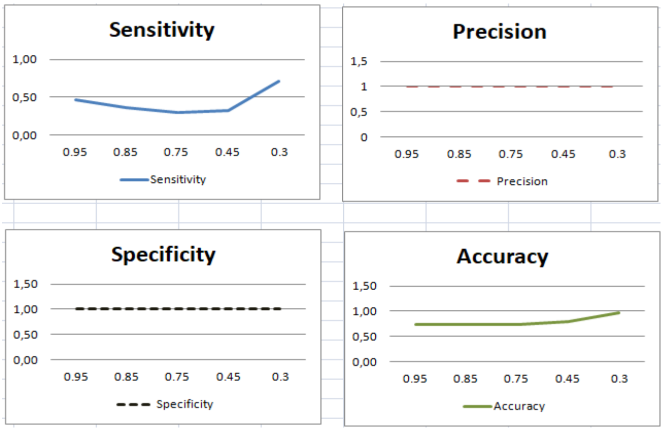


Fig. 15. Performance Indicators Evolution by Threshold.

As regards to the experimental results from above, it is clear that choosing the right value of threshold can be affected by the quality of the video. In our case, the change in selected line should be relatively high with the highest value of accuracy and low count of false detected frames FLC. Therefore, we conclude that the most accurate threshold value should not exceed **0.3** in order to achieve best results.

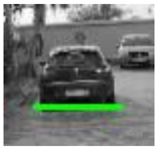
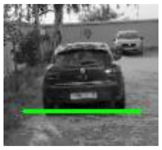
C. *The Result of the Proposed Solution*

In this section, we will use the same video as the subsection 4.2 above with a threshold value Th=0.3. The results displayed in Table V show that when using the chosen Th value, CRT System detects successfully the changes in start line and end line with an accuracy of 97%. The histogram represented in Figure 16 shows the variation of SSI score by the video frames where TLN, TLC and FLC values are highlighted. As regards the car crossing event, the graph above can be divided into three parts:

In the first part (1), the SSI score is greater than 0.3 so the system did not declare the change and the car is not arrived yet to the line (TLN). The highlighted part (2) represents the entering/leaving events of the car where the system declares the change (TLC). In the next part (3), the SSI score is under 0.3 in several frames which leads the system to declare false changes (FLC) and that affects the accuracy (the other missing 3%). It is due to the position of the camera, which shows the line been drawn upon the vehicle. Thus, to manage this problem, the CRT system should not monitor the line when a change has already detected for the first time.

In the Figure 17, Figure 18 and Figure 19, the proposed form represents the real scene view of CRT system through the use of captured images from the real scene where all parameters are drawn. In this stage, we tried to manage the monitoring of Start Line and End Line. In further researches, the focus will be on other techniques to process the Tracking Area and Limit Lines; these techniques would provide useful data such as calculating the distance between the vehicle and the limit lines, measuring the vehicle speed, which can be used to make decisions and helping the learning process of the program.

TABLE. V. EXPERIMENTAL RESULT OF IMPLEMENTATION OF LMA IN CRT

Scene	Accuracy	Snapshot
Start line	97 %	
End line	97%	

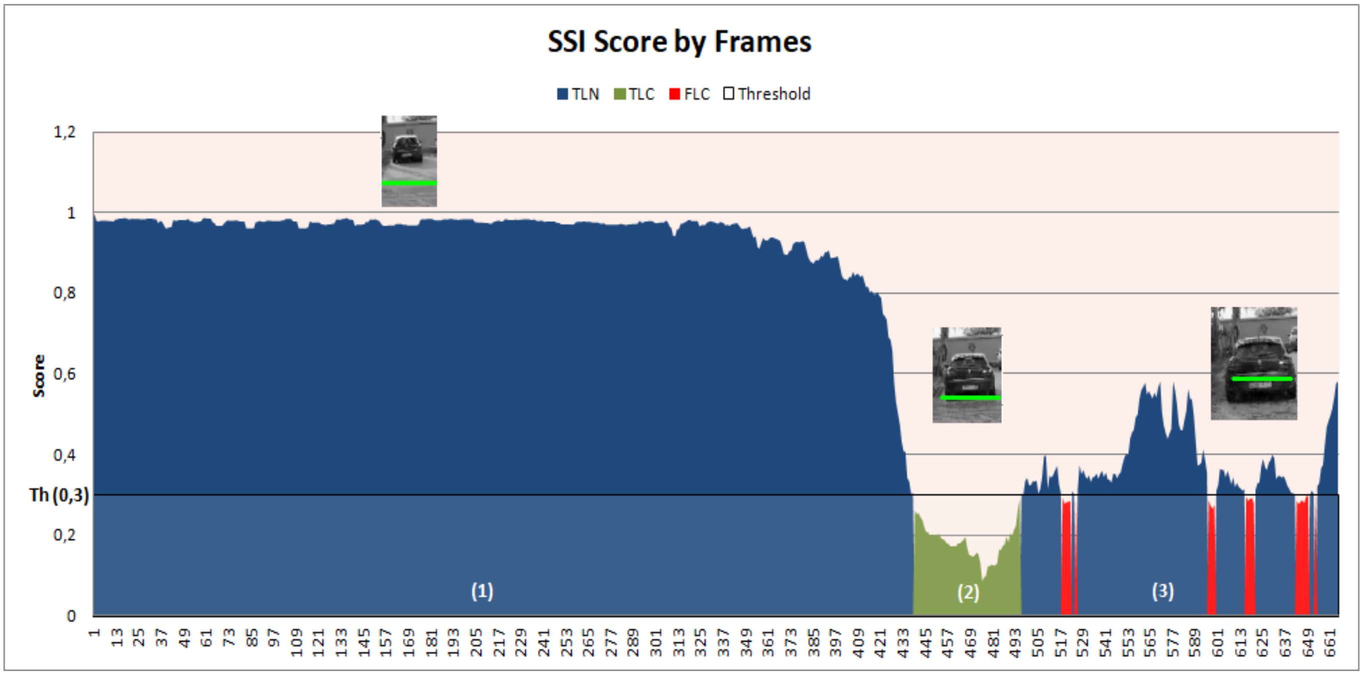


Fig. 16. Variation of SSI Score by Frame.

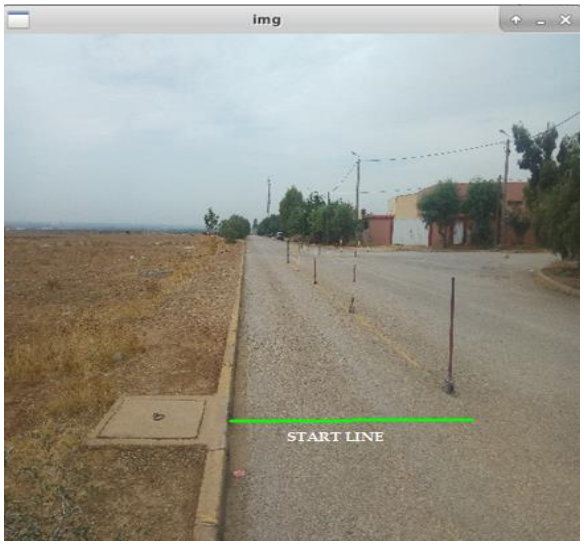


Fig. 17. Drawing Start Line

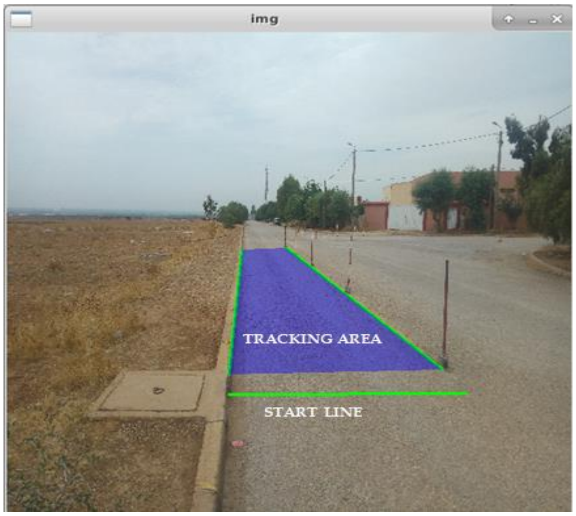


Fig. 18. Drawing Limit Lines and Tracking Area.

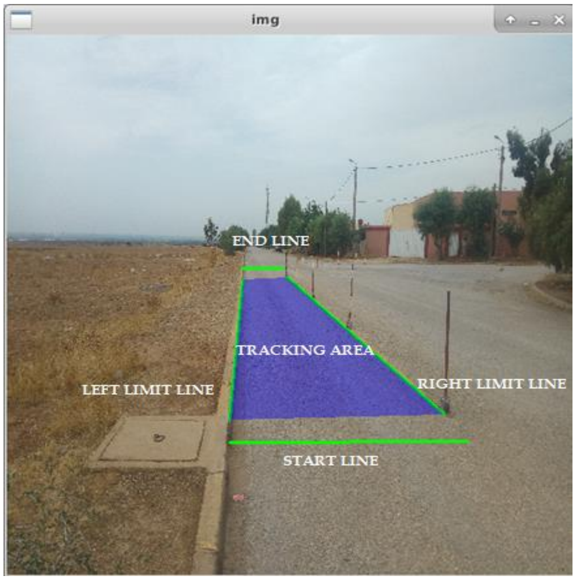


Fig. 19. Drawing End Line.

V. Conclusion and Perspectives

In this paper, in order to design and develop a CRT Supervisor System, which is one of several sub-systems composing a Driving License Supervising System that would give more objectivity to practical tests in driving license in Morocco. Therefore, the proposed Line Monitoring Algorithm based on Structural Similarity Index is investigated. The results exhibit that the proposed algorithm presents a high capacity in motion detection which would be a prominent resource of research.

The proposed LMA is based on extracting lines from successive frames and measures the similarity between them using Structural Similarity Index (SSI) and Bresenham's Algorithm. The proposed algorithm comes with the highest motion detection accuracy in most experiments, especially with the right threshold value and adapted camera view. The experimental results show that the proposed SSI based method has better performance comparing to monitoring the whole region or frame. This method also has shown to provide a high efficiency with a high speed and relatively a low storage, which makes it a great choice when it comes to dealing with limited resources. The main advantages of proposed method are higher detection accuracy and improved processing speed.

The implementation of the proposed method gives an exciting result for a start. Therefore, further researches will focus on providing more techniques to be integrated on the proposed system for the driving license test in order to serve the main cause which is improving the road safety in Morocco.

References

[1] W. H. Organizations, “WHO,” [Online]. Available: https://www.who.int/. [Accessed 10 1 2018].

[2] Decision of the Minister of Equipment and Transport No. 2709.10 issued on 20 Shawal 1431 (September 29, 2010) specifying the conditions under which the demand, Completion and delivery of driving licenses, Moroccan Official Bulletin [Online].Available: http://www.sgg.gov.ma/Portals/1/lois/A_2709.10_Ar.pdf?ver=2012-01-27-142440-000, p.p 15-24 [Accessed 9 10 2018].

[3] Mishra, Sumita, Prabhat, Chaudhary, Naresh K., Asthana, Pallavi, 2011. A novel comprehensive method for real time video motion detection surveillance. Int. J. Sci. Eng. Res. 2.

[4] Tang, Zhen, Miao, Zhenjiang, 2008. Fast Background Subtraction Using Improved GMMand Graph Cut. In: Congress on Image and Signal Processing, CISP ’08, p.p 181–185.

[5] Li, Hongyan, Cao, Hongyan, 2010. Detection and segmentation of moving objects based on support vector machine. In: 2010 Third International Symposium on Information Processing, Shandong China, p.p 193–197.

[6] Jung, Ho Gi, KyuSuhr, Jae, Bae, Kwanghyuk, Kim, Jaihie, 2007. Free parking space detection using optical flow-based euclidean 3D reconstruction. In: Proceedings of the IAPR Conference on Machine Vision Applications Tokyo Japan, p.p 16–18.

[7] Yu, Zhen, Chen, Yanping, 2009. A real-time motion detection algorithm for traffic monitoring systems based on consecutive temporal difference. In: Proceedings of the 7th Asian Control Conference, Hong Kong, China, August 27–29.

[8] Chan, S.C., Zhang, S., Wu, J., Tan, H, Ni, J.Q., Hung, Y.S. (2013) ‘On the Hardware/Software Design and Implementation of a High Definition Multiview Video Surveillance System’. IEEE Journal on Emerging and Selected Topics in Circuits and Systems. Vol. 3(2) p.p 248-262.

[9] Wang, Z., Bovik, A.C., 2002. Why is image quality assessment so difficult? Proc. IEEE Int. Conf., Acoustics, Speech and Signal Processing 4, p.p 3313–3316.

[10] Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004) Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing, 13, p.p 600-612.

[11] I. Joy Kenneth Breshenham's algorithm. In Visualization and Graphics Research Group, Department of Computer Science University of California, Davis (1999).

[12] Youtube video intituled: DOTS Parking Cameras Live Stream by DOTS Parking Cameras diffused live on Feb. 10th 2016 [Online].Available: https://www.youtube.com/watch?v=FuvhI9Vmek0/. [Accessed 15 10 2018].

[13] A. Buades, B. Coll, J. Morel, "Denoising image sequences does not require motion estimation", Proc. IEEE Conf. Adv. Video Signal Based Surveil. AVSS Palma de Mallorca Spain, pp. 70-74, September 2005.

[14] Z. Zhang, T. Jing, J. Han, Y. Xu, and X. Li, “Flow-Process Foreground Region of Interest Detection Method for Video Codecs,” IEEE Access, vol. 5, pp. 16263–16276, 2017.

[15] P. Banerjee, S. Sengupta, "Human motion detection and tracking for video surveillance", National Conference for Communication, 2008.

[16] M. Solokova, N. Japkowicz, S. Japkowicz, Beyond accuracy, F-score and ROC : a family of discriminant measures for performance evaluation , Australian Conference on Artificial Intelligence, vol. 4304, pp. 1015-1021, 2006.

Customers Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry

Yasser Khan[1], Shahryar Shafiq[2]
Abid Naeem[3], Sabir Hussain[6]
Department of Electrical Engineering, Iqra National University, Peshawar Pakistan

Sheeraz Ahmed[4]
Department of Computer Science
Iqra National University
Peshawar Pakistan

Nadeem Safwan[5]
Department of Management Science
Iqra National University
Peshawar, Pakistan

***Abstracts*—To survive in the fierce competition of telecommunication industry and to retain the existing loyal customers, prediction of potential churn customer has become a crucial task for practitioners and academicians through predictive modeling techniques. The identification of loyal customers can be done through efficient predictive models. By allocation of dedicated resources to the retention of these customers would control the flow of dissatisfied consumers thinking to leave the company. This paper proposes artificial neural network approach for prediction of customers intending to switch over to other operators. This model works on multiple attributes like demographic data, billing information and usage patterns from telecom companies data set. In contrast with other prediction techniques, the results from Artificial Neural Networks (ANN) based approach can predict the telecom churn with accuracy of 79% in Pakistan. The results from artificial neural network are clearly indicating the churn factors, hence necessary steps can be taken to eliminate the reasons of churn.**

Keywords—Neural Network; ANN; prediction; churn management

I. Introduction

To survive in dynamic, high service demanding sector of telecommunication and to achieve operational excellence, it is indispensable to maintain up to date customer relationship management system (CRM). This system plays pivotal role in developing customer satisfaction, loyalty and main interface to interact with our clients [1]. The CRM system could revolutionize the entire system by providing real time provisioning of information, improvement in sale process and, enhancement of customer loyalty, advertisements and increase the effectiveness of marketing tools [2]. This term is basically emerge from relationship marketing defined as – the type of marketing where a lot of emphasis is laid to improve the qualitative, strategic and supportive relationship with already working customers and devise strategy to attract new customers.[3]. In order to survive in extreme competitive market of telecom, the firms must have to adapt with external changes, so CRM is proved to be an important monitoring tools to detect environmental changes on business operation and take tough business decision. CRM system is providing both tangible and non-tangible benefits hence former are providing increasing productivity, increase retention rates and profitability, decrease the marketing cost and faster turnaround time while later benefit is responsible for customer loyalty and satisfaction, take the benefits of segmentation, increase customer experience, positive effect from word of mouth effects and excellence customer experience [4].

The income can be increased by improving service quality and customer satisfaction, redress the customer problems and timely tackle of complaints registered by customer [5]. The process can be more effective by introduction of automation through CRM system. The business capabilities can be improved by better and integrated knowledge in real time from CRM system and help significantly through better decision-making. By extracting detailed tables and report without spending much time which can better improve working productivity [6]. Excellent and comprehensive analysis can be executed through provisioning of customer service so individual planning and decision making can improved through better learning opportunity. The customer needs, expectations and behavior can be predicted through information derive from CRM [7].

The prime factors behind customer retention can be extracted in order to develop the profitable, loyal and long lasting relationship with their clients. Without running retention compaign, the telecom operators are consistantly losing significant numbers of customers that is 20% to 40% each years [8]. By bringing improvement in retention phenomena the aount £128 million rang from 20% to 10 % can anually be saved by one of british well know telecom operator Orange [9]. Many statistical and datamining techniques are introduced to investigate the customer churn prediction. In contrast to market survey, data mining techniques are analyzing the information obtained from both historical & current data in order to predict the patterns on historical data and future customer attitudes [10].

The most common techniques used for prediction are Decision Tree (DT), Logitic Regression (LR), Support Vector Machine (SVM) and Neural Network (NN). Further more, the decision tree is used for resolution of classificaiton problems to divide the instances into two or more than two classes. Similarly, logistics regression gives the probability by providing input/oupt fields, set of equation and factors causing customer churn [11].

Most of the companies in today's world are suffered badly due to switching over of dissatisfied customers famously know as customer churn and departure is mostly done to new competitor. The acquisition of new customer costs new company 6 to 7 times more than retaining the existing customer hence cause lot of profit lose [12]. The most probable reason behind the departure of customer is achieving of cheaper offer from another company, expression of dissatisfaction from existing operator or successful marketing strategy of new company [13].

Predictive Analytics can divided into classifications and prediction which can further be sub divided into decision tree, logistic regression, neural network and support vector machine given in Fig. 1. To protect the profit, brand name and assets of exiting operator, it is deemed necessary to retain the customers. Hence collection of these customer data and necessary prediction would help identification of these dissatisfied customer and would help utilization of resources for these targeted customers.

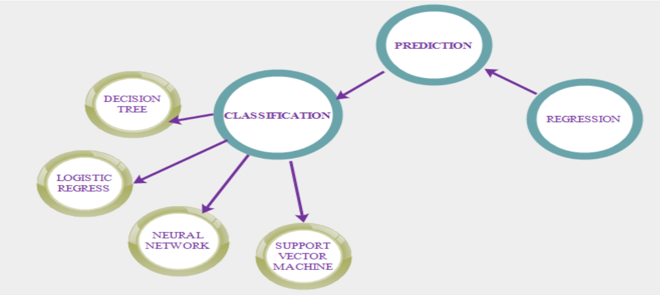


Fig. 1. Churn Prediction Techniques.

In the section-I brief introduction about customer churn is given which is followed by literature review in section-II and related work about churn and predictive analytics techniques is furnished. In the section-III deep learning is explained and illustrated followed by methodology in section-IV. In section-V, all the results of study are given in details.

II. Literature Review

The organization of recent times have significantly moved from product and service centric approach to customer centric behavior. To meet the needs of the customer is merely remain the core function of the organization but telecom operator are fast moving towards service improvement by enhancing customer loyalty and satisfaction. A lot of emphasis is laid to developed strategy to retain and categorize the customers according to the values they are providing to company [14].

Acquiring new customer is many times costlier than to retain the existing customer so strategic goal is set to work on dissatisfied customers who are intended to stop the service and leave the company and phenomena is termed as customer churn management. This involves prediction of those profitable customers in advance, are unable to continue with existing firm due to marketing strategy failure, company tariffs, poor customer care or frequent technical issues [15].

In this scenario data mining techniques are used to process and analyze the demographic details, customer care service data, billing and account receivables, customer credit scoring, customer usage behavior patterns , value added services data, customer relationship data. Hence model is developed and evaluated for potential churners and arrange to develop effective business solution to beat the competitor in the face of extreme competition. This activate the CRM department to convince churner customer by redressing their issues, offering extra service and discount in monthly bills so loss could be prevented. The six steps initiating from data collection to devising of churn policy [16] is depicted in Fig. 2.

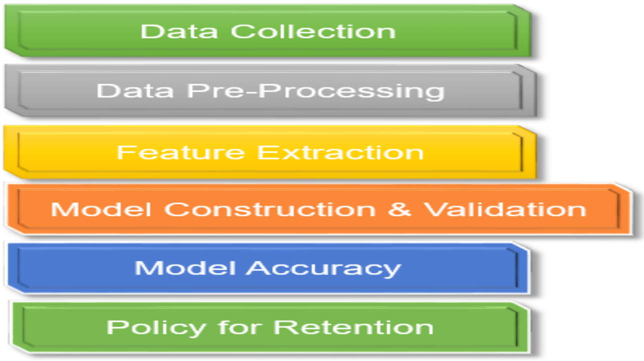


Fig. 2. The Six Steps for Customer Churn Prediction.

Churn prediction and analysis are performed through different techniques and covered mostly by data mining tools. Many different studies are conducted by researchers and telecom professional to construct churn prediction models with varying accuracy and precision on different data sets. Support Vector Machine (SVM), Logistic Regression (LR) and Multilayer Perceptron are being used for model creation from data set of Taiwan Logistic Company [17]. Similarly, decision tree algorithm (DT-ID3) is used for construction of prediction model for Malaysian Telecommunication Company [18]. Another study conducted on the basis of Decision Tree (DT) and logistic regression (LR) by taking the data from renowned Telecommunication Company by using survival analysis [19]. Research is conducted by using the data from Oracle Company's data set and techniques are Naïve Bayesian Theorem for building of churn prediction model [20]. In China Telecommunication Company, data was analyzed for churn prediction by using Decision Tree (DT), C5.0 algorithm BPN and Neural network [21].

III. Deep Learning Model

This is the phenomena involve the processing from features selection to ultimate process of classification or prediction. There are three main requirement for effective deep learning mode. These are availability of abundant data to train the model, adequate system computational power lead to introduction of GPU's and innovative algorithm to make the process fast [22].

A. *Artificial Neural Network*

In human brain, all decision are taken through neural networks provided naturally in our body which are composed of basic building block 'neuron'.

The biological neuron as given in Fig. 3 is composed of dendrites responsible for receiving of data from other neuron, cell body sums all the inputs received inside and output the data through axon outside from neuron [23]. All the communications and processing is performed in electrical signals through synapses–a connection point between dendrites and axon from preceding neuron.

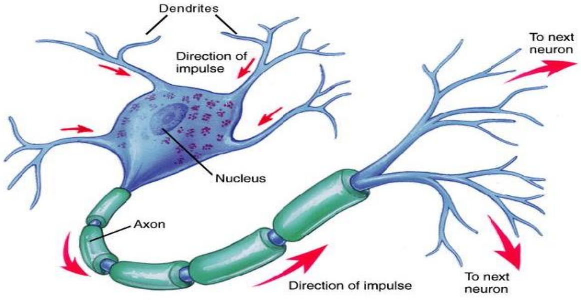


Fig. 3. The Biological Neuron.

Similarly, in artificial neuron inputs X1, X2… Xn are taken by each neuron and inserted to for summation and activation/ transfer function for decision making. The output is taken outside on basis of joint decision taken by entire neural network given in Fig. 4.

The neuron is composed of three main layers (input, hidden & output layers) provided in Fig. 5.When input Xi is applied to neuron, the weight is added to and results:

$$O_{k=} F\ (\sum W_i X_i + b_j) \quad (1)$$

W_i is weight for each input data X_i with b_j is the bias for each perceptron and O_k is the output.

The model is comprised of networks of neuron interconnecting with each other through weights and output from a particular perceptron is obtained from experience after completion of training process of a model [24].

Multilayer perceptron are perceptron consists of a scores of layer including a number of hidden layer in between input and output layer. In the training process, the error is consistently reduced calculated during comparison of actual and desired output and back feed to input side and weight for each input value is regularly adjusted.

$$E_i = A – \hat{A} \quad (2)$$

where E_i is the Error between actual and predicted outputs

A = Actual Output

Â= Predicted/ Modeled Output

B. Activation Function

In artificial neural network all the inputs are multiplied against their weights, sum and activation function is applied. There are lots of expectation from ANN to perform non-linear, complicated, high dimensional mapping between inputs and response output being considered as universal function approximates.

Sigmoid activation function or S-shaped curve range from 0 to 1. The function has some shortcomings of gradient problem vanishing and output is not zero centered hence push the gradient in far too different directions that’s why having slow convergence.

On another hand, Hyperbolic Tangent function (tanh) is zero centered and best for optimization which make it superior over sigmoid function. The third and most popular function is Rectified Linear Units (ReLU) which has improved the performance six times Tanh and completely remove the problem of vanishing gradient. The function is frequently used and limited to only hidden layer of neural network.

$$R(x)= \max(0,x) \quad (3)$$

If x<0, R(x) = 0 And x<0, R(x) = x

In addition, the modified version of ReLU is Leaky ReLU which has covered the shortcoming of ReLU and gradient stopped working during training phase and can die.

1) Sigmoid function :

$$f(\mathrm{z}) = \begin{cases} 0 & \text{for z} < 0 \\ \mathrm{z} & \text{for z} \geq 0 \end{cases}$$

2) Hyperbolic tangent function:

$$\tan\mathrm{h}(z) = \frac{2}{1+e^{-2x}} - 1$$

3) ReLU function:

$$f(z) = \frac{1}{1+e^{-z}}$$

C. Convolution Neural Network

The use of convolution neural network has largely being extended for understanding of image contents, image recognition, detection and segmentation.

The main reason behind this fact is capability of CoVNet depicted in Fig. 6, for scaling the network to millions of attributes and labeled data set help greatly in learning process. The technique has also caught the attention of classification of large scale videos classification where the network has excellent performance not only in static images but also complicated temporal evolution.

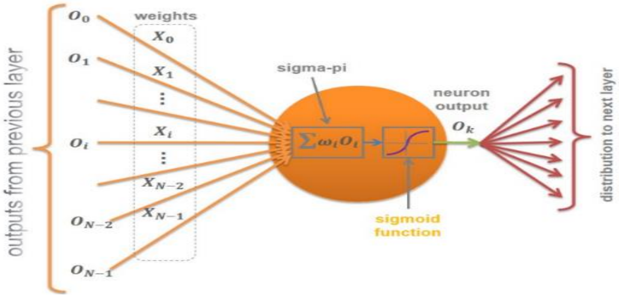


Fig. 4. The Artificial Neuron.

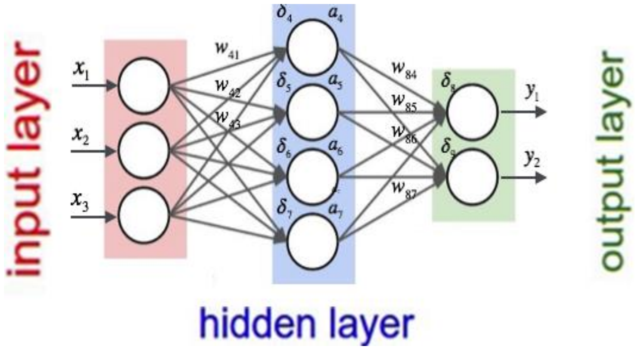


Fig. 5. The Layers of ANN.

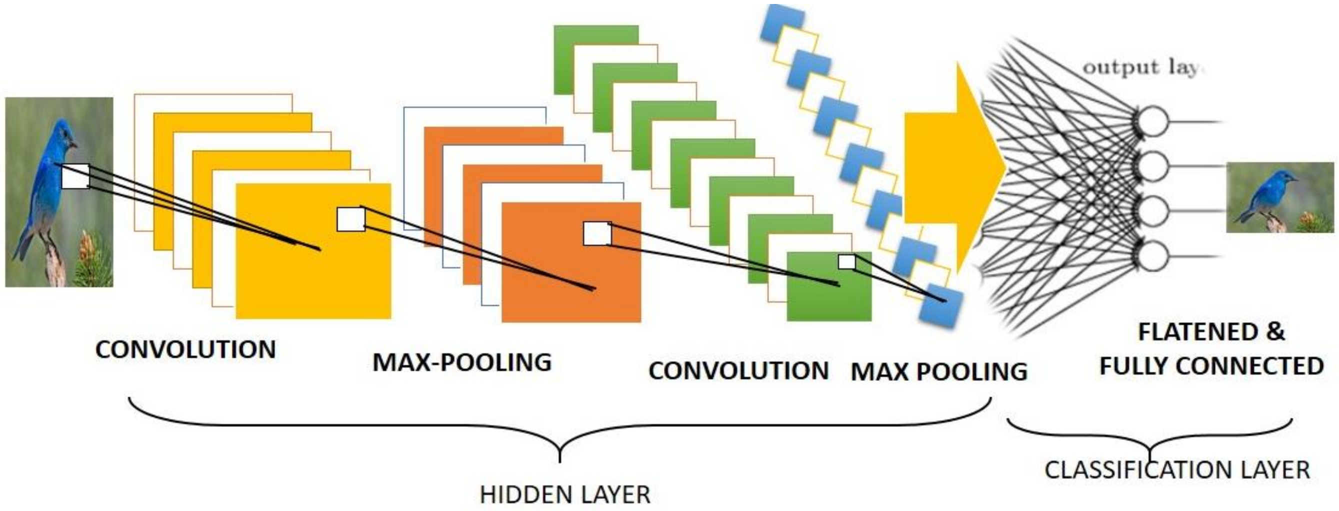


Fig. 6. Convolution Neural Network Layers Diagram.

IV. Methodology

A. Collection of Data

The collection of data were carried out from all cellular companies of Pakistan telecom market in Year 2018-19. There are 20468 instances comprise of 26 attributes selected during period of October to December 2018.The demographics information of the survey is illustrated in Fig. 7 indicating gender, education, marital and age information. However, Table I is detailed with descriptive statistics while the correlation of the data is furnished with another illustration, Table II.

B. Variable used for Design of Artificial Neural Network

1) Predictor (Independent) variables: The independent variable considered for this research studies is as below:

(i)Age (ii) Gender (iii)Monthly income (iv) Call Drop (v)call failure Rate (vi) Calling number (vii) customer ID (viii) customer suspension status (ix) Education (x) Home Owner (xi) Marital Status (xii) Monthly Billed Amount (xiii) Number of Complaints (xiv) Number of month Unpaid (xv) Number of days contract equipment's (xvi) Occupation (xvii) Area (xviii) Total minutes used in last minutes (xix) Unpaid Balance (xx) Either use internet service (xxi) Use voice service (xxii) Percentage of call outside network (xxiii) Total Call duration (xxiv) Avg. call duration (xxv) Churn (xxvi) Years& Months

2) Target variable: The dependent variable in this research is churn of customer and will exhibit dichotomous outcome with two variables i.e. 0 or 1. The value 1 indicate that customer has switched from the company and moved to another operator while value 0 shows customer is still stick to existing firm.

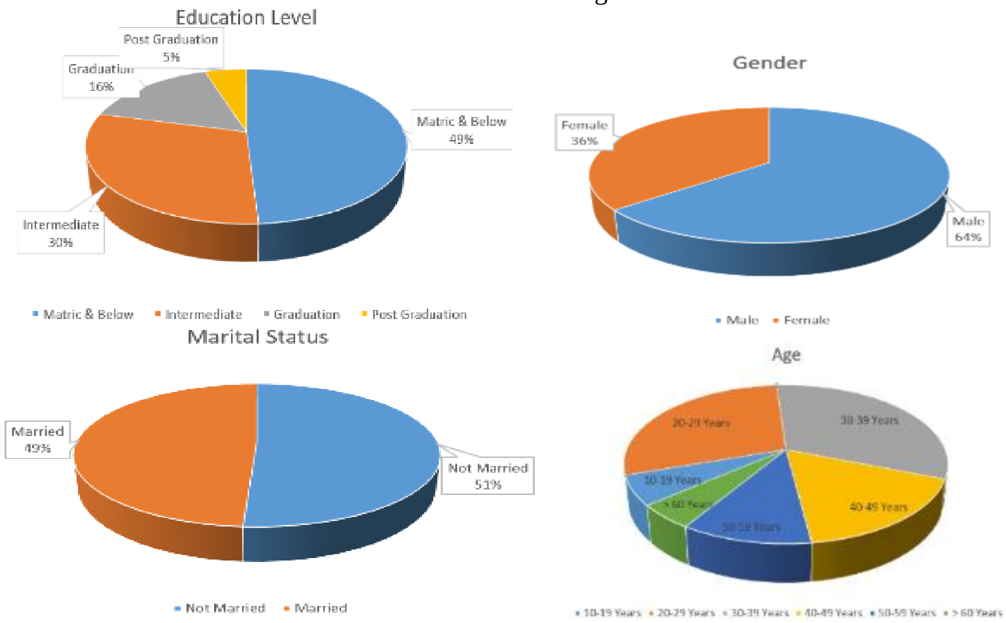


Fig. 7. The Pie Chart (Gender, Education Level, Marital Status and Age).

TABLE. I. DESCRIPTIVE STATISTICS

Descriptive Statistics									
	N	*Mean*		*Std. Deviation*		*Skewness*		*Kurtosis*	
	Statistics	*Statistics*	*Std. Error*	*Statistics*	*Std. Error*	*Statistics*	*Std. Error*	*Statistics*	*Std. Error*
Age	20468	45.33	.137	19.625	0.008	.008	.017	-1.199	.034
Education	20468	1.75	.006	0.887	.933	.933	.017	-.101	.034
Gender	20468	0.49	.003	0.500	.47	.047	.017	-1.998	.034
Marital Status	20468	0.49		0.500	.041	.041	.017	-1.998	.034
Valid N(listwise)	20468								

TABLE. II. THE CORRELATION AMONG AGE, GENDER, TENURE AND MONTHLY CHARGES

Correlations					
		age	gender	tenure	MonthlyCharges
age	Pearson Correlation	1	.661**	.623**	.791**
	Sig.(2-tailed)		.000	.000	.000
	N	20468	20468	20468	20468
gender	Pearson Correlation	.714**	1	.685**	.822**
	Sig.(2-tailed)	.000		.000	.000
	N	20468	20468	20468	20468
tenure	Pearson Correlation	.565**	.598**	1	.656**
	Sig.(2-tailed)	.000	.000		.000
	N	20468	20468	20468	20468
MonthlyCharges	Pearson Correlation	.786**	.822**	.756**	1
	Sig.(2-tailed)	.000	.000	.000	
	N	20468	20468	20468	20468
** Correlation is significant at the 0.01 level (2-tailed).					

C. Design and Setup of Artificial Neural Network

The artificial neural network model was built in multilayer perceptron (MLP) in IBM SPSS statistics 21. The ANN is trained through back-propagation learning algorithm and synaptic weights were adjusted through gradient decent for lowering the error through transformation function. The required data were assigned to training and testing in the percentage of 80% and 20% respectively. Special care were made not to over-train the model in the training mode and reduce the error to minimum possible level. Before initialization of training, all the covariates were normalized with values of 0 and 1.

The transformation function used for this model is hyperbolic tangent function returned the values all the ways between -1 and +1 efficiently.

Hyperbolic Tangent Function Output:

$$O_j = \tanh(H_j) = \frac{e^{Hj} - e^{-Hj}}{e^{Hj} + e^{-Hj}} \qquad (4)$$

ranges from [-1 , +1]

Soft max Function Output:

$$O_j = \sigma(H_j) = \frac{e^{Hj}}{\sum_{k=1}^{m} e^{Hk}} \qquad (5)$$

ranges between [0 , 1]

The function can be termed as probability distribution and the output Oj is called as network's pseudo-probability or estimated probability of input classification function. There are two mode for optimization through gradient descent algorithm and the most commonly used method is batch mode which updates all synaptic weights of all records in training data set. Another method is online or incremental mode which do the same work however with lower efficiency than preceding method. For solution of linear equations through iteration, the technique used for in the category of batch mode is scaled conjugate gradient has become more optimized. Hence by each repetitive iteration, the synaptic weights are continuously updated, the algorithm bring the error surface to more minimum values compare to previous iteration.

The scaled conjugate gradient algorithm has four parameters for development of model that is initial lambda, sigma, interval center and offset. The Hessian Matrix when acting as negative then controlling parameter is lambda. When size of weight change by sigma parameter, Hessian estimate change the derivative of first order of error function.

Similarly, the annealing algorithm boosts the interval center a0 and a_0 –a and a_0+a. So the initial setting of all parameters are:

Initial lambda =0.0000005, Initial sigma=0.00005 & Interval center =0 and Interval offset =0.5

The SPSS prefer to use cross-entropy error function rather than squared error function whenever softmax activation function is selected and applied on output layer. The mathematical form of cross entropy error function is as below:

$$\text{Error} = -\sum_{i=1}^{n} T_j \text{In} O_j \qquad (6)$$

where n= total number of output nodes

T_j=target value (output node j)

O_j= Actual output value (output node j)

V. Results

The main purpose of this research the generation of multilayer perceptron neural network (MLP-NN) model that can actually predict the telecommunication churn by processing of data obtained from telecom industry according to factors. Below mentioned Table III details the information of dataset applied for building ANN model.

In the Table IV provide the total number of neurons used in input, hidden & output layer and all predictor variables (income per month, drop call rate, failure call rate, education, monthly bill amounts, complaint lodges for resolution, unpaid numbers, call duration- total, call duration- average, internet subscribers, Voice service used, total calling, occupation and age).

Moreover, the number of nodes calculated in the hidden layer is 3 while in the output layer is 2, the activation function taken for hidden layer is selected as hyperbolic tangent (tanH) and softmax function in the output layer and error function used as cross-entropy of network information table.

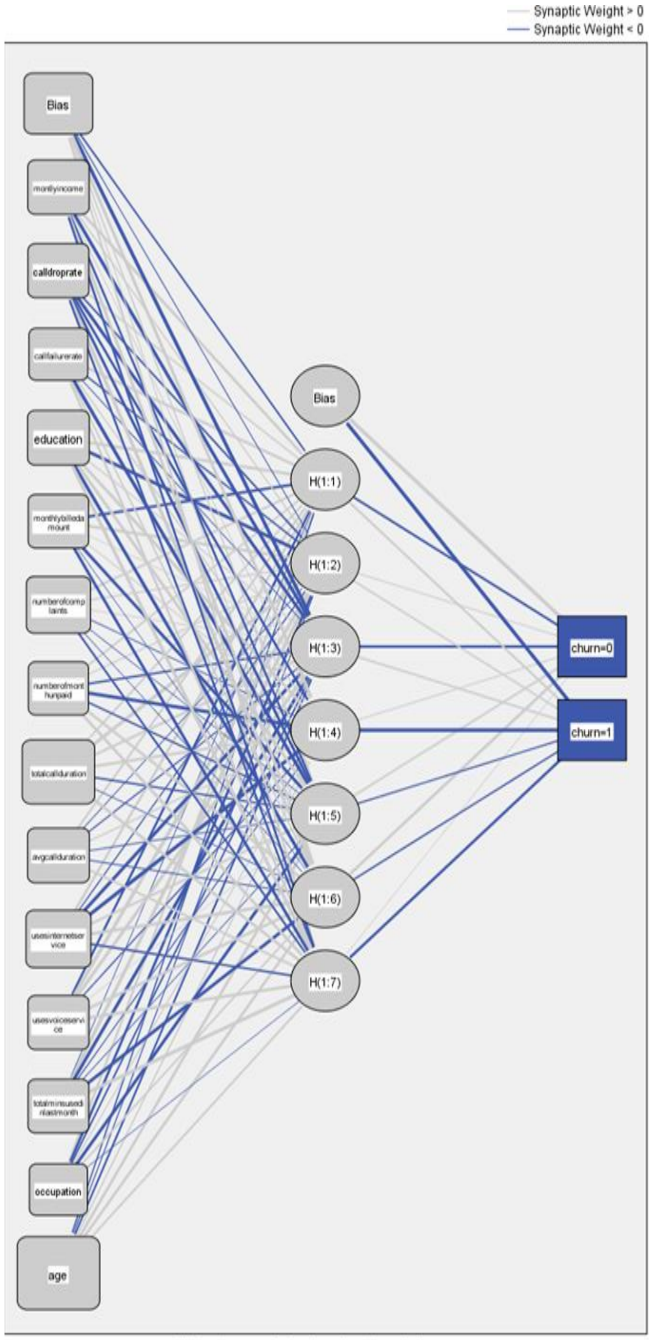
In network diagram extracted from SPSS results, the telecom customer churn (Churn=No, Churn=Yes) from inputs tenure.

Total charge. Similarly, there are three nodes in the hidden layer and two output nodes determined as churn, No-churn depend on results.

The network diagram of artificial neural network is furnished in Fig. 8 which are comprised of input layer, hidden layer and output layer after putting the data in software. The parameter estimates table is given in Table V while model summary in Table VI provide details information about training and testing data set. For both training and testing while the cross entropy error is brought down to minimum possible level during training processing. The power of the model to forecast the outcome of the model is calculated through small value of cross entropy error is 1623.861. The percentage of incorrect prediction is 22.7%. There are 10 consecutive steps completed was performed where there is no more reduction in error function during testing of sample is performed categorical dependent variable outcome.

TABLE. III. The Case Processing Summary

Case Processing Summary			
		N	Percent
Sample	Training	16421	80.20%
	Testing	4047	19.80%
Valid		20468	100.00%
Excluded		0	
Total		20468	



Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Softmax

Fig. 8. The Network Diagram of ANN.

TABLE. IV. THE NETWORK INFORMATION

Network Information			
Input layer	Covariates	1	monthly income
		2	Call drop rate
		3	Call failure rate
		4	education
		5	Monthly billed amount
		6	number of complaints
		7	Number of month unpaid
		8	Total call duration
		9	Avg. call duration
		10	Uses internet service
		11	uses voice service
		12	Total mins used in last month
		13	occupation
		14	age
	Number of Units		
	Rescaling Method of Covariates		Standardized
Hidden Layers(s)	Number of Hidden Layers		
	Number of Units in Hidden Layer1		
	Activation		Hyperbolic tangent
Output Layer	Dependent Variables	1	Customer churn
	Number of Units		
	Activation Function		Softmax
	Error Function		Cross-Entropy
a. Excluding the bias unit			

TABLE. V. THE PARAMETERS ESTIMATES

Parameter Estimates										
Predictor		Predicted								
		Hidden Layer1							Output Layer	
		H(1:1)	H(1:2)	H(1:3)	H(1:4)	H(1:5)	H(1:6)	H(1:7)	[churn =0]	[churn =1]
Input Layer	(Bias)	-.202	-.046	-.559	.401	.151	.051	.245		
	monthly income	.362	.130	-.457	.036	-.275	.370	-.278		
	calldroprate	.166	-.201	-.436	-.311	-.361	.459	-.263		
	callfailurerate	.313	-.164	-.038	.125	-.467	.200	.311		
	education	.353	-.426	.106	.384	.376	-.441	.454		
	monthlybilledamount	-.308	.349	.020	.065	-.394	-.002	-.372		
	Number of complaints	.249	.135	.044	-.050	-.017	-.008	-.247		
	numberofmonthunpaid	.205	.110	-.203	-.426	-.180	.345	.368		
	Total call duration	.236	.473	.279	.328	-.195	-.032	.364		
	avgcallduration	.255	-.037	-.117	.172	-.052	-.058	.262		
	usesinternetservice	.331	-.100	-.439	-.487	.278	.379	-.220		
	uses voice service	-.295	.492	.439	-.119	.002	.337	.363		
	totalminsusedinlastmonth	-.027	.488	-438	-.056	-.027	-.423	.443		
	occupation	.280	-.336	-.194	.075	-.406	.215	-.005		
	age	-.127	-.057	-.243	.164	.366	.355	.249		
Hidden Layer1	(Bias)								.799	-1.132
	H(1:1)								-.274	.258
	H(1:2)								.073	.183
	H(1:3)								-.291	.222
	H(1:4)								.086	-.490
	H(1:5)								.238	-.138
	H(1:6)								.367	-.180
	H(1:7)								.007	-.354

TABLE. VI. THE MODEL SUMMARY

Model Summary		
Training	Cross Entropy Error.	4856.634
	Percent Incorrect Predictions.	9.10%
	Stopping Rule Used.	1 consecutive step(s) with no decrease in error a
	Training Time.	00:00.7
Testing	Cross Entropy Error.	1196.824
	Percent Incorrect Predictions.	9.10%
Dependent Variable: Customer Churn		

For, classification (confusion matrix) is drawn in Table VII. When the predicted outcome is greater than 0.5 the outcome is determined as Yes (churn=Yes). The over percentage of training, 77.3% classification of training data was performed.

The pseudo-probability is calculated in box plot are drawn in Fig. 9. The dependent variable (Churn=Yes or No), the pseudo-probabilities obtained from whole dataset is displayed in box plots and each value greater than 0.5 display correct predictions. The first box plot extended from left to right are customer No churn in the category of No churn, the second plot indicates the classification of churn churn=No, although the values related are Yes category.

TABLE. VII. THE CLASSIFICATION

Classification				
Sample	Observed	Predicted		
		No	Yes	Percent Correct
Training	No	14926	0	100.00%
	Yes	1495	0	0.00%
	Overall Percent	100.00%	0.00%	90.90%
Testing	No	3679	0	100.00%
	Yes	368	0	0.00%
	Overall Percent	100.90%	0.00%	90.90%
Dependent Variable: Customer Churn				

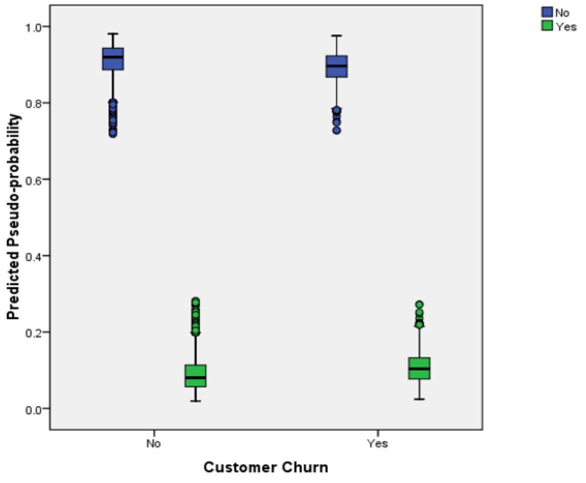


Fig. 9. The Box Plot.

Similarly, the third box plot actually relates to Yes category although predicted in Category churn=No and fourth plot indicated probability and actual are classified as customer churn=Yes.

To determine the possible cutoff point, the sensitivity versus specificity is classified through receiver operating characteristic curve (ROC) is illustrated in Fig. 10.

The combine sensitivity and specificity (1-false positive rate) is showing he random diagonal line is drawn from lower left side to upper right side at 45 degree and greater the accuracy can be achieved depend on the distance from 45 degree base line through classification process.

The mathematical determination of the area under the curve can determine in Table VIII. The probability predicted as 79% displays in the model having the customer churn= Yes and churn=No are randomly selected clearly reveal the pseudo-probability of churn prediction in the category churn=Yes.

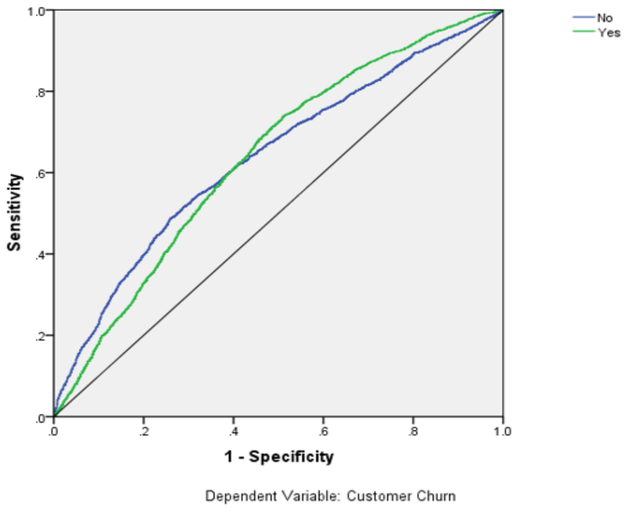


Fig. 10. The Receiver Operating Characteristic Curve.

TABLE. VIII. THE CURVE FOR AREA UNDER (AOC)

Area Under The Curve		
	Area	
Churn	No	0.79
	Yes	0.79

The cumulative gain that displays the classification of telecom customer churn calculated through artificial neural network against classification of prediction through chance. The fifth point in the curve at category churn=No (50% and 40%), meaning if dataset is rated and churn cases are predicted through pseudo-probability of category churn=No ,then it is not difficult to determine that top 40% cases contain 50% of total cases actually take churn=No.

Simply, the gain given in Fig. 11 is the determination of effectiveness of classification model that the correct prediction out of total model against the prediction determine without using a model. The greater the distance of baseline curve main incline line the greater the gain a model have, which measure the excellent performance.

On other hand, lift curve drawn in Fig. 12 also evaluate the performance of the model according to the portion of population and give clear view of benefit of using the model to scenario where there is no predictive modeling. By comparing the gain curve with lift curve, it is determine that 93% value of gain curve , the lift factor is determine as 2.5 on lift curve.

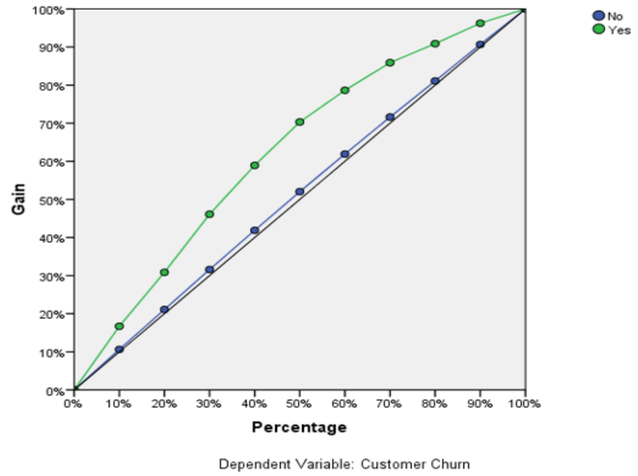


Fig. 11. The Gain Curve.

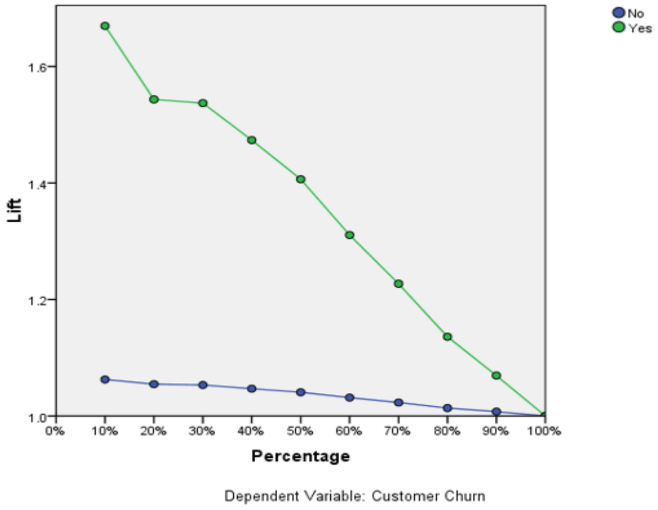


Fig. 12. The Lift Curve.

This figure indicate the sensitivity of model according to change of each input variable. The greatest impact from independent variable that is tenure/subscription of customer with company classify the customers to either churn =Yes or Churn= No. On second position, the churn is mostly affected by total charges.

Table IX is giving the importance of each variable used in main data set. The importance along with percentage is given in more detail for each variable where age has got the highest values equal to 100% followed by call duration (Total) equal to 63%, complaint lodge 33.3%, call duration(Avg) 25%, education 24%, income per month 23.30%, Drop call rate 21.9%, monthly income 20.1%, unpaid numbers 19.1% and occupation 11.4% at lowest level. All these percentage values are illustrated in bar chart provided in Fig. 13 for better understanding and arrange in descending orders for easy understanding.

TABLE. IX. THE INDEPENDENT VARIABLE IMPORTANCE

Independent Variable Importance		
	Importance	Normalized Importance
Income Per Month	.053	23.30%
Drop Rate Call	.050	21.9%
Failure Call Rate	.032	14.0%
Education	.057	24.7%
Monthly Bill Amount(Rs.)	.046	20.1%
Complaints Lodge	.076	33.3%
Unpaid Numbers	.044	19.1%
Call Duration(Total)	.147	63.9%
Call Duration (Avg.)	.057	25.0%
Internet Subscribers	.083	36.2%
Voice Service Used	.051	22.1%
Total Calling Minutes	.049	21.3%
Occupation	.026	11.4%
Age	.229	100.0%

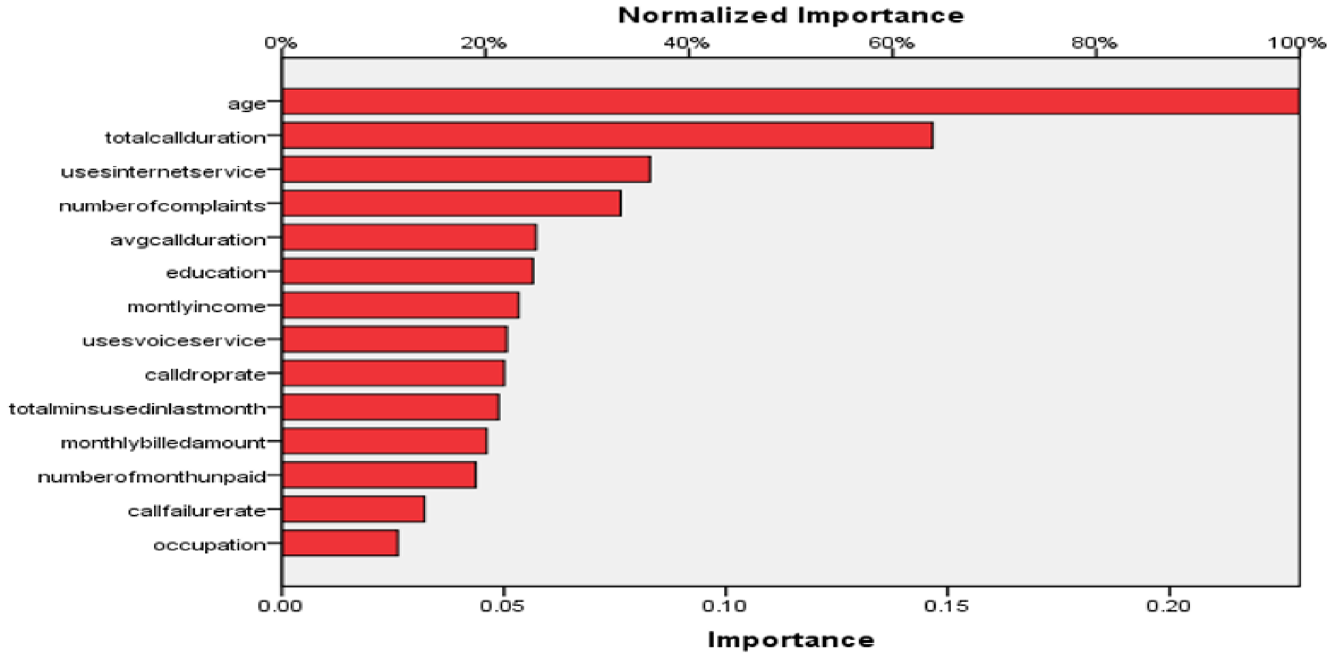


Fig. 13. The Normalized Importance.

VI. Conclusion

The prediction and management of customer churn has become more important task due to liberalization of cellular market. Timely prediction of loyal customers intended to leave the company can help identification and subject to the proactive action in order to retain them. Therefore, building of accurate and precise churn model is necessary not only for telecom companies' owner and practitioners as well. To determine a promising solution for maintaining strong customer baseline, telecom churn prediction has taken a shape of modern day research problem to issue an early warning system for switching over subscribers. The main theme of this paper is application of artificial neural network modeling in customer churn prediction in most volatile market of Pakistan telecommunication industry of the data obtained from all cellular and fixed operators of Pakistan. The research is already in line with literature review and has declared the Artificial Neural Network (ANN) performing excellent than other classification techniques. The already prevailing technique of back propagation algorithm is used to train the model for prediction of telecom customer churn. In the last part of the research the impact of each variable on telecom customer churn in term of normalized importance is provided to calculate the impact of each aspect.

References

[1] Hung, Chihli, and Chih-Fong Tsai. "Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand." Expert systems with applications 34, no. 1 (2018): 780-787.

[2] Berry, Michael JA, and Gordon S. Linoff. Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons, (2013) pp. 12-19.

[3] Laudon, Kenneth C., and Jane P. Laudon. "Managing the digital firm." Managing Information Systems (2014): 197-200.

[4] Kim, Sangkyun. "Assessment on security risks of customer relationship management systems." International Journal of Software Engineering and Knowledge Engineering 20, no. 01 (2010): 103-109.

[5] Burez, Jonathan, and Dirk Van den Poel. "CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services." Expert Systems with Applications 32, no. 2 (2017): 277-288.

[6] Khodakarami, Farnoosh, and Yolande Chan. "Evaluating the success of customer relationship management (CRM) systems." In Proceedings of the European Conference on Information Management & Evaluation,. (2015) pp. 253-262.

[7] Uturytė-Vrubliauskienė, Laura, and Mantas Linkevičius. "Application of customer relationship management systems in Lithuanian mobile telecommunications companies/Science–Future of Lithuania 5, no. 1 (2013): 29-37.

[8] Lee, Jonathan, Janghyuk Lee, and Lawrence Feick. "The impact of switching costs on the customer satisfaction-loyalty link: mobile phone service in France." Journal of services marketing 15, no. 1 (2011): 35-48.

[9] Aydin, Serkan, and Gökhan Özer. "The analysis of antecedents of customer loyalty in the Turkish mobile telecommunication market." European Journal of marketing39, no. 7/8 (2015): 910-925.

[10] Wei, Chih-Ping, and I-Tang Chiu. "Turning telecommunications call details to churn prediction: a data mining approach." Expert systems with applications 23, no. 2 (2012): 103-112.

[11] Ahn, Jae-Hyeon, Sang-Pil Han, and Yung-Seop Lee. "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry." Telecommunications policy 30, no. 10-11 (2016): 552-568.

[12] Van den Poel, Dirk, and Bart Lariviere. "Customer attrition analysis for financial services using proportional hazard models." European journal of operational research 157, no. 1 (2014): 196-217.

[13] Petrova, Ekaterina, Pieter Pauwels, Kjeld Svidt, and Rasmus Lund Jensen. "In search of sustainable design patterns: Combining data mining and semantic data modelling on disparate building data." In Advances in Informatics and Computing in Civil and Construction Engineering, Springer, Cham, (2019) pp. 19-26.

[14] Nabavi, Sadaf, and Shahram Jafari. "Providing a customer churn prediction model using random forest and boosted trees techniques (case study: Solico Food Industries Group)." vol 3 (2013): 1018-1026.

[15] Mestre, Maria Rosario, and Pedro Vitoria. "Tracking of consumer behaviour in e-commerce." In Proceedings of the 16th International Conference on Information Fusion, pp. 1214-1221. IEEE, 2013.

[16] V. Umayaparvathi, K. Iyakutti,, “Attribute Selection and Customer Churn Prediction in Telecom Industry”, Proceedings of the IEEE International Conference On Data Mining and Advanced Computing,(2019).

[17] Chen, Kuanchin, Ya-Han Hu, and Yi-Cheng Hsieh. "Predicting customer churn from valuable B2B customers in the logistics industry a case study." Information Systems and e-Business Management 13, no. 3 (2015): 475-494.

[18] Arsanl, Taner, and Safa Çimenli. "Churn Analysis and Prediction." International Journal of Computer Science and Information Security 14, no. 8 (2016): 550.

[19] Junxiang Lu, "Predicting Customer Churn in the Telecommunications Industry -- An Application of Survival Analysis Modeling Using SAS", In SAS Proceedings, SUGI27, pages 114-127, 2012.

[20] Nath, Shyam V., and Ravi S. Behara. "Customer churn analysis in the wireless industry: A data mining approach." In Proceedings-annual meeting of the decision sciences institute, vol. 561,. (2013) pp. 505-510.

[21] Liu, Peng, Naijun Wu, Chuanchang Huang, Bingrong Wei, Libo Wang, and Zhen’an He. "Research on Applying Data Mining to Telecom CRM." In International Forum of Information System Frontiers-2016 Xian International Symposium.

[22] Jayne, Chrisina, Andreas Lanitis, and Chris Christodoulou. "Neural network methods for one-to-many multi-valued mapping problems." Neural Computing and Applications 20, no. 6 (2016): 775-785.

[23] Adebayo, adelaja oluwaseun, and mani shanker chaubey. "data mining classification techniques on the analysis of student’s performance." gsj 7, no. 4 (2019).

[24] Andreu, David, and Marc Torrent. "Prediction of bioactive peptides using artificial neural networks." In Artificial Neural Networks, Springer, New York, (NY, 2015) pp. 101-118.

Modified Seam Carving by Changing Resizing Depending on the Object Size in Time and Space Domains

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

***Abstract*—Modified seam carving by switching from the conventional method to resizing method depending on the object size is proposed. When the object size is dominant in the scene of interest, the conventional seam carving shows deformation of components in the object. To avoid the situation, resizing method is applied rather than the conventional seam carving in the proposed method. Also, the method for video data compression based on the seam carving not only in image space domain but also in time domain is proposed. It is specific feature that original quality of video picture can be displayed when it is replayed. Using frame to frame similarity defined with histograms distance between the neighboring frames, frames which have great similarity can be carved results in data is compressed in time domain. Moreover, such carved frame can be recorded in the frame header so that the carved frame can be recovered in reproducing the compressed video. Thus, video quality can be maintained, no degradation of video quality at all. Compression ratio is assessed with the several video data. It is obvious that data compression ratio of the proposed space and time domain seam carving is greater than that of the conventional space domain seam carving.**

Keywords—Seam carving; data compression in time and space domains; video data compression

I. Introduction

Image data compression methods which allow compression using removing content less portion of image from original images (Seam Carving[1]) are proposed [1]- [8]. Seam carving with OpenCV, Python, and scikit-image[2] is available. The graph cut concept was proposed together with dynamic graph cut based on Markov random fields [7]. These methods allow image segmentation, texture extraction, image mosaic, image energy minimization, etc. Also, video cutout method is proposed for targeting an image portion through panning and scanning together with a content based video retargeting [9]. Then content resizing based method of the well-known seam carving is proposed [10]. Furthermore, video carving method is proposed [11] together with acceleration algorithm for video carving [12]. These methods are referred to "the conventional seam carving method (See Appendix).

Seam carving is supported by Adobe[3] , GIMP[4] , digiKam[5] and Image Magick[6] already [2]. Video carving allows targeting to content rich time periods in the video stream so that video content is shortened. Meanwhile, some portions of time series of video contents are deleted so that not natural object movement appears some time.

Video carving method proposed here is based on the distance between histograms of the objects in concern in the two adjacent frames, the current frame and the next frame. If the distance is not less than a prior determined threshold, then such frame can be deleted for video data compression. Also, the deleted frame number is stored in the header of the video content so that object movement is much natural in playing the compressed video contents by referring the deleted frames (the deleted frame is replaced to the previous frame) in comparison to the conventional video carving method [13]. On the other hand, improved seam carving by switching from the conventional method to resizing method depending on the object size is proposed. When the object size is dominant, the conventional seam carving shows deformation of components in the object. Sometimes it would be fanny shapes of the objects. To avoid the situation, image size change method is applied rather than the conventional seam carving in the proposed method.

The method proposed here is preprocessing of the object size detection before applying space domain of seam carving. If the object size is dominant in the whole image, object shape cannot be maintained its shape is the conventional seam carving applied to the image of concern. To avoid the situation, object size is detected prior to the seam carving in the proposed method. Also, it is possible to compress the video data in concern by considering the object moving speed. If the moving speed is below a threshold, then such frames can be removed. It is called time domain seam carving in the proposed method.

The following section describes the proposed method followed by some experimental results with video contents. Finally, conclusions with some discussions are described.

[1] https://www.pyimagesearch.com/2017/01/23/seam-carving-with-opencv-python-and-scikit-image/

[2] https://scikit-image.org/

[3] https://www.adobe.com/

[4] https://www.gimp.org/

[5] https://www.digikam.org/

[6] https://imagemagick.org/index.php

II. Proposed Method

A. Proposed Seam Carving Method

It is not always that the conventional space domain seam carving works well. If the major portion of image which must be maintained is shared dominantly in the almost over the original image, then seam carving does not work as is shown in Fig. 1, namely, the object size is dominant in the image, the conventional seam carving method shows such like very fanny object image shown in Fig. 1. To avoid such situation, the preprocessing which is shown in Fig. 2 is proposed.

If the approximated circle or rectangle shares dominant area of the image in concern, then seam carving is applied to the entire image and if not, then seam carving is applied to the rest of the image. Example of the edge detected result of the image of Fig. 1 is shown in Fig. 3(a). Fig. 3(b) shows the approximated edge image with ellipsoids.

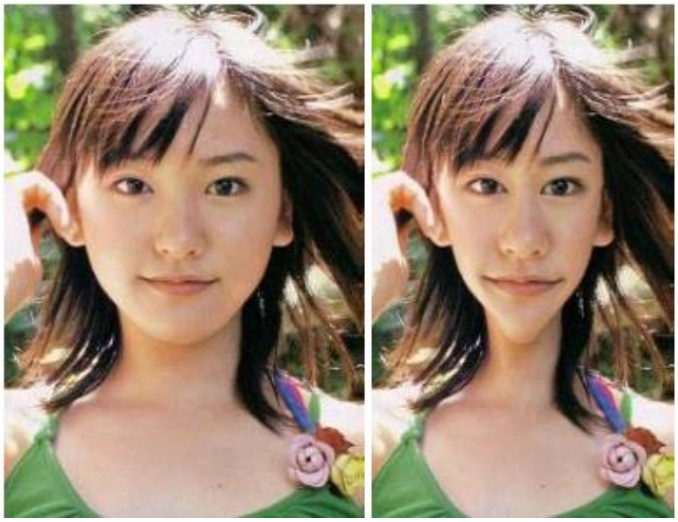

(a) Original Image (b) Seam carved image

Fig. 1. Image after the Seam Carving Referring to Energy List and m List in Accordance with the Conventional Seam Carving Algorithm.

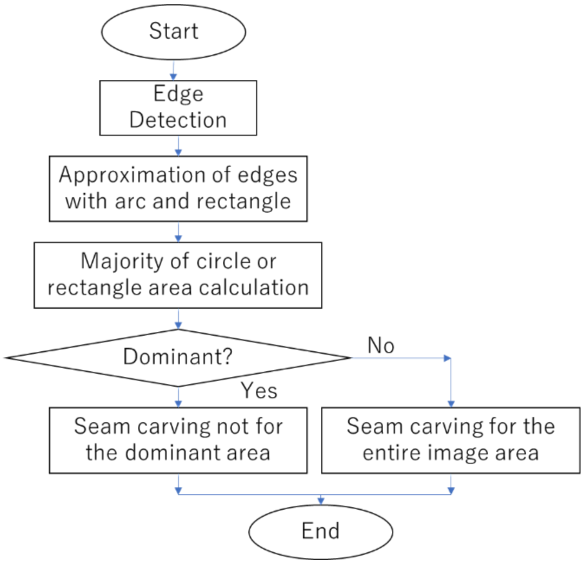

Start
Edge Detection
Approximation of edges with arc and rectangle
Majority of circle or rectangle area calculation
Dominant?
No
Yes
Seam carving not for the dominant area
Seam carving for the entire image area
End

Fig. 2. Proposed Preprocessing for Seam Carving in Spatial Domain.

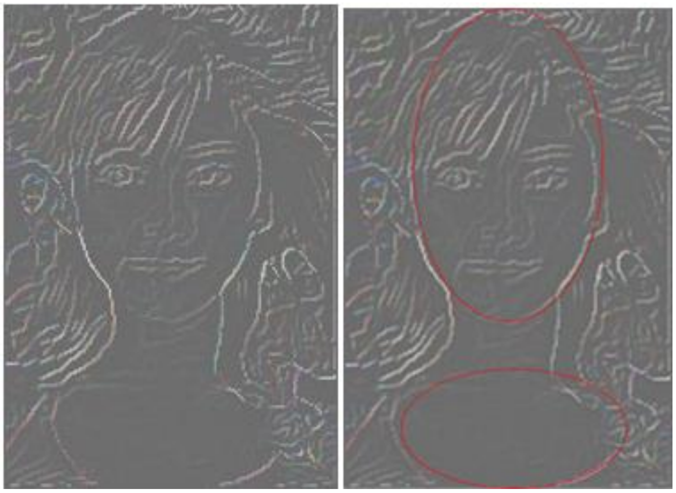

(a) Edge detected image (b) Approximated with ellipsoids

Fig. 3. Example of the Edge Detected Result of the Image of Fig. 1.

In this case, resizing method is applied to the object image. Namely, image aspect ratio is changed depending on the designated compression ratio. Then interpolation is applied to the resized image when it is reproduced. Therefore, original shapes of objects are maintained when it is reproduced. The reproduced images with the conventional and the proposed seam carving methods are shown in Fig. 4. It is obvious that the reproduced image with the proposed seam carving method is much natural than that with the conventional method. The compression ratio of the proposed method is almost same comparing to that of the conventional method.

B. Proposed Space and Time Domain Seam Carving

The proposed video seam carving method applies another seam carving in time domain. Process flow of the proposed time domain seam carving is shown in Fig. 5. Objected regions are already selected in the space domain seam carving.

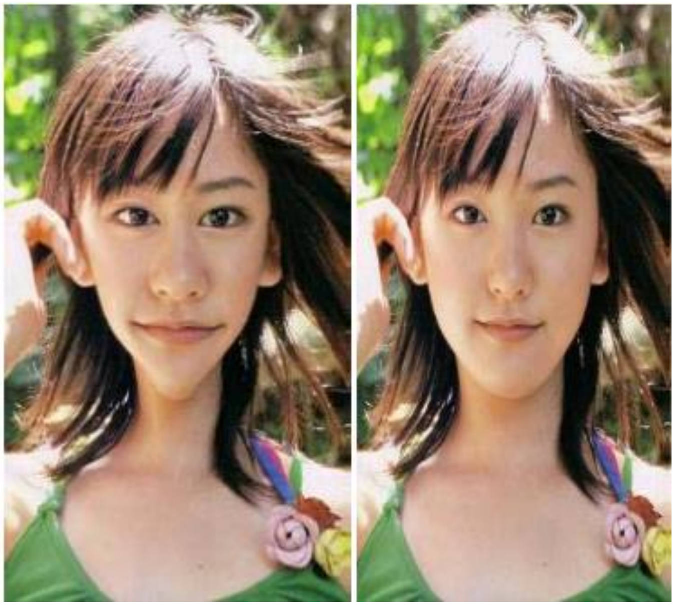

(a) Conventional (b) Proposed

Fig. 4. Reproduced Images with the Conventional and the Proposed Seam Carving Methods.

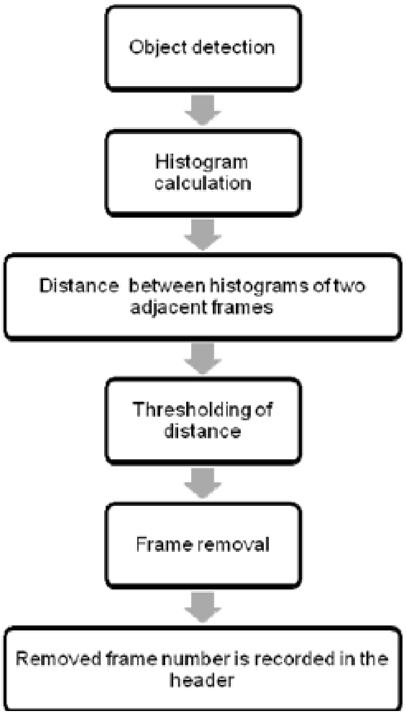

Object detection
Histogram calculation
Distance between histograms of two adjacent frames
Thresholding of distance
Frame removal
Removed frame number is recorded in the header

Fig. 5. Process Flow of the Time Domain Seam Carving.

Object shape is changed in natural by frame by frame because the object moves typically. The proposed method uses object image histogram change between two adjacent frames. It is assumed that histograms between two adjacent frames are not changed then such two frames are redundant so that it may be removed because such frames are redundant.

Contour of the object can be defined after the object detection. Then histogram of intensity component of the detected object is calculated after the RGB to HIS conversion process. After that comparison of current histogram and the previous histogram is made. Histogram distance is defined with Bhachattaryya distance[7], D which is expressed as follows [14]:

$$D = \sum_{i=1}^{256} \sqrt{p_i q_i} \qquad (1)$$

where p and q denote frequency of the intensity i for the current and the previous object images. By removing the frames which have a small D, under a prior determined threshold, data compression can be done. The removed frame number is recorded in the header. For instance, bitmap image format has 14 byte of header region. In the region, there is two byte of file type information followed by four byte of file size information. Then two sets of two byte of preserved regions are followed by together with four bytes of offset information.

In the preserved region, it is possible to put in the information of the removed frames. By using this information, it can be refrained the previous frames when it is played onto display.

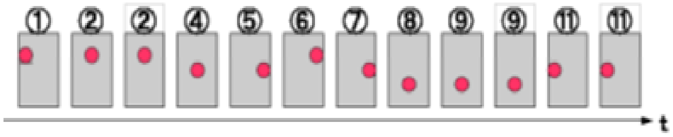


Fig. 6. Procedure of the Proposed Time Domain Seam Carving Method.

Because the conventional video carving makes shortened the video contents in time domain, it may give some strange impressions of object movement. Meanwhile, the proposed time domain seam carving does not give such an impression at all. Fig. 6 shows illustrative schematic view of the proposed time domain seam carving.

Red circle shows the extracted object. Histogram of the intensity of the pixels in the contour of the extracted object is calculated followed by Bhachattaryya distance between histogram calculations. Then the distance D is below threshold, in this case, the frame number 2 and 3, 9 and 10 as well as 11 and 12 of D are below threshold so that the frame number 3, 10 and 12 are removed. Then compressed data, in this case, compression ratio is 3/4 is stored together with the removed frame numbers in the header. When the stored compressed video data is played, the removed frames are added to the compressed video content so that the object movement is somewhat natural. Therefore, the proposed method is for storage volume saving not for time saving.

C. *Specific Feature of the Proposed Space and Time Domain Seam Carving*

The most specific feature of the proposed method is that space and time domain seam carving portions can be coded and stored in the storage memory. The information of seam carving portion can be stored in the header information of coded video picture. Therefore, original video picture can be replayed referring to the seam carving portions in the storage memory.

III. Experiments

A. *Space Domain Seam Carving*

Fig. 7 shows the procedure of the proposed space domain seam carving method. Red circle indicates object which is detected with OpenCV library of cascade of boosted classifiers based on Haar-like features[8] which is provided by

OpenCV library of CvHaarFeature, CvHaarClassifier,

CvHaarStageClassifier, CvHaarClassifierCascade.

Then object can be tracked by frame by frame. The detected object is meaning full so that it is remained through space domain seam carving while the background is removed by the space domain seam carving. The rectangle areas in the time series of image data after the space domain seam carving show the detected object. Location of the detected object can be stored so that space domain seam carving can easily be done.

[7] https://en.wikipedia.org/wiki/Bhattacharyya_distance

[8] https://algorithm.joho.info/image-processing/haar-like-feature-value/

Fig. 8 shows the test image generated with the sphere function provided by the well-known free software of the PovRay[9] of computer graphic software tool [15]. 45 frames of 512 by 384 pixels of test images are created. The location of object is changed by frame by frame. Fig. 9 shows the process flow of the proposed space domain seam carving method.

Object detection in the case is easy because only thing OpenCV must do is to detect “Sphere”. Using the simulated time series of test image, data compression ratio is evaluated.

The experimental results of space domain seam carving are as follows:

Image size: 512 by 384 pixels→100 by 100 pixels

Data volume a frame: 576KB→48KB

The number of frames: 45→45

Data volume after the space domain seam carving: 26MB →2.6MB

In this case, 1/10 of data compression is confirmed with same image quality.

B. *Time Domain Seam Carving*

The experiments with two video contents which are shown in Fig. 10 are conducted. Fig. 10(a) is the video of the remains in Greek and (b) the video the man walking the street situated in front of big trees. Old remain is captured from the different aspect angles so that not only clouds in the background but also object of the old remain is changed in shape together with shadow in the scene number 1. Meanwhile, the walking man of the scene number 2 moves from right hand side to left direction so that not only waving leaves of the big trees by the winds but also the man is changed in location together with their shadows.

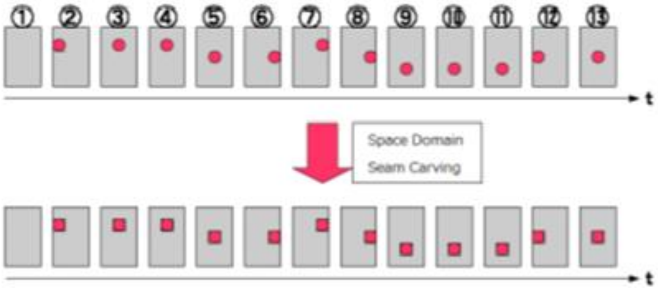


Fig. 7. Procedure of the Proposed Space Domain Seam Carving Method.

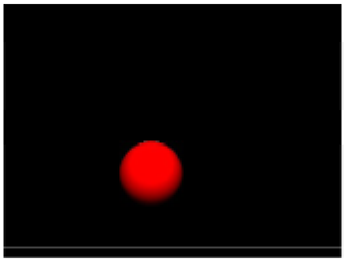


Fig. 8. Simulated Test Image for Evaluation of Data Compression Ratio of the Proposed Space Domain Seam Carving (Image Size is 512 by 384 Pixels).

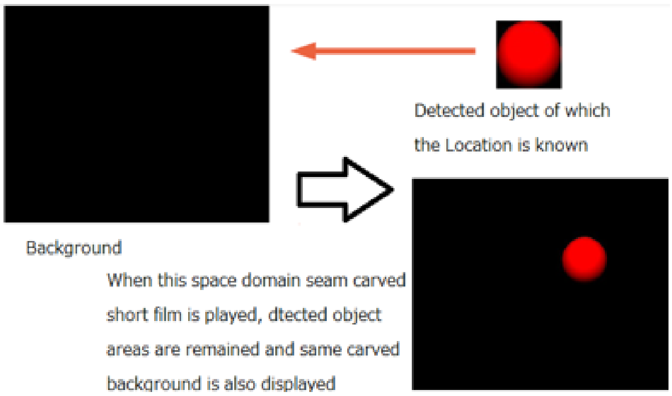
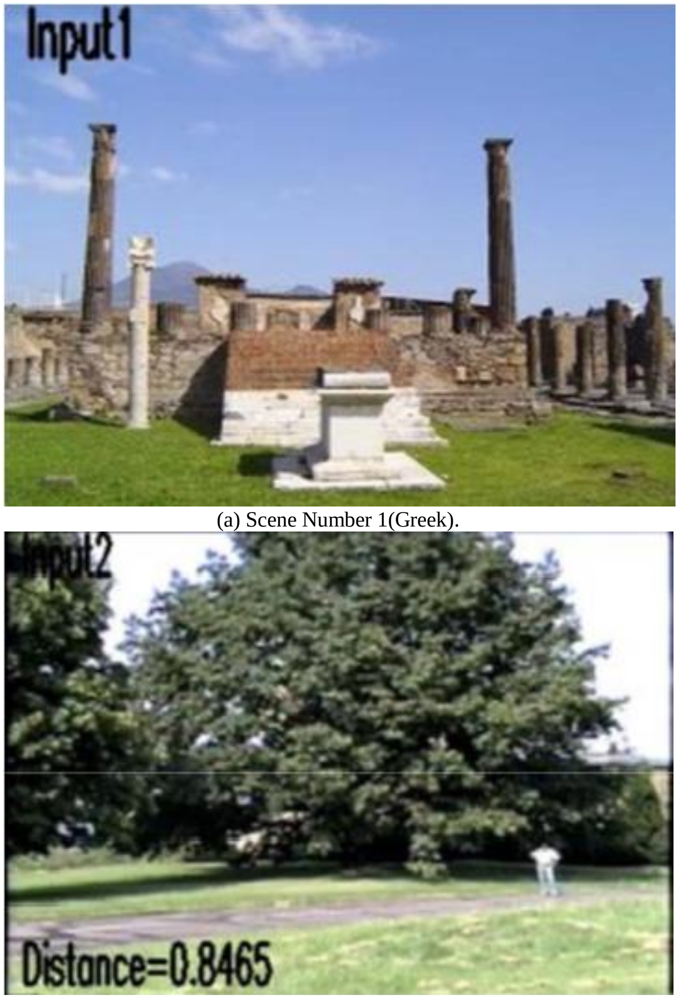

Detected object of which
the Location is known
Background
When this space domain seam carved
short film is played, dtected object
areas are remained and same carved
background is also displayed

Fig. 9. Process Flow of the Proposed Space Domain Seam Carving.



(a) Scene Number 1(Greek).

(b) Scene Number 2(Walking).

Fig. 10. Test Scenes used for Evaluation of Data.

Results of moving picture data compression for test scene number 1 is as follows:

Image size: 244 by 360 pixels

Data volume a frame: 257KB

Threshold: 0.0479

[9] http://www.povray.org/download/

The number of removed frames: 22 frames out of 132 frames so that 11/61 of data compression ratio is accomplished which is corresponding to the compressed data volume of 5654KB. On the other hand, the results of moving picture data compression for test scene number 2 is as follows,

Image size: 240 by 360 pixels

Data volume a frame: 253KB

Threshold: 0.0435

The number of removed frames: 47 frames out of 182 frames so that 47/182 of data compression ratio is accomplished in this case which is corresponding to the compressed data volume of 11891KB.

The conventional video seam carving requires much shorter time for play the compressed video with a little bit funny impression. Meanwhile, the proposed time domain seam carving requires the completely same time for play the compressed video without any defect because the removed or carved frames are replaced with the previous frame in concern.

IV. Conclusion

It is confirmed that the proposed video seam carving method is effective; in particular, time domain seam carving achieves around 1/5-1/6 of data compression ratio. It is obvious that the data compression ratio depends on the changes in the moving pictures. Once the object is detected with OpenCV library through training with the objective moving picture itself, then time domain seam carving is performed effectively.

The conventional image space domain seam carving by resizing algorithm shows 1/2-1/10 of data compression ratio depending on redundant image areas in the background in the image. Therefore, 1/10-1/600 of significantly high data compression can be achieved with a slight image degradation if the proposed space and time domain seam carving data compression is applied.

The specific feature of the proposed time domain seam carving is that it requires the completely same time for play the compressed video without any defect because the removed or carved frames are replaced with the previous frame in concern.

Further experimental studies are required for validation of the proposed seam carving method in time and space domains.

Appendix: Conventional Space Domain Seam Carving

Firstly, seam carving is assumed to be applied to the video contents in image space domain already. In the space domain seam carving, objected image portion is extracted first. The method for object extraction is used to be based on OpenCV library [13]. How-to build a cascade of boosted classifiers based on Haar-like features is provided by OpenCV library of. CvHaarFeature, CvHaarClassifier, CvHaarStageClassifier, CvHaarClassifierCascade.

Then the energy concentration in the other image regions is calculated for seam carving. Poor energy regions, then, are removed after that. This procedure is the conventional seam carving method. An example of the conventional image space seam carving is shown in Fig. 11. In this case, 400 by 300 pixels of the original image are resized in 200 by 300 pixels. It is also possible to determine remaining objects and removing objects. In this example, the green portion of image would like to be remained while the red portion of image would like to be removed. Thus, image is resized results in image data compression.

There are some web sites which allow space domain seam carving. Using web site provided seam carving tool, image resizing based on space domain seam carving can be done. Fig. 12 shows an example of seam carving image resizing through the web site.

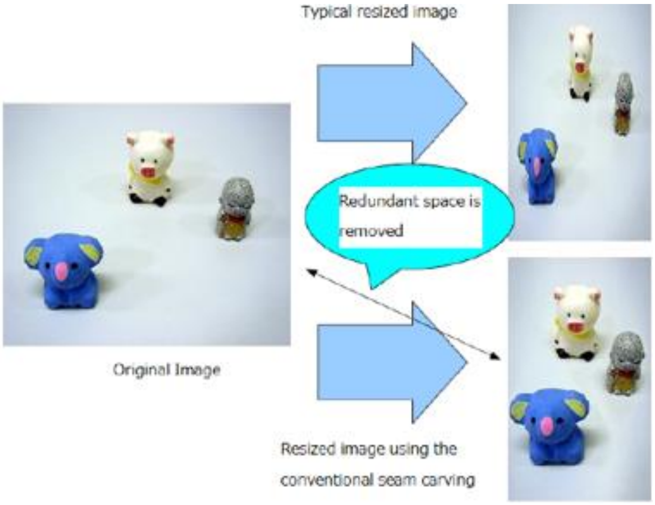

Typical resized image
Redundant space is removed
Original Image
Resized image using the conventional seam carving

Fig. 11. Example of the Conventional Image Space Domain of Seam Carving.

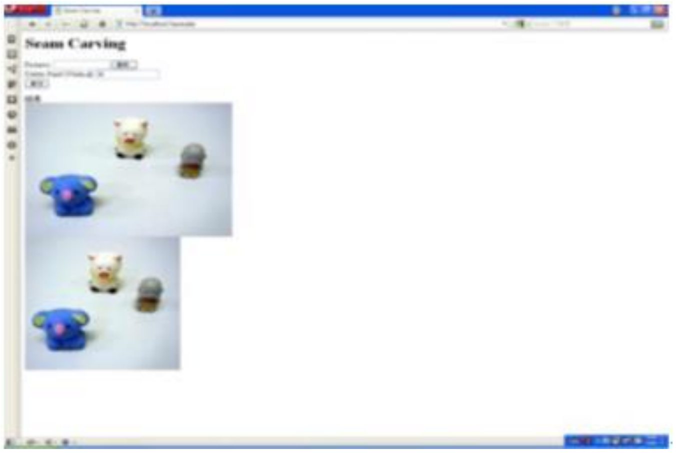

Seam Carving

Fig. 12. Example of Seam Carving Image Resizing through the Web Site.

Space domain seam carving allows image resizing with some intentional conditions. From the web site of Dr. Arial Shamir's paper derived PHP top page, Micro Soft Windows version of space domain seam carving tool is created.

The algorithm is as follows:

1) Get the original image together with the image size

2) Get the color index of each pixel in the image

3) Calculate energy list using Euclidian distance between the color index of the pixel in concern and the four neighboring pixels shown in Fig. 13.

4) Take an average over Euclidian distance between pixel #1 and #4 and that between #1 and #5 results in a distance in vertical direction.

5) Take an average over Euclidian distance between pixel #1 and #2 and that between #1 and #3 results in a distance in horizontal direction.

6) Replace the pixel value of concern (#1) with summation of the distances in vertical and horizontal directions

Then energy list is converted to the m list as follows:

①'=①+min(⑦,⑧)
②'=②+min(⑦,⑧,⑨)
⑥'=⑥+min(⑪,⑫)
⑧'=⑧+min(⑬,⑭,⑮)

Where the pixels in the m list is aligned as shown in Fig. 14.

Thus, y pixels in the m list are calculated. To identify the pixels for seam carving, the following algorithm is applied to the m list,

1) Extract minimum value of pixel in the calculated max_y

↓

2) From the max_y-1, the following calculation is made
m_list[x][max_y-1] energy_list[x][max_y-1]

↓

3) The result is corresponding to the following two candidates:

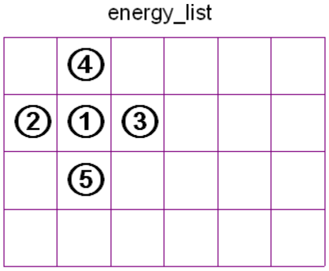
m_list[x-1] [max_y-2], m_list[x][max_y-2],

m_list[x+1] [max_y-2]

4) Go back to (2), and refrain the process (3), then the seam carving portion is extracted referencing to energy list and m list like as shown in Fig. 15.

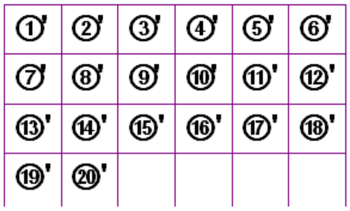
In the above figure, green line shows the highest energy. Low energy pixel which is situated at the far from the highest energy pixel is removed by seam carving. It looks like an onion pealing. Other examples are shown in Fig. 16.

Fig. 17 shows an example of image resizing with intentional condition on whether remove the image portions. In this case, the green rectangle in Fig. 17(b) shows the image portion which must be remained while the red rectangle is the image portion which must be removed. Thus, resultant image of Fig. 17(c) is reduced with space domain seam carving.


energy_list

	④				
②	①	③			
	⑤				

Fig. 13. Calculate Energy List using Euclidian Distance between the Color Index of the Pixel in Concern and the Four Neighboring Pixels.



①'	②'	③'	④'	⑤'	⑥'
⑦'	⑧'	⑨'	⑩'	⑪'	⑫'
⑬'	⑭'	⑮'	⑯'	⑰'	⑱'
⑲'	⑳'				

Fig. 14. Pixels in the m List is Aligned.

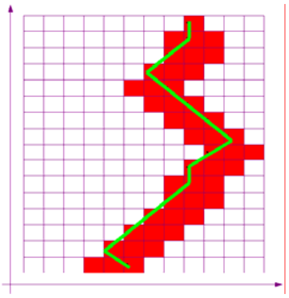
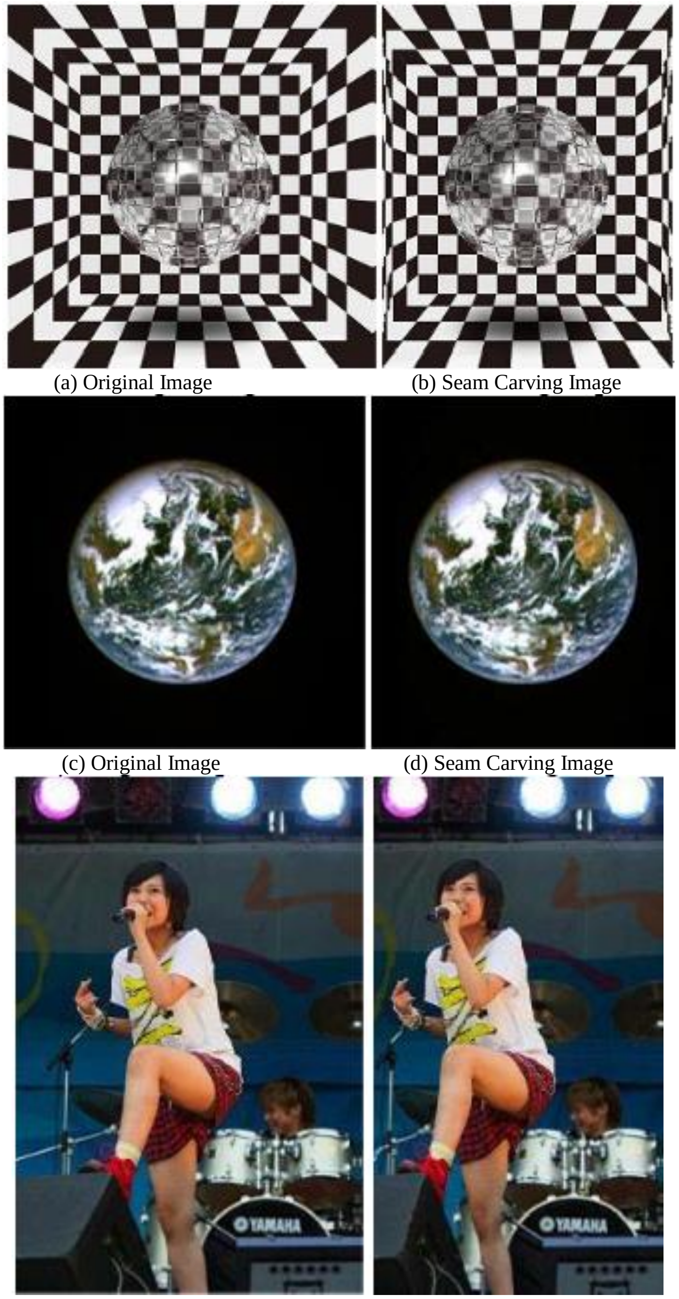


Fig. 15. Seam Carving Portion is Extracted Referencing to Energy List and m List.

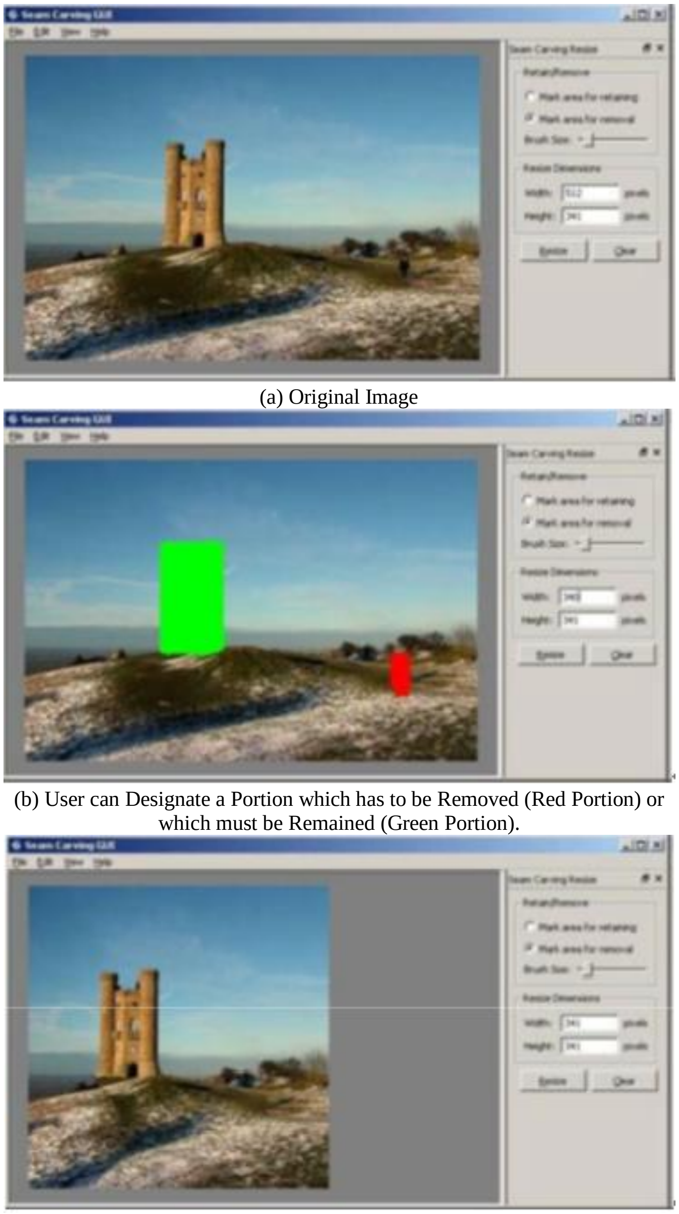


(a) Original Image (b) Seam Carving Image

(c) Original Image (d) Seam Carving Image

(e) Original Image (f) Seam Carved Image

Fig. 16. Other Examples of Space Domain Seam Carving.



(a) Original Image

(b) User can Designate a Portion which has to be Removed (Red Portion) or which must be Remained (Green Portion).

(c) Resized Image

Fig. 17. Example of Image Resizing with Intentional Condition on whether Remove the Image Portions.

ACKNOWLEDGMENT

The author would like to thank Mr. Hiroya Inoue for his effort through the experiments of seam carving programming.

REFERENCES

[1] Kwatra V., Schodl A., Essa I., Turk G., and Bobick A., Graphcut textures: image and video synthesis using graph cuts, ACM Trans. Graph. 22, 3, 277–286, 2003.

[2] Image Retargeting was invented by Vidya Setlur, Saeko Takage, Ramesh Raskar, Michael Gleicher and Bruce Gooch in 2005 http://en.wikipedia.org/wiki/Seam_carving.

[3] Boykov Y., and Kolmogorov V., An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision, IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 9, 1124–1137, 2004.

[4] Wang J., Bhat P., Colburn R. A., Agrawala M., and Cohen M. F., Interactive video cutout, ACM Trans. Graph. 24, 3, 585–594 2005.

[5] Lombaert H., Sun Y., Grady L., and Xu C., A multilevel banded graph cuts method for fast image segmentation, In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), vol. 1, 259–265, 2005.

[6] Liu F., and Gleicher M., Video retargeting, automating pan and scan, In MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia, ACM, 241–250, 2006.

[7] Kohli P., and Torp P. H. S., Dynamic graph cuts for efficient inference in Markov random fields, IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 29, 12, 2079–2088, 2007.

[8] Rav-Acha A., Pritch Y., Lischinski D., and Peleg S., Dynamosaicing: Mosaicing of dynamic scenes, IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 29, 10, 1789–1801, 2007.

[9] Wolf L., Guttmann M., and Cohen-or D., Nonhomogeneous content-driven video-retargeting, In Proceedings of the Eleventh IEEE International Conference on Computer Vision (ICCV '07), 1–6, 2007.

[10] Shai Avidan Ariel Shamir, Seam carving for content-aware image resizing, ACM Transactions on Graphics (TOG) - Proceedings of ACM SIGGRAPH 2007, 26, 3, 10, 2007.

[11] Chen B., and Sen P., Video carving, In Short Papers Proceedings of Eurographics. 2008.

[12] Chen-Kuo Chiang, Shu-Fan Wang, Yi-Ling Chen, Shang-Hong Lai, Fast JND-Based Video Carving With GPU Acceleration for Real-Time Video Retargeting, IEEE Transactions on Circuits and Systems for Video Technology,, 19, 11, 1588-1597, 2009.

[13] The Open Computer Vision Library has more than 500 algorithms, documentation and sample code for real time computer vision. Tutorial documentation is in O'Reilly Book: Learning OpenCV http://www.amazon.com/Learning-OpenCV-Computer-Vision-Library/dp/0596516134

[14] Bhattacharyya, A., On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society 35: 99–109, 1943.

[15] Starting with version 3.7, POV-Ray is released under the AGPL3 (or later) license and thus is Free Softwar http://www.povray.org/download/

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

A Novel Approach to Rank Text-based Essays using Pagerank Method Towards Student's Motivational Element

M Zainal Arifin[1], Naim Che Pee[2], Nanna Suryana Herman[3]
Department of Electrical Engineering, State University of Malang, Indonesia[1]
Faculty of Information & Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia[1, 2, 3]

***Abstract*—Learning outcomes is one of the important factors to measure student achievement during the learning process. Today's learning is more focused on problem-solving and reasoning to existing problems than an ordinary problem. Most exams have been directed to analysis questions for students to think and synthesize. As such is troublesome for most students, they are not ready to answer the question, thus, their answers almost similar to their friends. This implies that the teacher has tried to guide students to work professionally and originally. However, the Teacher facing difficulty assessing student's work, especially if the assignments are conducted online without face-to-face instructions/discussions. To bridge such a gap, the teacher needs a method or algorithm to measure their rank to encourage students making an original answer. This research provides a solution in calculating students' ranks based on the similarity score of the essay answers. Pagerank is a ranking algorithm used by Google, this algorithm utilizes a Markov matrix that contains the direction of similarity score for each student. These scores are computed by the power method until converging. Rank is displayed to the teacher to review the similarity level of students' answers. As such is presented a line chart in which the x-axis refers to the students and the y-axis depicts the level of similarity. Ranking computation in Matlab produces an Eigen vector which acts as the rank measure. The higher the rank, the more similar is their answers to others. Hence, students with high ranks to work their answers more seriously ensure their original thoughts. In conclusion, the similarity score matrix using the PageRank algorithm can contribute to the teacher in providing peer motivation and encouraging student's internal motivation by presenting the ranked-answers presentation.**

***Keywords*—Pagerank; learning outcomes; similarity**

I. Introduction

Online Learning becomes the requirement in industrial era 4.0 to gain the easiest and fastest way of the learning process based on technological development. Teachers need it to get the student's exam results promptly and manageable in digital format. An essay test is still the best method to assess a student's analytical ability to the given problem. The students to practice his/her skills to analyze and synthesize, as a way to develop their knowledge.

In the modern learning goal, a teacher administers an exam at the end of the semester and she or he reviews those exams accordingly, often the teacher finds it difficult to identify the originality of student response and especially their similarity score. Teachers need an alternative method to measure student's scores and their ranks. Thus one can anticipate to whom student needs extra motivation during the learning process.

E-learning application provides similarity checking on essay responses. Unfortunately, as such is not quite informative for the teacher such as an unknown source of students copied work and a similarity relationship between one another. Based on the observations conducted for about a month on several campuses in Indonesia, teachers claim that motivation is compulsory for students, by presenting similarity scores as a visual report, i.e. graph, chart or diagram.

PageRank is a ranking algorithm used by Google to calculate web page rank based on a directed relationship of hyperlinks among others. It uses a power method formula which calculates a Markov matrix to produce Eigen vector. An Eigen value is related to its convergences, As a result, Google named it PageRank. PageRank computations take several hours, depends on the Markov matrix size, to reach convergence with initial Eigen vector related to Markov matrix character.

This research is important to do because lecturers find it difficult to motivate students based on their work as has been investigated by [1], besides that research objective focuses on developing a relationship between students based on the similarity score test using the PageRank method. It consists of parameters: including similarity score and student location (latitude and longitude). Meanwhile, the research scopes are: 1) exams take place is an essay type, and 2) the student's location is based on the current local address detected by GPS or browser with location plugin support.

II. Literature Review

In this section we cover some foundations that grounds this research consisting of student ranking, plagiarism among students, students' motivation and page rank method.

A. Student Ranking

Vieira (2015) claims the superiority of the strategic value is given by the academic ranking process as globally and influences academic quality in scientific activities c. This article accentuates that student rank has a positive impact to the quality and learning process. Unfortunately, the author does not mentions the cause and effect in a case where a student was ranked at the low level. As such should lead to the strategic action to be taken, in which the paper is also lacking of.

Furthermore, Keller (2016) explains that in order to improvinge students' intelligence affective factor takes control, the teacher must have both knowledge and the right strategy to motivate students [2]. Affective intelligence and intellectual intelligence are inseparable to ensure students achieving good learning outcomes.

Surveys carried out by Buckley (2016) reveals that gamified learning has a good impact for the student's learning process. Rosyid (2018) also reveals that game-based learning strategy can creates a fun-learning environment. However, it is inevitable that participation may varies depending on intrinsic or extrinsic motivation of the student concerned. The results of these study are interesting to analyse by educators, the fact that students at any level of education can take part in such learning methods[3]. But it is not discussed if students dislike a learning material, to which strategy the teacher must apply to present the motivation.

Marks (2017) said that education process which includes the effectiveness of learning will affects the quality of learning and increases learning achievement by factors such as course recommendations, student skills, and behavior detection [4].

Denning (2018) said students with the top three academic rank, potentially obtains high test score in the coming exams. Even, they are found to be easily enrolled the college level. This is influenced by the quality and ranking of the school [5]. Students in the higher-ranked schools have higher average exam score than students from lower-ranked schools.

However, Murphy (2018) explained that the potential benefits of schools that have high rankings will be influenced by the confidence ranking of students who are poorly ranked. With the influence of this ranking, investment in education will decline [6]. Meanwhile, student self-motivation is still a new research area. Unfortunately, the author did not explain that good school rank is influenced by students rank well. And there is no relationship between low-ranked students and school's rank. In facts, there are other factors, involved as well such as school's accreditation.

B. *Plagiarisme Over Students*

According to Hu (2015), rules that are made based on detection and focus on giving penalties are more effectively applied to Chinese students. On the contrary, based on other findings, it shows that an education-based approach that motivates them is better than giving a punishment [7].

Cronan (2018) mentions behavior patterns, morals, attitudes are the biggest factors that influence a person to academic integrity especially when not to do plagiarism and share homework. The author gives an overview of 33% and 35% of students working on academic task do plagiarism or sharing homework [8]. The author did not clearly state what type of motivation had been given to students and how many students did plagiarism at the school.

Ehrich (2016) conducted research on 131 Australian students and 173 Chinese students to compare policies against plagiarism, there was no difference in plagiarism but rather understanding and plagiarism behavior [9]. There is no discussion in the paper about the response of students when the teacher motivated them regarding this plagiarism, but the results of this study were quite clear.

Khlifi (2017) conducted a simulations that the increasing quality of the learning process can be achieved by doing authentication security during the task process, so the possibility of plagiarism is minimal [10].

Sprajc (2017) in this study states that plagiarism is currently using technology that facilitates copy-paste and is transferred with students with low motivation to study well. Such a lacking motivation is not only due to plagiarism, but can be the product of a poor teaching method. Another research finds that students who spend time on the internet has no relationships to plagiarism scores [11]. Technology is a tool for facilitating goals, but plagiarism is more about self-awareness. Therefore technology should be the driving force for students to learn, where they can find sources of knowledge such as from the internet.

Kashian (2015) discusses that plagiarism is an ethical act carried out by students. This happens even though the teacher has provided advice and guidance to not perform such an action. But students, mostly, kept doing it. The results of this study indicate that students with awareness about plagiarism have a low plagiarism score. A teacher can help students to avoid plagiarism by using Turnitin application to check the results of his work. If there is found plagiarism, students are expected to act accordingly, for instance rephrasing the statements in his own words [12]. In some countries this method is not necessarily applicable, due to application licenses and fees. Some developing countries built their own plagiarism detection system and fully supported by the government.

Ba (2016) conducts research on 681 student articles at one of the universities in Vietnam using Turnitin and found that the level of plagiarism in that campus was higher than outside Vietnam with a percentage of 29.06%. Plagiarism will lead to low academic value. In contrast, it will produce a positive effect when done honestly, although this task can take some times. So, it can be concluded that involving a plagiarism detection tool is very helpful in developing campus policies related to this [13]. In addition, the study did not explain that the acquired plagiarism included self-plagiarism or not. This is because when students writing articles, they sometimes copy-paste statements from another section of the same article.

C. *Students Motivation*

Gianna (2017) provides an interesting picture that the teachers' motivation is actually contagious to their students. On the other hand, a teachers who is lacking teaching motivation can still improve students achievement [14]. This finding should enforce schools to encourage teachers teaching motivations. As such is beneficial to the school and the people involved.

Samir (2014) said that teachers need to understand the students' motivation when teaching online and it is very difficult to get because the interaction is not face-to-face. Teachers need to use another strategies including demographic variables and the use of technology to prepare the 21st century generation [15].

Darwis (2016) study conducted on EFL fall students during the 2015/2016 semester shows that, plagiarism conducted by students was caused by the overly-difficult task material and students desire to graduate soon. Another cause is the openness of information stream where students can easily retrieve contents from the internet [16]. One could prohibit the use of internet-connected devices in class. However, it is all driven by the student's motivation in learning and his self-awareness.

Harits (2018) proposed approach in developing a SEG game, named Chem Dungeon, as a case study in order to demonstrate the effectiveness of media. This research making a good contribution to motivate the student via game learning [17].

Rocher (2018) reflects cognitive factor as the important aspect for a teacher to control his students, which including student involvement in learning, assignments and others. Plagiarism is a concern that students are not aware of. The paper emphasize on in class learning strategy where the teaching learning process becomes active in motivating students to avoid plagiarism [18]. However, the author fails to affirms that the easiest step is to announce plagiarism via a banner.

Johnson (2018) finds that plagiarism is a complex problem in a human state neither he is conscious or not, It is compulsory to seriously prevent plagiarism in order to maintain academic integrity on campus rather than penalizing plagiarists [19]. Johnson however, does not assert the prevention properly and the method to motivate students to get their hands off plagiarism.

D. PageRank

Amjad (2018) models an academic object ranking, including articles, journals, and conference proceedings. This ranking is based on the method of link analysis based on publication, number of citations, author's position, influence of the co-author, and topics in the scientific work. Each parameter influences rank scores to some degrees [20]. Link analysis is the basis of the page rank algorithm that uses the markov matrix. The weight of each citation is calculated based on the count and the list of references.

Massucci (2019) use pagerank algorithms to study scientific articles in a campus. Researchers believe that increasing a rank scores can be done by increasing academic ranking scores, such as via conducting internal or external citations between campuses, and the geographical location of the campus [21]. However, internal citation is susceptible to self-plagiarism that can drop the campus rank. A better alternative is conducting campus-to-campus citations and collaborates in the research community.

Gao (2016) explain that Pagerank index is the reflection of the popularity of publications, in the form of scientific works based on the article collection database. The author states that pagerank index is more suitable to use and is the main prediction parallel to h-index [22]. The h-index value is also influenced by the number of articles that Google recognize.

Hu (2015) applies pagerank algorithm to determine the referee in a match. It uses sensitivity analysis and general analysis methods in football matches, basketball, and baseball in determining coaches improvement analysis method for better result [23]. In addition such a method should fit for finding the best trainer for Olympics athletes or the likes.

Nykl (2015) aims to rank authors based on journal impact value using PageRank algorithm sourced from citation tracking. The result of this calculation is used to reward the author for the former year successful inclusion in the database. This algorithm solely use pagerank and the number of publications by eliminating self-citation [24].

Pagerank, which was developed by Google, has been used in various fields such as biology, computer science, physics, ecology, chemistry, sports and medicine to find ranks according to the research objectives. In addition, understanding pagerank charts makes it easy to visually understand the target set of interest [25]. PageRank algorithm selection is based on previous research which gives the fact that this algorithm is suitable for searching ranking based on direction or contribution from other nodes.

Let A is adjacent matrix for graph with vertex $V = \{1, \dots., n\}$ and the graph is directed graph which $A_{i,j}$ is probability value of weight (i, j). Suppose that x is Eigen vector, λ is Eigen value related to x, then power method denote as :

$$Ax = \lambda x \qquad (1)$$

Iterate that formula until convergence.

III. Methodology

In this research, we apply several procedures elaborated in the following section.

In general, out method refers to causal research where the problem has been clearly defined in the beginning. The main problem is how to calculate student ranking based on similarity score using power methods accurately.

Based on Fig. 1, there are several steps to follow (1) preparing dataset, (2) conducting primary data, (3) data analysis, (4) running power method and (5) ranks report. The following subsections elaborate each step in details.

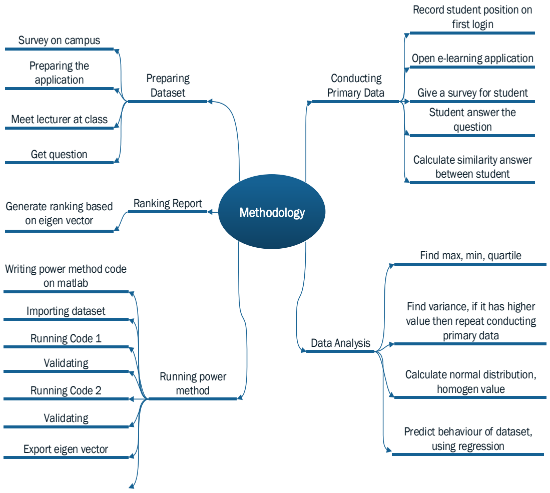

Methodology
Preparing Dataset
Survey on campus
Preparing the application
Meet lecturer at class
Get question
Conducting Primary Data
Record student position on first login
Open e-learning application
Give a survey for student
Student answer the question
Calculate similarity answer between student
Data Analysis
Find max, min, quartile
Find variance, if it has higher value then repeat conducting primary data
Calculate normal distribution, homogen value
Predict behaviour of dataset, using regression
Running power method
Writing power method code on matlab
Importing dataset
Running Code 1
Validating
Running Code 2
Validating
Export eigen vector
Ranking Report
Generate ranking based on eigen vector

Fig. 1. Research Methodology (Clockwise Direction).

A. Dataset Preparation

Dataset is the key factor of research, without a valid dataset the research is useless. There are two types of datasets, primary data and secondary data. Secondary data is to use existing data, while primary data is to collect data directly as conducted by this research. Preparing for primary dataset takes several steps.

1) Preparing the application: The e-learning application at UM campus has been developed since 2016. It focused on developing essays and the plagiarism detection application for used in student assignment assessment. Researchers collaborate with laboratory assistants in the Department of Electrical Engineering to prepare the applications prior to research. The application outputs students responses, similarity scores between students, student locations coordinates while doing the assesment (GPS latitude and longitude), response time for each answer, the timestamp when an answer is stored in the database table and the timestamp the response are submitted to the system.

2) Meet the teacher: The database course used as the experimental sample consisted of three classes, namely classes A, B, C, and D for each class containing approximately 40 students. Lecturer in this course, Ms Sophie gave permission to conduct research in his class on January 21, 2019. The lecturer said that students work on assignments online from their respective locations and some students work together in a certain location.

3) Course question: Questions are arranged in accordance with the material and syllabus in the course. The questions are as follows:

a) According to your understanding, how do you implement Hadoop on an e-commerce site? Describe your answers in detail starting from requirements engineering to reporting.

b) On the Hadoop platform there is a term map reduce that is used as a resource router on the Hadoop platform. Describe in details how the algorithm works.

c) Describe in details according to your understanding of the application of a relational database in the banking system where a good level of security and indexing is needed.

The questions take part in the e-learning database and will be displayed to students through an online learning system.

B. Conducting Primary Data

Primary data retrieval is carried out by researchers collaborating with lecturers at the UM campus. Some steps for taking primary data are as follows:

1) Record student position: Firstly, When students log in to e-learning system, they are required to enter a username and password, when they enter data correctly then the system will track the GPS coordinates in real, as well as the time login into the system.

2) Application start: After students have successfully logged into the system, then students take the exam accordingly with the schedule. Students work in two hours and they can do it anywhere.

3) Student answer the question: When students responds to exam questions, the system automatically save student answers every 5 minute, then, this answer will be stored in the server with including student answers, answering duration, and submission timestamp.

4) Student participate a survey: After student completed the exam, the students complete some items in a survey. This questionnaire contains questions related to this research including similarity answer, ranks and learning outcomes.

5) Similarity computation: Student answers in the database are then calculated using a formula developed by the researcher to product a similarity score.It has a unit of percentage from 0-100 and each answer has a similarity score except the uncomplated one. These results are then stored in the database for students to find out their similarity scores.

C. Data Analysis

We apply data analysis technique to measure the accuracy of a result. If the results of the analysis are reliable, the algorithm used is said to be efficient. Otherwise if the results of the analysis are not good, then the primary data retrieval is repeated.

1) Simple statistic requirement: The simplest analysis in looking at plagiarism scores is to use the minimum value, mode, mean, maximum value. The mean is used to determine the average plagiarism scores.

2) Homogen and variance: Ensuring homogeneous distribution is useful to see how the data is against other data, this is also aided by variance analysis that assesses each data to the average distance of all data. If the variance is small, then the data is spread evenly.

3) Normal distribution: Normal distribution is used to see the distribution of data at 5% beginning and end. Data that is normally distributed has a curve shape like a bell.

4) Regression analysis: The final part of this data analysis stage is knowing the behavior of the data using regression. The type of regression used depends on the characteristics of the data which will determine the trend of data based on other variables.

D. Running the Pagerank

The pagerank algorithm developed by Google adopted the power method formula by iterating the matrix Markov to converge. The calculation results are the Eigen vector and Eigen value values.

1) Writing Matlab code: Writing the Matlab code for power methods is actually not too complex, from Fig. 2 can be seen that there are some things that need to be considered so that the code matches the power method algorithm.

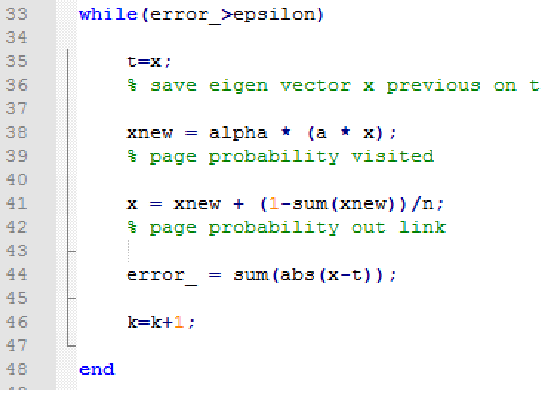
2) Importing dataset: In implementing this power method code, the input is the matrix markov where the data is taken from the student plagiarism score. as Fig. 3 describes direction factor of markov. The import dataset can be done by typing the command: A = loadMatrix;

3) Running code: After the dataset is uploaded into Matlab memory, the next step is to run the Matlab code. The following example on Fig. 4 runs the Matlab code.

4) Validating: After the code is run, the next step is to validate the results of the calculation of the power method in the form of Eigen values and Eigen vectors. Validation is done using SPSS.

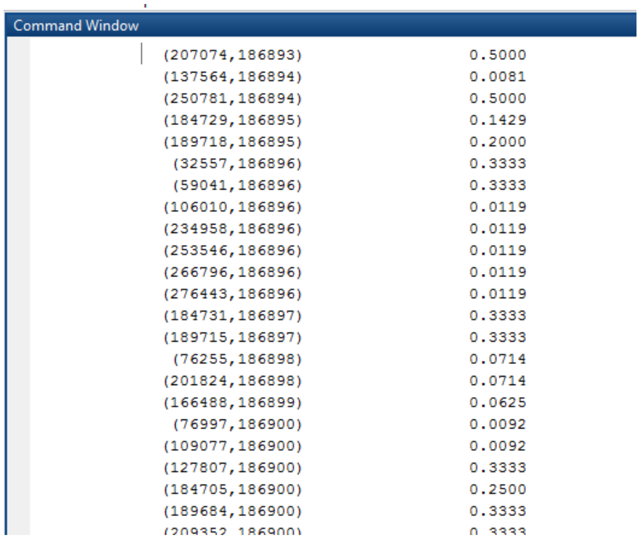
5) Export Eigen vector: In ranking calculation, the calculation results in Fig. 5 from the form of Eigen vectors stored in txt extension files. The example file is as follows.

6) Export Eigen value: The output of this calculation other than the Eigen vector also produces an Eigen value. This Eigen value is symbolized by λ. As an example, $\lambda =$ 0.0000438.



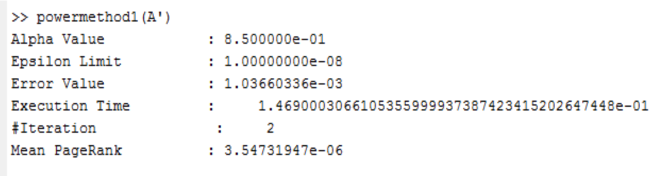
```
while(error_>epsilon)

    t=x;
    % save eigen vector x previous on t

    xnew = alpha * (a * x);
    % page probability visited

    x = xnew + (1-sum(xnew))/n;
    % page probability out link

    error_ = sum(abs(x-t));

    k=k+1;

end
```

Fig. 2. Power Method Code on Matlab.



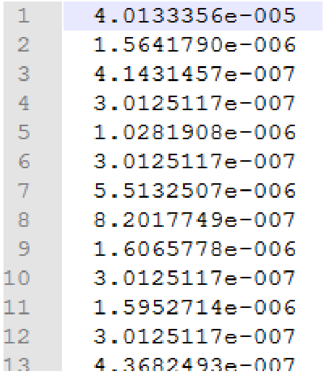
```
   (207074,186893)             0.5000
   (137564,186894)             0.0081
   (250781,186894)             0.5000
   (184729,186895)             0.1429
   (189718,186895)             0.2000
    (32557,186896)             0.3333
    (59041,186896)             0.3333
   (106010,186896)             0.0119
   (234958,186896)             0.0119
   (253546,186896)             0.0119
   (266796,186896)             0.0119
   (276443,186896)             0.0119
   (184731,186897)             0.3333
   (189715,186897)             0.3333
    (76255,186898)             0.0714
   (201824,186898)             0.0714
   (166488,186899)             0.0625
    (76997,186900)             0.0092
   (109077,186900)             0.0092
   (127807,186900)             0.3333
   (184705,186900)             0.2500
   (189684,186900)             0.3333
   (209352,186900)             0.3333
```

Fig. 3. Example of Process on Importing Dataset into Matlab Memory.



```
>> powermethod1(A')
Alpha Value              : 8.500000e-01
Epsilon Limit            : 1.00000000e-08
Error Value              : 1.03660336e-03
Execution Time           :     1.4690003066105355999937387423415202647448e-01
#Iteration                :     2
Mean PageRank            : 3.54731947e-06
```

Fig. 4. Result of Power Method Calculation.



```
4.0133356e-005
1.5641790e-006
4.1431457e-007
3.0125117e-007
1.0281908e-006
3.0125117e-007
5.5132507e-006
8.2017749e-007
1.6065778e-006
3.0125117e-007
1.5952714e-006
3.0125117e-007
4.3682493e-007
```

Fig. 5. Example of Eigen Vector.

E. Ranking Report

After the Eigen vector has been generated, then the next is the association based on the order of the student plagiarism score, the way to add it is in the database by importing the Eigen vector to the corresponding table. After that the results can be displayed in the e-learning application.

IV. Result and Discussion

This research provides several research results which will be discussed in this chapter. In addition to the results of the study, it is also discussed about the analysis of each outcome.

A. Primary Data Acquisition

Retrieval of primary data through several steps that begin from the preparation of the application, enter the question, retrieve the answer, until the results of the plagiarism score.

B. Importing Question into Database

In this study, the lecturer made a question for the test database subject as Fig. 6, while the questions were as follows.

a) According to your understanding, how do you implement Hadoop to e-commerce site? Describe your answers in detail starting from requirements engineering to reporting.

b) On the Hadoop platform there is the term map reduce that is used as a resource router on the Hadoop platform. Describe in detail how the algorithm works.

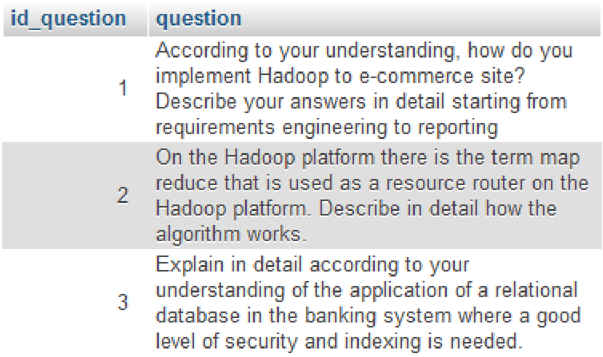
c) Explain in detail according to your understanding of the application of a relational database in the banking system where a good level of security and indexing is needed.

These questions are entered in the database in the question table.

After the questions are entered in the database, the questions can be used by the e-learning system. The online test is conducted on January 14, 2019 with a duration of 2 hours.

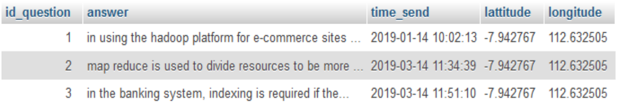
C. Student Answer Result

The online examination process has been carried out and runs smoothly, after students work on the exam questions then the next is to see the complete answers in the database even though students are not allowed to send blank answers, but need to be checked manually.



id_question	question
1	According to your understanding, how do you implement Hadoop to e-commerce site? Describe your answers in detail starting from requirements engineering to reporting
2	On the Hadoop platform there is the term map reduce that is used as a resource router on the Hadoop platform. Describe in detail how the algorithm works.
3	Explain in detail according to your understanding of the application of a relational database in the banking system where a good level of security and indexing is needed.

Fig. 6. Question Already Inserted in Database.



id_question	answer	time_send	lattitude	longitude
1	in using the hadoop platform for e-commerce sites ...	2019-01-14 10:02:13	-7.942767	112.632505
2	map reduce is used to divide resources to be more ...	2019-03-14 11:34:39	-7.942767	112.632505
3	in the banking system, indexing is required if the...	2019-03-14 11:51:10	-7.942767	112.632505

Fig. 7. Student Response Captured.

In Fig. 7, it can be seen that the system records the time of sending the answer, latitude and longitude, besides it also stores student information, subjects and other attributes.

D. Execute Plagiarism Detection

When the test time has been completed, the student cannot work again and the system will be closed, the answer at that time will be automatically saved in the database. Plagiarism calculations are then carried out automatically by the system, so researchers wait until the calculation is complete because this calculation runs cron jobs at midnight to save server load. The results of the calculation appear in Fig. 8.

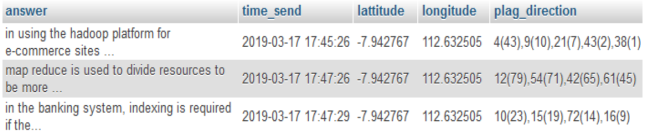
At Fig. 8 it appears that plagiarism is stored in the plag_direction field as example 4 (43), 9 (10) means that the student has the same plagiarism as the student with id 4 of 43% and has plagiarism with students id 9 with a plagiarism score of 10%.

E. Transforming to Directed Graph

When plagiarism has been obtained along with the direction of the sensitivity, then the next step is to change to graph form which facilitates analysis. The calculated similarity direction is the exit direction; the similarity of the dangling node type is ignored. Based on the number of students of 120 people, the number of graphs produced is 120 graphs. In Fig. 9, is given a brief explanation of the calculation results in the direction of graph.

F. Generating Markov Matrix

After all graphs are depicted, the next step is to change into the markov matrix. This markov matrix will then be processed using the power method algorithm. How to get matrix markov elements are as follows in Fig. 10.



answer	time_send	lattitude	longitude	plag_direction
in using the hadoop platform for e-commerce sites ...	2019-03-17 17:45:26	-7.942767	112.632505	4(43),9(10),21(7),43(2),38(1)
map reduce is used to divide resources to be more ...	2019-03-17 17:47:26	-7.942767	112.632505	12(79),54(71),42(65),61(45)
in the banking system, indexing is required if the...	2019-03-17 17:47:29	-7.942767	112.632505	10(23),15(19),72(14),16(9)

Fig. 8. Plagiarism Detection Result on Right Table.

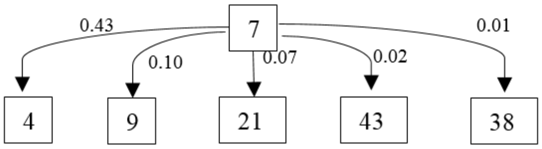


Fig. 9. Directed Graph.

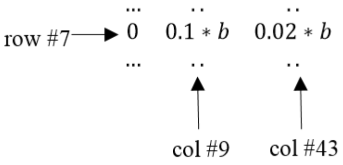


Fig. 10. Probability if Each Node.

From node 7 to node 4, it has a plagiarism score level of 0.43, while node 7 has similarities in answers with nodes 4, 9, 21, 43, 38. The number of inlinks to node 7 is 16 nodes, then the probability value in the markov matrix element is as with the value *b* is the weight of node 7 for the inlink and outlink probabilities.

G. Running Power Method

This part is the part that determines the Eigen value and vector Eigen, before running the power method first write the code in Matlab. The code can be seen in Fig. 11.

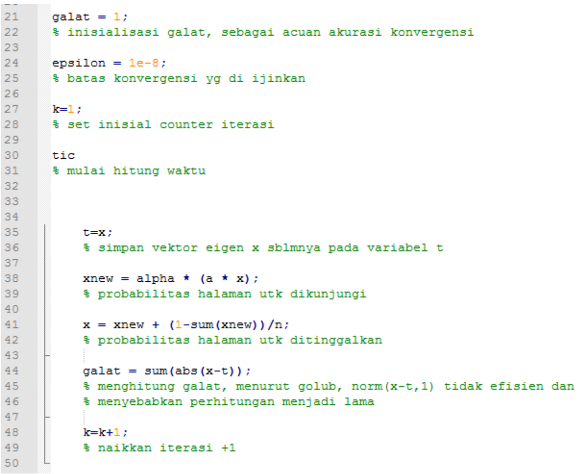
H. Eigen Vector

After the calculation using power method is complete, Eigen values and Eigen vectors will be produced as show on Fig. 12. Eigen vectors are *nx1*-size vectors where each row is the ranking value of each student.

I. Student Ranking

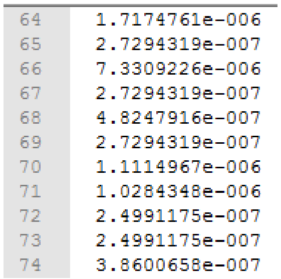
After the pagerank calculation is complete and the Eigen vector is generated, the next step is to save the data into the database in the remaining answer table as Fig. 13. The elements in the Eigen vectors will be stored and then sorted by the ranking of the Eigen vectors that have been stored.

The greater the Eigen value, the greater the similarity level so that the ranking displayed is a sequence on the Eigen vector element which is reversed from low to high.



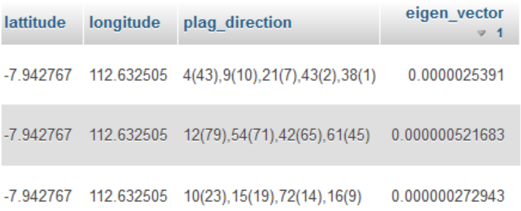
```
galat = 1;
% inisialisasi galat, sebagai acuan akurasi konvergensi

epsilon = 1e-8;
% batas konvergensi yg di ijinkan

k=1;
% set inisial counter iterasi

tic
% mulai hitung waktu



    t=x;
    % simpan vektor eigen x sblmnya pada variabel t

    xnew = alpha * (a * x);
    % probabilitas halaman utk dikunjungi

    x = xnew + (1-sum(xnew))/n;
    % probabilitas halaman utk ditinggalkan

    galat = sum(abs(x-t));
    % menghitung galat, menurut golub, norm(x-t,1) tidak efisien dan
    % menyebabkan perhitungan menjadi lama

    k=k+1;
    % naikkan iterasi +1

```

Fig. 11. Power Method Code on Matlab.



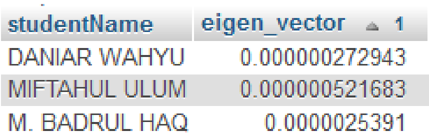
64	1.7174761e-006
65	2.7294319e-007
66	7.3309226e-006
67	2.7294319e-007
68	4.8247916e-007
69	2.7294319e-007
70	1.1114967e-006
71	1.0284348e-006
72	2.4991175e-007
73	2.4991175e-007
74	3.8600658e-007

Fig. 12. Eigen Vector as Result of Power Method Computation.



lattitude	longitude	plag_direction	eigen_vector 1
-7.942767	112.632505	4(43),9(10),21(7),43(2),38(1)	0.0000025391
-7.942767	112.632505	12(79),54(71),42(65),61(45)	0.000000521683
-7.942767	112.632505	10(23),15(19),72(14),16(9)	0.000000272943

Fig. 13. Ranking Shows on the Right Hand.



studentName	eigen_vector 1
DANIAR WAHYU	0.000000272943
MIFTAHUL ULUM	0.000000521683
M. BADRUL HAQ	0.0000025391

Fig. 14. Student Ranking Result.

In Fig. 14, it can be seen that students with the name DANIAR WAHYU have the first rank in the lowest plagiarism score, meaning that the student has a slight similarity, and is based on a rank value of 2.7e-7. These students were attended by MIFTAHUL ULUM and M. BADRUL HAQ. Students who need to be motivated by the teacher are M. BADRUL HAQ, the teacher can provide motivation or together with other friends (peer motivation).

V. Conclusion

Student ranking is important for teachers to know students who need motivation to learn well. The ranking results are used by teachers in supporting learning but teachers often find it difficult to make rankings based on student work results. Teachers often give online examinations through e-learning sites and assess one by one the test results and then make a ranking.

PageRank algorithm that is used by Google gives the best ranking on Google search engines by providing document citations as a determinant of ranking. PageRank algorithm with a power method base is applied in ranking students based on similarity answers to the essay exam. Similar answers are a subset of other answers that are more complex and ranking calculations are done using Matlab.

The results of the calculations carried out provide vector Eigen results as the basis for determining ranking. This research provides a very useful implication that teachers are facilitated in determining groups of students with the same level of learning motivation so that teachers can provide appropriate advice. Teachers feel the ease in providing motivation to students because the teacher will definitely know which students will get more attention to motivate learning so that their learning outcomes increase.

Further research can be done by speeding up the iteration time on the power method, so the calculation time of iteration until convergent will be efficient.

Acknowledgment

I would like to thanks to Universiti Teknikal Malaysia Melaka, Universitas Negeri Malang for providing Laboratory assistance for PageRank computing.

References

[1] R. Vieira, M. L.-H. E. Studies, and undefined 2015, “Academic Ranking--From Its Genesis to Its International Expansion.,” ERIC.

[2] M. M. Keller, K. Neumann, and H. E. Fischer, “The impact of physics teachers’ pedagogical content knowledge and motivation on students’ achievement and interest,” J. Res. Sci. Teach., vol. 54, no. 5, pp. 586–614, May 2017.

[3] P. Buckley, E. D.-I. L. Environments, and undefined 2016, “Gamification and student motivation,” Taylor Fr.

[4] A. Marks, M. Al-Ali, M. Majdalawieh, and A. Bani-Hani, “Improving academic decision-making through course evaluation technology,” Int. J. Emerg. Technol. Learn., 2017.

[5] J. Denning, R. Murphy, and F. Weinhardt, “Class Rank and Long-Run Outcomes,” 2018.

[6] R. Murphy and F. Weinhardt, “Top of the class: The importance of ordinal rank,” 2018.

[7] G. Hu, J. L.-E. & Behavior, and undefined 2015, “Chinese university students’ perceptions of plagiarism,” Taylor Fr.

[8] T. Cronan, J. Mullins, D. D.-J. of B. Ethics, and undefined 2018, “Further understanding factors that explain freshman business students’ academic integrity intention and behavior: Plagiarism and sharing homework,” Springer.

[9] J. Ehrich, S. J. Howard, C. Mu, and S. Bokosmaty, “A comparison of Chinese and Australian university students’ attitudes towards plagiarism,” Stud. High. Educ., vol. 41, no. 2, pp. 231–246, Feb. 2016.

[10] Y. Khlifi and H. A. El-Sabagh, “A novel authentication scheme for E-assessments based on student behavior over E-learning platform,” Int. J. Emerg. Technol. Learn., 2017.

[11] P. Šprajc, M. Urh, J. Jerebic, D. Trivan, E. J.- Organizacija, and undefined 2017, “Reasons for plagiarism in higher education,” degruyter.com.

[12] N. Kashian, S. Cruz, J. Jang, K. S.-J. of A. Ethics, and undefined 2015, “Evaluation of an instructional activity to reduce plagiarism in the communication classroom,” Springer.

[13] K. Do Ba et al., “Student plagiarism in higher education in Vietnam: an empirical study,” High. Educ. Res. Dev., vol. 36, no. 5, pp. 934–946, Jul. 2017.

[14] B. Gianna, R. Claudio, S. P.-P. economica, and undefined 2017, “Teacher Motivation and Student Learning,” ideas.repec.org.

[15] M. Samir Abou El-Seoud, I. A. T. F. Taj-Eddin, N. Seddiek, M. M. El-Khouly, and A. Nosseir, "E-learning and students' motivation: A research study on the effect of e-learning on higher education," Int. J. Emerg. Technol. Learn., 2014.

[16] S. Al Darwish, A. S.-I. E. Studies, and undefined 2016, "Reasons for College Students to Plagiarize in EFL Writing: Students' Motivation to Pass.," ERIC.

[17] H. A. Rosyid, M. Palmerlee, and K. Chen, "Deploying learning materials to game content for serious education game development: A case study," Entertain. Comput., vol. 26, pp. 1–9, May 2018.

[18] A. R. du Rocher, "Active learning strategies and academic self-efficacy relate to both attentional control and attitudes towards plagiarism," Act. Learn. High. Educ., p. 146978741876551, Mar. 2018.

[19] E. Johnson, "Situational Cheating Assessment of Motivation (SCAM): A Model for Understanding Student Plagiarism," 2018.

[20] T. Amjad, A. Daud, and N. R. Aljohani, "Ranking authors in academic social networks: a survey," Libr. Hi Tech, vol. 36, no. 1, pp. 97–128, Mar. 2018.

[21] F. Massucci, D. D.-J. of Informetrics, and U. 2019, "Measuring the academic reputation through citation networks via PageRank," Elsevier.

[22] C. Gao, Z. Wang, X. Li, Z. Zhang, W. Z.-P. one, and undefined 2016, "PR-index: using the h-index and PageRank for determining true impact," journals.plos.org.

[23] Z.-H. Hu, J.-X. Zhou, M.-J. Zhang, and Y. Zhao, "Methods for ranking college sports coaches based on data envelopment analysis and PageRank," Expert Syst., vol. 32, no. 6, pp. 652–673, Dec. 2015.

[24] M. Nykl, M. Campr, K. J.-J. of Informetrics, and U. 2015, "Author ranking based on personalized PageRank," Elsevier.

[25] D. G.-S. Review and undefined 2015, "PageRank beyond the Web," SIAM.

Optimal Control and Design of Electrical Machines

Wissem BEKIR[1]
Research Laboratory Smart Electricity and ICT
SEICT, LR18ES44
National Engineering School of Carthage
Université de Carthage
Tunis, Tunisia
Univ. Lille, Arts et Metiers
ParisTech, Centrale Lille, HEI, EA
2697, L2EP, F-59000
Lille, France

Lilia EL AMRAOUI[2]
Research Laboratory Smart Electricity and ICT
SEICT, LR18ES44
National Engineering School of Carthage
Université de Carthage, Tunis, Tunisia

Frédéric GILLON[3]
Univ. Lille, Arts et Metiers
ParisTech, Centrale Lille, HEI, EA
2697, L2EP, F-59000, Lille, France

***Abstract*—This paper presents a global optimization approach aiming to improve the energy efficiency of electrical machines. The process is made on a hybrid stepper motor allowing to simultaneously optimize design and command. This approach is axed around Pontryagin's maximum principle, which is applied to a magnetodynamic model based on permeances network model. The originality of the proposed approach is to obtain in the same process, the minimization of the energy by optimal control and the minimization of the energy by optimal sizing.**

Keywords—Optimal control; optimal sizing; Pontryagin's maximum principle; permeances network; hybrid stepper motor; energetic efficiency

I. Introduction

Currently, improving the energy efficiency of electric machines is a subject of high interest. Indeed electrical machines are widely used in industries, transportation and home applications. Thus, electric machines consume the largest amount of energy in the world (i.e. 46% of global consumption resulting in about 6040 megatonnes of CO2) [1,2].

The energy efficiency problem is studied in two different areas that require different skills. Firstly, automaticians deal with this problem as an optimal control issue. They seek to find the optimal control that allows minimizing either the energy consumption, subject to some constraints on the control, and/or the performances. Among the performance constraints, we can mention the constraints on the torque or for displacement problems, the constraints of positioning and speed [3]. Thus, the optimal control theory, which is a part of applied mathematics and automatic, is used in dynamic operation of the machine to find the trajectory of the command. However, the machine models used in the automatic field are coarse models [4-7] that do not take into consideration the geometric design parameters or not fully the magnetic phenomena.

Secondly, design specialists tackle the energy efficiency problem as an optimal design issue. Therefore, they seek to find the optimal design to achieve the required performance while minimizing also energy consumption. Thus, the models used are more complex and take into account with more accuracy the magnetic phenomena [8], the nature of the materials and the design parameters [9]. However, the models are quite complicated, the number of design parameters may be high and the search for optimality is carried out through optimization algorithms. Nevertheless, it would be absurd to seek for the optimal design of an electric machine without studying his command. Therefore, the command is one more parameter that is added to the optimization problems [10]. To overcome this difficulty the machine is optimized for operating points and imposing a form of control [11]. For example, by imposing a sinusoidal current control for a given machine, the problem is to find the optimum amplitude and angle of the control to have a certain torque and speed for a given operating point. The solution of this type of problem requires significant computation time, because of the extended model and the optimization algorithms used. Therefore, the difficulty is major if it is to find the optimal value of the command at any time, and in this case, the calculation time will be dissuasive.

In this paper, a method is proposed to solve this difficult problem. The idea is to merge the two domains and to develop a global optimization approach for design and control by applying optimal control theory [12-16] and nonlinear optimization algorithms on magnetodynamic models of an electric machine. This work is applied on a hybrid stepper motor [17-22], to prove the feasibility of the approach on a realistic case.

The paper is organized as follow; in the first part, a magnetic model of the machine is developed based on a permeances network. This model is coupled thereafter to a dynamic model that describes the electrical and mechanical behavior of the motor. In the second part, an optimal control theory is implemented, based on Pontryagin's maximum principle to this coupled magnetodynamic model. At first, an optimal control problem is posed. Then the optimality conditions are exploited to conclude on a Hamiltonian model that presents a two point boundary value problem. After that, a numerical method is proposed to solve this problem. Finally, optimal controls are calculated for a positioning problem. Results are then compared with a classical control. In the third part, a global optimization approach is proposed. The link between the different models and the resolution loops are presented. Then a global optimization problem is proposed followed by a study on the design parameter influence on the energy consumption of the machine. Finally, the global optimization problem is solved and the results are discussed.

II. Modélization

The study is applied to a two-phase hybrid stepper motor, with 1024 steps per revolution illustrated in Fig. 1. This motor is composed of two rings; each one has 50-tooth. The two rings are angularly offset by a tooth half step. They are interconnected by a permanent magnet. The stator has 8 plots each having Z_s teeth. The motor flux distribution is three-dimensional.

A. Magnetic Model

The magnetic model is based on a permeances network method, which consists in decomposing the magnetic device into a set of flux sources and passive elements. Fig. 2 shows a front view and a rear view of the MPPH. $\mathcal{P}_\alpha$, $\mathcal{P}_\beta$, $\mathcal{P}_\alpha'$, and $\mathcal{P}_\beta$' represents the permeances between the different stator plots and the two rotor rings. As the structure is symmetrical, this model focuses on the half of the machine.

Fig. 3 shows the equivalent magnetic circuit. The reluctance of iron is assumed infinite and the magnet is modeled by an ideal flux source Fm. The phases are modeled by the flux sources F_α and F_β.

The magnetic circuit resolution aims to determine the flux flowing in the branches. Equations are performed using Kirchhoff's laws; the fluxes generated are multiplied by 2 to represent the entire machine and by the number of coil, N_s, to describe the flux seen by the coils. The analytical expressions of the phases flux fed by currents I_α and I_β generated are given by:

$$\Phi_\alpha = 2\left(\mathcal{P}_\alpha + \mathcal{P}_\alpha'\right) N_s^2 I_\alpha + \left(\mathcal{P}_\alpha - \mathcal{P}_\alpha'\right) N_s F_m \tag{1}$$

$$\Phi_\beta = 2\left(\mathcal{P}_\beta + \mathcal{P}_\beta'\right) N_s^2 I_\beta + \left(\mathcal{P}_\beta - \mathcal{P}_\beta'\right) N_s F_m \tag{2}$$

The inductance L_α (resp. L_β) of the phases α (resp. β) is then deduced:

$$L_\alpha = 2\left(\mathcal{P}_\alpha + \mathcal{P}_\alpha'\right) N_s^2 \tag{3}$$

$$L_\beta = 2\left(\mathcal{P}_\beta + \mathcal{P}_\beta'\right) N_s^2 \tag{4}$$

As well as the mutual flux between the phases and the magnet which are expressed by:

$$\varphi_\alpha = \left(\mathcal{P}_\alpha - \mathcal{P}_\alpha'\right) N_s F_m \tag{5}$$

$$\varphi_\beta = \left(\mathcal{P}_\beta - \mathcal{P}_\beta'\right) N_s F_m \tag{6}$$

The magnetomotive force expression of a magnet as a function of the length of the magnet l_m and the coercive field H_c is given by:

$$F_m = l_m H_c \tag{7}$$

An analytic method is used to calculate air gap permenances in order to have a fairly fast model. The method aims to represent flux lines by tubes formed with straight lines and arcs and to calculate the permeance of each flux tube as a function of displacement. Calculations are performed for a tooth step, the displacement is assumed linear since the length of the angular displacement of a step is negligible in front of the rotor radius. Fig. 4 describes the tooth structure and the approximation of the flux tubes for a given position. In this figure, g represents the gap length, t_w the width of a tooth, l_r the ring length, P_i the permeances of flux tubes and x is the linear displacement.

A linear displacement $x=2t_w$ is equivalent to a rotation of a mechanical angle $\theta=7.2°$. In [22] the permeance expressions are available and the results of this method have been validated by finite element method.

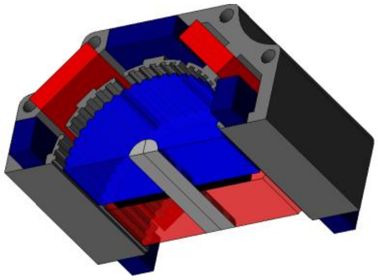


Fig. 1. Structure of the Hybrid Stepper Motor.

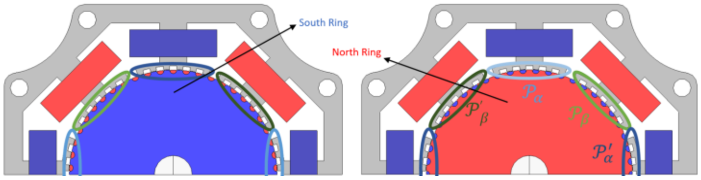


Fig. 2. HSM, Front and Rear View.

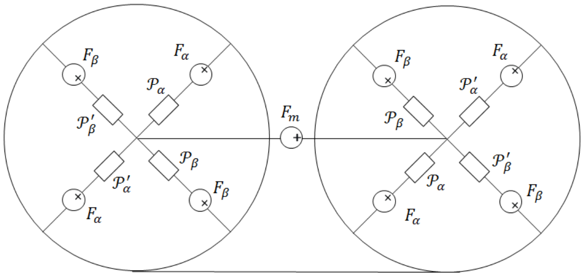


Fig. 3. Equivalent Magnetic Circuit.

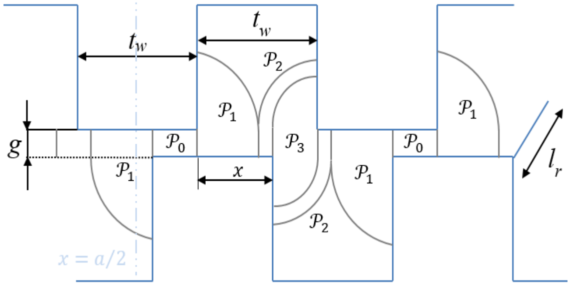


Fig. 4. Tooth Structure and the Approximation of the Flux Tubes.

B. Dynamic Model

The equations of the voltages induced by the phases are expressed by:

$$V_\alpha(t) = RI_\alpha(t) + I_\alpha(t)\frac{dL_\alpha(\theta)}{dt} + L_\alpha(\theta)\frac{dI_\alpha(t)}{dt} + \frac{d\varphi_\alpha(\theta)}{dt} \tag{8}$$

$$V_\beta(t) = RI_\beta(t) + I_\beta(t)\frac{dL_\beta(\theta)}{dt} + L_\beta(\theta)\frac{dI_\beta(t)}{dt} + \frac{d\varphi_\beta(\theta)}{dt} \tag{9}$$

where R is the resistance of the phases.

The fundamental principle of dynamics gives the following mechanical equation of motion (10):

$$J_r\frac{\partial^2\theta(t)}{\partial t^2} = C_{em}(t) - C_r(t) \tag{10}$$

With:

J_r : The rotor inertia,

C_{em} : electromagnetic torque,

C_r : resistant couple.

The inertia J_r is bounded to the ring length l_r and the magnet length l_m as expressed in equation (11).

$$J_r = 2\rho\pi r_r^4 l_r + \frac{1}{2}\rho_m\pi r_m^4 \tag{11}$$

With:

ρ : Density of motor iron,

ρ_m : density of neodymium magnet,

r_r : rotor radius,

r_m : magnet radius.

The fundamental principle of energy conservation allows to find the expression of the power absorbed and to deduce the expression of the electromagnetic torque (12):

$$C_{em}(t) = \frac{1}{2}I_\alpha^2\frac{dL_\alpha}{d\theta} + \frac{1}{2}I_\beta^2\frac{dL_\beta}{d\theta} + I_\alpha\frac{d\varphi_\alpha}{d\theta} + I_\beta\frac{d\varphi_\beta}{d\theta} \tag{12}$$

Equations (8), (9), (10) and (12) allow us to write the following state model (13):

$$\begin{pmatrix} \frac{dI_\alpha}{dt} \\ \frac{dI_\beta}{dt} \\ \frac{d\Omega}{dt} \\ \frac{d\theta}{dt} \end{pmatrix} = \begin{pmatrix} -\frac{1}{L_\alpha}\left(R + \frac{\partial L_\alpha}{\partial\theta}\Omega\right) - \frac{1}{L_\alpha}\frac{\partial\varphi_\alpha}{\partial\theta}\Omega \\ -\frac{1}{L_\beta}\left(R + \frac{\partial L_\beta}{\partial\theta}\Omega\right) - \frac{1}{L_\beta}\frac{\partial\varphi_\beta}{\partial\theta}\Omega \\ \frac{1}{J_r}\left(\frac{1}{2}I_\alpha^2\frac{\partial L_\alpha}{\partial\theta} + \frac{1}{2}I_\beta^2\frac{\partial L_\beta}{\partial\theta} + I_\alpha\frac{\partial\varphi_\alpha}{\partial\theta} + I_\beta\frac{\partial\varphi_\beta}{\partial\theta} - C_r\right) \\ \Omega \end{pmatrix} + \begin{pmatrix} \frac{1}{L_\alpha} & 0 \\ 0 & \frac{1}{L_\beta} \\ 0 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} V_\alpha \\ V_\beta \end{pmatrix} \tag{13}$$

with Ω is the rotation speed.

C. Coupled Model

The coupling is carried out as follows: first the air gap permeances are calculated, then the magnetic circuit is solved to find the vectors of the variations of the inductances and flux as a function of the mechanical angle. These vectors are then injected into the solver of the dynamic model. The inductances and flux values for each instant are then generated with an interpolation on the vectors considering the value of the instantaneous mechanical angle. Fig. 5 describes the magnetodynamic coupling.

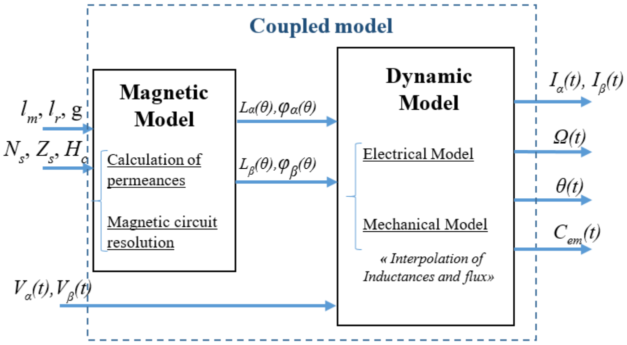

Coupled model
Magnetic Model
Calculation of permeances
Magnetic circuit resolution
Dynamic Model
Electrical Model
Mechanical Model
« Interpolation of Inductances and flux»
l_m, l_r, g
N_s, Z_s, H_a
$V_\alpha(t)$,$V_\beta(t)$
$L\alpha(\theta)$,$\varphi\alpha(\theta)$
$L_\beta(\theta)$,$\varphi_\beta(\theta)$
$I_\alpha(t)$, $I_\beta(t)$
$\Omega(t)$
$\theta(t)$
$C_{em}(t)$

Fig. 5. Magnetodynamic Coupling.

III. Optimal Control

The motor is now in load, driving a wheel having an inertia noted J_{roue} and a viscous friction coefficient, k_{roue}. The main purpose is to turn the wheel (initially in a position θ_0 at time t_0) to reach the position θ_f at a speed Ω_f and in a time t_f with a minimum energy. The problem could be formulated as following:

$$\begin{gathered} \min_{V_\alpha(t),V_\beta(t)} Obj(t) = \int_{t_0}^{t_f}\left(|V_\alpha| + |V_\beta|\right)dt \\ With: \\ V_{\alpha\min} \le V_\alpha \le V_{\alpha\max} \\ V_{\beta\min} \le V_\beta \le V_{\beta\max} \\ \theta(t_0) = \theta_0 \quad , \quad \theta(t_f) = \theta_f \\ \Omega(t_0) = \Omega_0 \quad , \quad \Omega(t_f) = \Omega_f \\ I_\alpha(t_0) = I_{\alpha 0} \quad , \quad I_\alpha(t_f) \in \square \\ I_\beta(t_0) = I_{\beta 0} \quad , \quad I_\beta(t_f) \in \square \end{gathered} \tag{14}$$

In equation (14), Obj represents the objective function to be minimized. This problem can be solved using the Pontryagin's maximum principle.

A. Hamiltonien and Costate Vector

The Hamiltonian of the system is given by:

$$\begin{aligned} H = {} & \left(|V_\alpha| + |V_\beta|\right) - \psi_1\frac{1}{L_\alpha}\left(RI_\alpha + I_\alpha\frac{\partial L_\alpha}{\partial\theta}\Omega + \frac{\partial\varphi_\alpha}{\partial\theta}\Omega - V_\alpha\right) \\ & -\psi_2\frac{1}{L_\beta}\left(RI_\beta + I_\beta\frac{\partial L_\beta}{\partial\theta}\Omega + \frac{\partial\varphi_\beta}{\partial\theta}\Omega - V_\beta\right) \\ & +\psi_3\frac{1}{J}\left(\frac{1}{2}I_\alpha^2\frac{\partial L_\alpha}{\partial\theta} + \frac{1}{2}I_\beta^2\frac{\partial L_\beta}{\partial\theta} + I_\alpha\frac{\partial\varphi_\alpha}{\partial\theta} + I_\beta\frac{\partial\varphi_\beta}{\partial\theta} - k\Omega\right) \\ & +\psi_4\Omega \end{aligned} \tag{15}$$

With ψ_1, ψ_2, ψ_3 and ψ_4 are costate variables, J is the sum of the rotor and wheel inertia, and k is the sum of the viscous friction coefficient of the rotor and the wheel.

According to the maximum principle, the costate vector must verify the following relation:

$$\begin{pmatrix} \dfrac{d\psi_1}{dt} \\ \dfrac{d\psi_2}{dt} \\ \dfrac{d\psi_3}{dt} \\ \dfrac{d\psi_4}{dt} \end{pmatrix} = \begin{pmatrix} -\dfrac{\partial H}{\partial I_\alpha} \\ -\dfrac{\partial H}{\partial I_\beta} \\ -\dfrac{\partial H}{\partial \Omega} \\ -\dfrac{\partial H}{\partial \theta} \end{pmatrix} \tag{16}$$

These relations give the first optimality condition.

B. Optimal Control Expresssion

The second optimality condition of the maximum principle indicates that the optimal control minimizes the Hamiltonian. Therefore, to find the command expression the sign of the functions corresponding to each command and derived from H has been studied. For example, the function derived from H with respect to V_a is given by:

$$H_{V\alpha} = \frac{V_\alpha}{|V_\alpha|} + \psi_1 \frac{1}{L_\alpha} \tag{17}$$

For $\psi_1(1/L_\alpha)>0$, $H_{V\alpha}$ is strictly positive, then $H(V_\alpha)$ is strictly increasing and the minimum of the Hamiltonian is reached in $V_\alpha=V_{min}$. Thus, studying the sign of the derivative allowed to obtain optimal command expressions. For $V_{\alpha min}=-V_{\alpha max}$ we get the expressions of the optimal controls V_α^* and V_β^* are:

$$V_\alpha^* = \begin{cases} V_{\alpha \min} sign\left(\psi_1 \dfrac{1}{L_\alpha}\right) & if \ \left|\psi_1 \dfrac{1}{L_\alpha}\right| > 1 \\ 0 & if \ \left|\psi_1 \dfrac{1}{L_\alpha}\right| \le 1 \end{cases} \tag{18}$$

$$V_\beta^* = \begin{cases} V_{\beta \min} sign\left(\psi_2 \dfrac{1}{L_\beta}\right) & if \ \left|\psi_2 \dfrac{1}{L_\beta}\right| > 1 \\ 0 & if \ \left|\psi_2 \dfrac{1}{L_\beta}\right| \le 1 \end{cases} \tag{19}$$

The command with a minimum of energy for this problem involving constraints on the commands is therefore Bang-off-bang type. In (18) and (19) the command is expressed in terms of the costate variables ψ_1, ψ_2 which in turn are expressed as a function of all the parameters of the machine according to the relation (16). The control is expressed explicitly according to all the parameters taken into account by the model.

The equations (16), (18), (19) and the machine state allow (13) us to obtain the Hamiltonian model (20):

$$\begin{cases}
\dfrac{d\psi_1}{dt} = \psi_1 \dfrac{1}{L_\alpha}\left(R + \dfrac{\partial L_\alpha}{\partial\theta}\Omega\right) - \psi_3 \dfrac{1}{J}\left(I_\alpha \dfrac{\partial L_\alpha}{\partial\theta} + \dfrac{\partial\varphi_\alpha}{\partial\theta}\right) \\
\dfrac{d\psi_2}{dt} = \psi_2 \dfrac{1}{L_\beta}\left(R + \dfrac{\partial L_\beta}{\partial\theta}\Omega\right) - \psi_3 \dfrac{1}{J}\left(I_\beta \dfrac{\partial L_\beta}{\partial\theta} + \dfrac{\partial\varphi_\beta}{\partial\theta}\right) \\
\dfrac{d\psi_3}{dt} = \psi_1 \dfrac{1}{L_\alpha}\left(I_\alpha \dfrac{\partial L_\alpha}{\partial\theta} + \dfrac{\partial\varphi_\alpha}{\partial\theta}\right) + \psi_2 \dfrac{1}{L_\beta}\left(I_\beta \dfrac{\partial L_\beta}{\partial\theta} + \dfrac{\partial\varphi_\beta}{\partial\theta}\right) + \psi_3 \dfrac{k}{J} - \psi_4 \\
\dfrac{d\psi_4}{dt} = \psi_1 \dfrac{1}{L_\alpha}\Omega\left(-I_\alpha \dfrac{1}{L_\alpha}\left(\dfrac{\partial L_\alpha}{\partial\theta}\right)^2 + I_\alpha \dfrac{\partial^2 L_\alpha}{\partial\theta^2} - \dfrac{1}{L_\alpha}\dfrac{\partial L_\alpha}{\partial\theta}\dfrac{\partial\varphi_\alpha}{\partial\theta} + \dfrac{\partial^2\varphi_\alpha}{\partial\theta^2}\right) \\
\qquad + \psi_1 \dfrac{1}{L_\alpha}\left(-RI_\alpha \dfrac{1}{L_\alpha}\dfrac{\partial L_\alpha}{\partial\theta} + V_\alpha^* \dfrac{1}{L_\alpha}\dfrac{\partial L_\alpha}{\partial\theta}\right) \\
\qquad + \psi_2 \dfrac{1}{L_\beta}\Omega\left(-I_\beta \dfrac{1}{L_\beta}\left(\dfrac{\partial L_\beta}{\partial\theta}\right)^2 + I_\beta \dfrac{\partial^2 L_\beta}{\partial\theta^2} - \dfrac{1}{L_\alpha}\dfrac{\partial L_\beta}{\partial\theta}\dfrac{\partial\varphi_\beta}{\partial\theta} + \dfrac{\partial^2\varphi_\beta}{\partial\theta^2}\right) \\
\qquad + \psi_2 \dfrac{1}{L_\beta}\left(-RI_\beta \dfrac{1}{L_\beta}\dfrac{\partial L_\beta}{\partial\theta} + V_\beta^* \dfrac{1}{L_\beta}\dfrac{\partial L_\beta}{\partial\theta}\right) \\
\qquad - \psi_3 \dfrac{1}{J}\left(\dfrac{1}{2}I_\alpha^2 \dfrac{\partial^2 L_\alpha}{\partial\theta^2} + \dfrac{1}{2}I_\beta^2 \dfrac{\partial^2 L_\beta}{\partial\theta^2} + I_\alpha \dfrac{\partial^2\varphi_\alpha}{\partial\theta^2} + I_\beta \dfrac{\partial^2\varphi_\beta}{\partial\theta^2}\right) \\
\dfrac{dI_\alpha}{dt} = -\dfrac{1}{L_\alpha} I_\alpha\left(R + \dfrac{\partial L_\alpha}{\partial\theta}\Omega\right) - \dfrac{1}{L_\alpha}\dfrac{\partial\varphi_\alpha}{\partial\theta}\Omega + \dfrac{1}{L_\alpha}V_\alpha^* \\
\dfrac{dI_\beta}{dt} = -\dfrac{1}{L_\beta} I_\beta\left(R + \dfrac{\partial L_\beta}{\partial\theta}\Omega\right) - \dfrac{1}{L_\beta}\dfrac{\partial\varphi_\beta}{\partial\theta}\Omega + \dfrac{1}{L_\beta}V_\beta^* \\
\dfrac{d\Omega}{dt} = \dfrac{1}{J}\left(\dfrac{1}{2}I_\alpha^2 \dfrac{\partial L_\alpha}{\partial\theta} + \dfrac{1}{2}I_\beta^2 \dfrac{\partial L_\beta}{\partial\theta} + I_\alpha \dfrac{\partial\varphi_\alpha}{\partial\theta} + I_\beta \dfrac{\partial\varphi_\beta}{\partial\theta} - k\Omega\right) \\
\dfrac{d\theta}{dt} = \Omega
\end{cases} \tag{20}$$

The Hamiltonian model obtained is a two point boundary value problem. Indeed, it is necessary to find the initial conditions of the costate variables allowing bringing the system from its initial state to the desired final state. As for the final state of the current, it was left free. This implies transversality conditions on the costate vector. In fact, if the final state is free, the corresponding costate vector must be equal to zero:

$$\begin{pmatrix} \psi_1(t_f) \\ \psi_2(t_f) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{21}$$

To solve this kind of problem of boundary condition type, the so-called shooting method has been used, which aims to create a function S that takes as inputs the initial conditions of the variables and returns the difference between the final state obtained and the desired final state and the transversality conditions.

$$S: \begin{pmatrix} \psi_1(t_0) \\ \psi_2(t_0) \\ \psi_3(t_0) \\ \psi_4(t_0) \end{pmatrix} \to \begin{pmatrix} \psi_1(t_f) \\ \psi_2(t_f) \\ \Omega(t_f) - \Omega_f \\ \theta(t_f) - \theta_f \end{pmatrix} \tag{22}$$

The root of the S function should be determined. This can be done with an algorithm based on the newton method, the *fsolve* routine of *Matlab*®. However, one must be able to estimate the initial conditions to converge towards the solution. In our case, initial conditions chosen at random will give a solution to a displacement problem that we note Pb_0 with an error at startup. Indeed, a differential equation solver, such as *ode15s* of *Matlab*®, tends to correct the initial condition error and to find a trajectory. Moreover, since the system is repetitive (step by step) the pace of the variation of the costate variables allows finding a good estimate of the initial conditions for Pb_0. Fig. 6 and Fig. 7 show the variation of the costate variables for randomly chosen conditions and the estimation of the initial conditions.

The shooting method is subsequently launched with the estimated initial conditions to get the solution of the problem Pb_0. From this solution, a dichotomy technique is applied to the most influential costate variables and allow us to find the solution to this displacement problem, in our case it is the variable $\psi_4(t_0)$. This solution is fast and well adapted to our study. The following problem has been considered (23).

$$\min_{V_\alpha, V_\beta} Obj(t) = \int_0^{1.2s} \left(|V_\alpha| + |V_\beta|\right) dt$$

With;

$$-5V \le V_\alpha \le +5V$$
$$-5V \le V_\beta \le +5V$$
$$\begin{aligned} &\theta(t_0) = 0\,rd &&, \quad \theta(t_f) = 0.5835rd \\ &\Omega(t_0) = 0rd/s &&, \quad \Omega(t_f) = 0rd/s \\ &I_\alpha(t_0) = 0A &&, \quad I_\alpha(t_f) \in \Box \\ &I_\beta(t_0) = 0A &&, \quad I_\beta(t_f) \in \Box \end{aligned} \tag{23}$$

The resolution of this problem takes about 10 minutes. Fig. 8 and Fig. 9 describe respectively the optimal commands V_α^* and V_β^* and the corresponding switching functions.

The pulse widths gradually decrease until reaching the end position. This behavior is due to the effect of inertia, which has a great impact on energy consumption. Considering thus inertia the motor consumes less and less energy to reach the final position. Table I gives the different values of the positive and negative pulse widths of the optimal control V_α^* of Fig. 8.

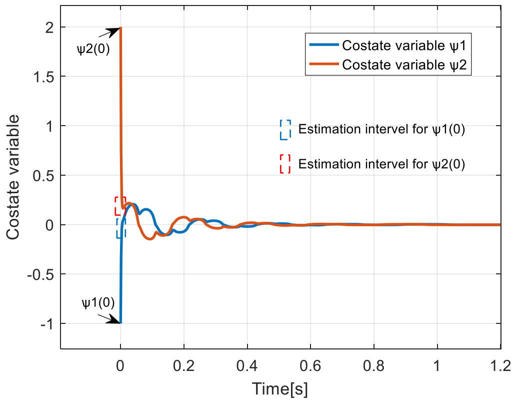


Fig. 6. Costate Variables ψ_1(t) and ψ_2(t) and Estimation of the Initial Conditions.

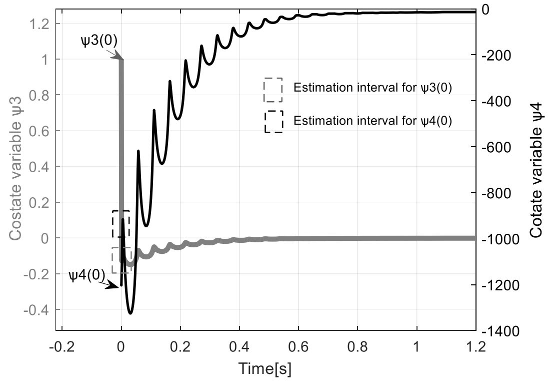


Fig. 7. Costate Variables ψ_3(t) and ψ_4(t) and Estimation of the Initial Conditions.

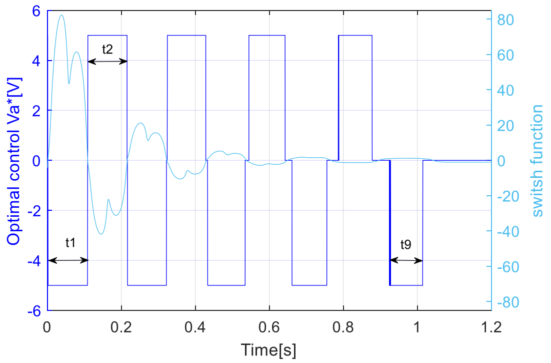


Fig. 8. Optimal Control V_α*(t).

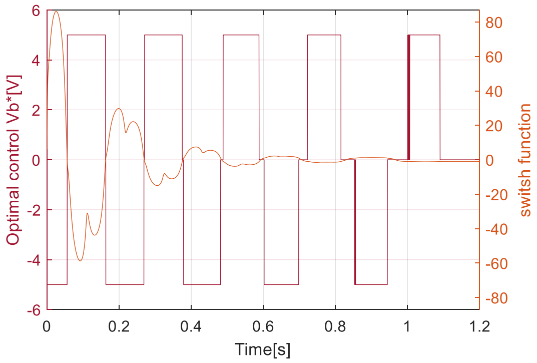


Fig. 9. Optimal Control V_β*(t).

TABLE. I. PULSE WIDTH

Pulse width	Value [s]
t1	0.1082
t2	0.1060
t3	0.1052
t4	0.1038
t5	0.1013
t6	0.0979
t7	0.0943
t8	0.0903
t9	0.0878

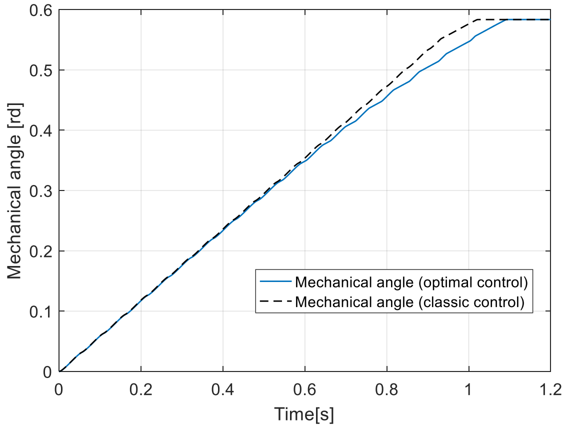


Fig. 10. Evolution of Mechanical Angle for Both Methods.

Fig. 10 shows the evolution of the mechanical angle obtained with the optimal control, in continuous line and a conventional control in dashed line. In the conventional control, the pulses have the same width, optimized to have a minimum pulse width.

The variation of the mechanical angle for the optimal control presents a curvature, which is due to the effect of the taking into account of the inertia. For the classic control, the position is reached more quickly but the motor in this case consume more energy. The obtained optimal control offers a gain of absorbed power of 5,6%.

Fig. 11 presents the evolution of the costate variables $\psi_3(t)$ and $\psi_4(t)$ related respectively to the speed $\Omega(t)$ and to the position θ (t). The costate variables $\psi_1(t)$ and $\psi_2(t)$ will have the same pace as the switch functions with a difference in amplitude due to the terms $1/L_\alpha$ and $1/L_\beta$ in (8) and (9).

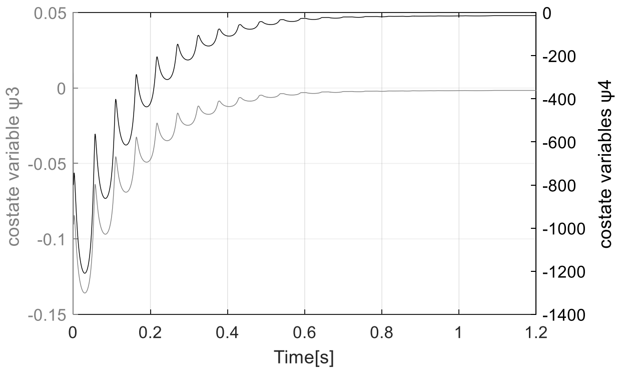


Fig. 11. Costate Variables ψ3(t) and ψ4(t).

IV. Global Optimization Approach

The application of the Pontryagin's maximum principle gives a formulation that allows the automatic generation of the optimal control through costate equations. This optimal control takes into account all input machine dynamic model parameters. Therefore, it is possible to vary the design inputs of the machine and find the optimal control for each geometric configuration. The idea is to add an optimization loop on design parameters. This will give an overall optimization approach to design and control.

A. Global Model and Optimisation Loop

Fig. 12 describes the interaction between the different models, the dichotomy approach and the optimization loop whenever both the rotor rings length and the magnet length should be opimized. The magnetic model is now coupled to the Hamiltonian model. The dichotomy method allows finding initial conditions of the costate model. Then an optimization loop is applied on the design parameters.

B. Effet De La Longeur De L'aimant Et Des Coronnes

In this section, the influence of the rings and magnet length on both the energetic performances machine and on the optimal control has been studied.

Fig. 13 describes a section of the rotor and the two design parameters that we study.

An unloaded machine has been simulated for a 5 steps displacement, i.e. $\theta(t_f)=0.157$[rd], and a final time $t_f=0.1s$. The objective function and the constraints on the controls, speed and currents have been kept as in the equation (23). However, we vary the length of the crown between 8 mm and 22mm, and the magnet length between 0.5mm and 1.6mm. An objective function denoted *Obj'* is defined such that:

$$Obj'(t) = \int_0^{0.1s} \left(V_\alpha^2(t) + V_\beta^2(t)\right) dt \tag{24}$$

Minimizing the objective function *Obj'* is the minimization of the *Obj* function. The only difference is that the *Obj'* function would have a smoother evolution and would be more efficient we use a gradient descent.

Fig. 14 describes the evolution of the function *Obj'* as a function of l_m and l_r.

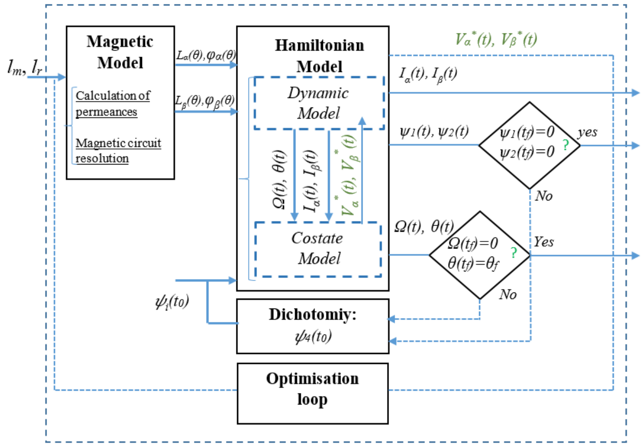


Fig. 12. Resolution Process.

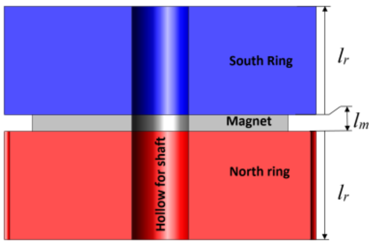


Fig. 13. Rotor Section.

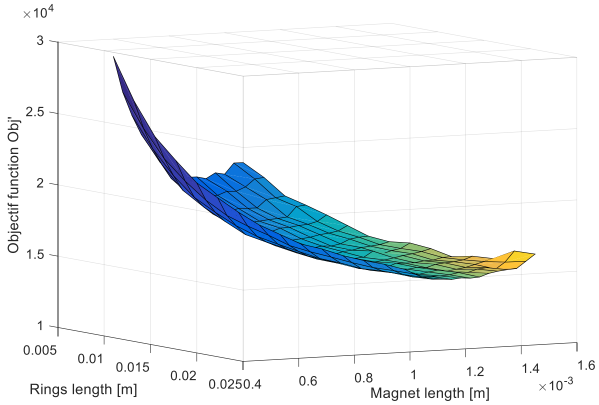


Fig. 14. Evolution of Objective Function Obj.

The evolution of the surface is inversely proportional to the magnet thickness. This is explained by the fact that a thicker magnet provides a larger magnetomotive force and thus the machine needs less energy to turn.

In the direction of the increase of the length of the crowns, the surface is decreasing then increasing this is explained by two phenomena. Taking the cases where l_m = 0.5mm, there exists a value of l_r that we note l_r^* for which the function *Obj'* is minimal. From this value if the value of l_r is decreased there will be less interaction between the rotor and the stator and therefore the motor will need more energy to turn. In the other hand, if from l_r^*, the value of l_r is increased, the inertia of the rotor increases and the motor will need more energy to turn. So for each value of the magnet thickness there exists a value of lm for which the *Obj'* function is minimal. This minimum presents a compromise between rotor-stator interaction and rotor inertia. Fig. 15 describes the evolution of optimal control V_β^* as a function of l_r Note that the pulse widths are minimal for the value l_r^* which explains the behavior of *Obj'*.

C. Global Optimisation Problem

The purpose here is to find the optimal values of l_r and l_m that minimize the consumption consideration that minimizing *Obj'* is minimizing *Obj*. The global optimization problem could be written as follows:

$$\min_{l_m,\, l_r}\left(\min_{V_\alpha(t),V_\beta(t)}\left(\int_0^{0.1\,s}\left(\left|V_\alpha\right|+\left|V_\beta\right|\right)dt\right)\right)$$

$$\begin{array}{l}
with: \\
\qquad 8mm \le l_r \le 22mm \\
\qquad 0.5mm \le l_m \le 1.6mm \\
\qquad -5V \le V_\alpha \le +5V \\
\qquad -5V \le V_\beta \le +5V \\
\theta(t_0) = 0\,rd \quad , \quad \theta(t_f) = 0.157rd \\
\Omega(t_0) = 0rd/s \quad , \quad \Omega(t_f) = 0rd/s \\
I_\alpha(t_0) = 0A \quad , \quad I_\alpha(t_f) \in □ \\
I_\beta(t_0) = 0\,A \quad , \quad I_\beta(t_f) \in □
\end{array} \qquad (25)$$

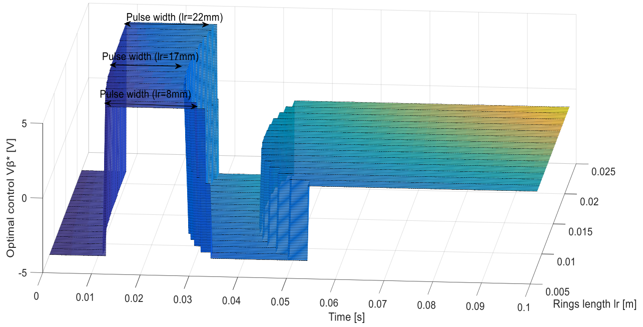


Fig. 15. Evolution of Optimal Control V_β*(t).

The optimization loop is launched on the design with the *fmincon* routine of *Matlab*®. Table II shows the initial and the optimal sizing.

Table III illustrates the objective function and the input power for both configurations.

The optimal sizing provides a gain of 34.6% for input power compared to the initial configuration. The obtained gain is significant mainly because of the magnet length that has increased.

TABLE. II. INITIAL AND OPTIMAL SIZING

	Initial sizing	**Optimal sizing**
Length of rings [mm]	13	11.7
Length of magnet [mm]	1	1.6

TABLE. III. OBJECTIVE FUNCTION AND INPUT POWER

	Objective function	**Input power [w]**
Initial sizing	13250	5.43
Optimal sizing	10650	3.55

V. CONCLUSION

The main contribution of this paper is to obtain in the same process, the minimization of the energy by optimal control and the minimization of the energy by optimal sizing. First, a magnetodynamic model based on a permeances network was developed. Then the Pontryagin Maximum Principle was applied to the magnetodynamic model in order to find the optimal control minimizing the energy. The application of the PMP allows us to explicitly express the command according to all model parameters and to have a Hamiltonian model that automatically generates optimal control. The study has shown that boundary value problem encountered at the resolution level of the Hamiltonian model can be solved by a simple dichotomy when it is a control problem of an electric machine. The results showed a gain of 5.7% compared to a conventional control for a given positioning problem. Finally, adding an optimization loop on design inputs to give an overall optimization approach.

References

[1] G. Lei, J. Zhu, and Y. Guo “Multydisciplinary Design Optimisation Methods for Electrical Machines and Drive Systems”, Springer-Verlag Berlin Heidelberg, 2016.

[2] D. Dorrell “A Review of the Methods for improving the Efficiency of Drive Motors to Meet IE4 Efficiency Standards”, Journal of power electronics, pp. 842-851, September 2014.

[3] H. Alihalli, M. Ilyas Bayindir, “Time-energy optimal control of vector controlled induction motor”, COMPEL: Int. J. Computation and Mathematics in Electrical and Electronic Engineering, pp. 235-251, 2002.

[4] V. P. Lapshin, I. A. Turkin and V. V. Khristoforova, "Synthesis of Electromechanical Position Control System by Means of Maximum Principle," 2018 International Russian Automation Conference (RusAutoCon), Sochi , pp. 1-4, 2018.

[5] C. M. Vega, J. R. Arribas and J. Herrero, "Optimal-time control of squirrel cage induction motors with constant load torque," IEEE 2002 28th Annual Conference of the Industrial Electronics Society. IECON 02, Sevilla, pp. 2039-2044 vol.3, 2002.

[6] R. Zheng, R. Cai and M. Li, "Energy management of electric vehicles with permanent magnet synchronous in-wheel motors using pontryagin's minimum principle," IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society, Beijing, pp. 2275-2280, 2017.

[7] A. S. Revko and R. D. Yershov, "Control Rapidity Optimization Technique of DC-Motor Driven by Quasi-Resonant Converter Using Pontryagin's Maximum Principle," 2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO*)*, Kievpp. 705-710, , 2018.

[8] L. Roubache, K. Boughrara, F. Dubas and R. Ibtiouen, ”New Subdomain Technique for Electromagnetic Performances Calculation in Radial-Flux Electrical Machines Considering Finite Soft-Magnetic Material Permeability”, in IEEE Transactions on Magnetics, vol. 54, no. 4, pp. 1-15, April 2018.

[9] Z. Djelloul-Khedda, K. Boughrara, F. Dubas and R. Ibtiouen, "Nonlinear Analytical Prediction of Magnetic Field and Electromagnetic Performances in Switched Reluctance Machines," in IEEE Transactions on Magnetics, vol. 53, no. 7, pp. 1-11, July 2017.

[10] M. H. Mohammadi, R. C. P. Silva and D. A. Lowther, "Finding Optimal Performance Indices of Synchronous AC Motors," in IEEE Transactions on Magnetics, vol. 53, no. 6, pp. 1-4, June 2017.

[11] M. H. Mohammadi, R. C. P. Silva, D. A. Lowther,”Incorporating Control Strategies Into the Optimization of Synchronous AC Machines: A comparison of Methodologies”, IEEE Transaction on magnetics, vol. 54, no. 3, March 2018.

[12] H. Geering, “Optimal Control with Engineering Application”, IEEE Control Systems, vol. 31, no. 5, pp. 115-117, 2011.

[13] Zhu, J. Trélat, E. & Cerf, M. Pac,” Geometric optimal control and applications to aerospace”, J. Math. Ind, 2198-4115, 2017.

[14] S. Uebel, N. Murgovski, C. Tempelhahn and B. Bäker, "Optimal Energy Management and Velocity Control of Hybrid Electric Vehicles," in IEEE Transactions on Vehicular Technology, vol. 67, no. 1, pp. 327-337, Jan. 2018.

[15] N. Kim, S. Cha and H. Peng, "Optimal Control of Hybrid Electric Vehicles Based on Pontryagin's Minimum Principle," in IEEE Transactions on Control Systems Technology, vol. 19, no. 5, pp. 1279-1287, Sept. 2011.

[16] H. Dagdougui, “Optimal Control of a Network of Power Microgrids Using the Pontryagin’s Minimum Principle”, IEEE Transactions on Control Systems Technology, vol. 22, no. 5, pp. 1942-1948, 2014.

[17] M. Matsuri, M. Nakamura, T. Kosaka, “Instantaneous torque analysis of hybrid stepping motor,” IEEE Transactions on Industry Applications, vol. 34, no°5, pp. 1176-1182, 1996.

[18] C. Kuert, M. Jufer and Y. Perriard, "New method for dynamic modeling of hybrid stepping motors," Conference Record of the 2002 IEEE Industry Applications Conference. 37th IAS Annual Meeting (Cat. No.02CH37344), Pittsburgh, PA, pp. 6-12 vol.1, USA, 2002.

[19] P. P. Acarnely, “Stepping Motors: A Guide to Theory and Practice 4th edition ”, IET Control Engineering Series 63, London, 2007.

[20] I. Ionica, M. Modereanu, A. Morega, C Boboc, “Design and modeling of a hybrid stepper motor”, IEEE International Symposium on Advanced Topics in Electrical Engineering, pp. 192-195, 2017.

[21] C. Stuebig, B. Ponick, “Comparision of Calculation Methods for Hybrid Stepping Motors”, IEEE Transaction on Industry Applications, Vol. 48, no°6, pp2182-2189, Nov-Dec 2012.

[22] W. Bêkir, L. E. Amraoui and F. Gillon, “Dynamic performances determination of a HSM using a finite element validation of the magnetostatic model”, 2016 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM), pp. 1-6, Marrakech, 2016.

The Model of Game-based Learning in Fire Safety for Preschool Children

Nur Atiqah Zaini[1], Siti Fadzilah Mat Noor[2], Tengku Siti Meriam Tengku Wook[3]
SOFTAM, Research Center for Software Technology and Management
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor
Malaysia

***Abstract*—The Model of Game-based Learning in Fire Safety developed for preschool children to educate them in learning fire safety issues. Due to the lack of awareness towards fire hazard, there are few factors that have arisen regarding this issue such as children's ages, experiences and knowledge. The main objective of this study is to identify the user requirements of preschool children in developing the Model of Game-Based Learning in Fire Safety. This study involved six preschool children of Tabika Kemas Kampung Berawan, Limbang Sarawak by using User-Centered Design method. The ability of cognitive, behavior and psychomotor skills are the main aspects to develop the model. Thus, to lower the risk of injuries during practical training in real situation, there is a need to educate them using the technology of tablet. Therefore, a prototype has been developed known as APi Game-Based Learning as a platform for children to learn about fire safety issues. Hence, this APi prototype developed to validate the Model of Game-Based Learning in Fire Safety development for preschool children. Thus, the finding of the study showed the engagement of children in learning fire safety through game improved their knowledge, behavior and psychomotor skills. Overall, this study makes an important contribution in determining the usability on the level of effectiveness towards preschool children through active learning.**

***Keywords*—Game-based learning; fire safety; user-centered design; effectiveness**

I. Introduction

Lack of awareness in potential fire hazards as the concern issues bring threat for the children. Addressing this need, the exposure of fire safety issues on the children led them to learn survival skills. In fact, fire hazard affects their lives, education and abilities. Many cases in Malaysia as reported by Fire Rescue Department of Malaysia showed the annual report of fire cases related with houses and buildings [1]. Due to the faulty of electricity as the highest factor of deadly fire occurred every year. Thus, by providing the information of fire hazard especially to children plays an important role to educate them and save lives as well. The importance of learning and teaching fire safety issues should be taken seriously to prevent injuries to the children.

These are the most common problems occurred that related with fire safety. There were caused by limited fire safety awareness, less training, fire protection systems were not provided and delayed on notifying the fire brigades [2]. Fires are always life-threatening and dangerous that risk people to save lives. Addressing these problems, a widespread implementation of teaching and learning using technology was focused to reduce injuries and fatalities among the people especially children.

In addition, there is a growing need for the preschool children to learn using technology instead of conveying the information through conventional way of teaching and learning [3]. Subsequently, technologies have been exposed all over the world and spread like wildfire by providing a lot of functionalities that contribute in helping people. Therefore, the development of interactive technology such as tablet led to continuous innovation of teaching and learning. Apparently, the technology itself stimulates creativity and often draws the attention of preschool children [4]. However, children are facing difficulties when using technology because of some constraints due to their abilities and growth development. Many applications developed to ease them in using technology and its use in education trained their cognitive and psychomotor skills [5], [6].

The implementation of applications that support learning helped the children to interact well with the systems. For example, the application of fire safety education in the form of gaming environment. Thus, the interactive learning such as game-based learning enhanced the children to improve interaction and learning in fire safety issues [7]. On the other hand, to provide them clearly on the risk of fire hazard that caused a lot of damages and death required an efficient way to identify their needs.

II. Literature Review

Game-based learning promoted interactive learning, which, derived from the use of computer games in delivering the educational value [8]. The main purpose of developing educational games and training simulators were using the same technical elements with different purpose of delivering information to the users. There were many applications developed in the form of gaming to educate the children in teaching and learning [9], [10], [11]. Realizing the best ways to attract the children to stay focus during learning session by providing multimedia elements such as animation and audio [12], [13], [14].

Getting feedback from the system while interacting, the children gave either positive or negative responses and kept on playing continuously. It showed that the engagement of the children by using their abilities to think and operate the technology themselves. Engagement obviously affect the ways of teaching and learning which, emphasizing the enjoyment and motivation of the users [15]. Meanwhile, the children who are actively participating in learning must be motivated to engage them with the use of sound [16]. It helped the learning environment to be more fun and interesting. To improve their interests in learning serious issue such as fire safety should be delivered through interactive learning to let them interact well with the system without feeling scared.

Therefore, fire safety is a growing concern issue among the youth especially children. They were highly-risk of exposure towards fire hazard and led into danger because of lack of knowledge and delay actions [17]. Thus, the fire safety education should be delivered in the form of gaming to lower the risk of injuries [18]. Serious games, virtual reality, mobile and computer games were used as a platform to promote the fire safety education. Hence, the learning session would be more entertaining at the same time delivering the main purpose of fire safety issues. It is crucial to provide high quality of education where they learned on the importance of evacuation time and basic skills to escape. Their responses towards the fire situations should be taken seriously. By providing an effective way to learn with the use of interactive devices absolutely attracts the children's interests in learning.

On the other hand, interactive devices demonstrated the children's creativity in solving problems. Due to their growth of psychomotor, age is the main element that will affect their performances in operating the devices [19]. However, the functionalities of the devices should be suitable for their ages and capabilities to use the technology. The fact that the children were still in growth development phase showed they were still having difficulties to control their hands and fingers movements precisely [20]. Focusing on the specific devices provided for children to convey the information will ease them to interact and give the positive responses with the system as well. In fact, the children's capabilities are also affecting their involvement and achievements [21]. The most concern issue was promoting fire safety education through interactive learning by using technology that suitable for the fine motor skills development of preschool children.

In this research, a technology of tablet supported learning where, it considered as one of the various ways that able to teach the children about fire safety education. As children nowadays have early contact with the tablet that led them to deal faster with the functionalities [22]. Thus, there were three main aspects focused that were cognitive, psychomotor and behaviour. These aspects were observed on preschool children to validate the effectiveness of fire safety education through the technology of tablet.

III. Method

Overall method used in this study was User-Centered Design (UCD) that focused on the involvement of the users from the early stage until the end of the experiment [23]. By using UCD method, four distinct phases involved the users throughout the design and development process. The following are the general phases of UCD process as shown in Fig. 1.

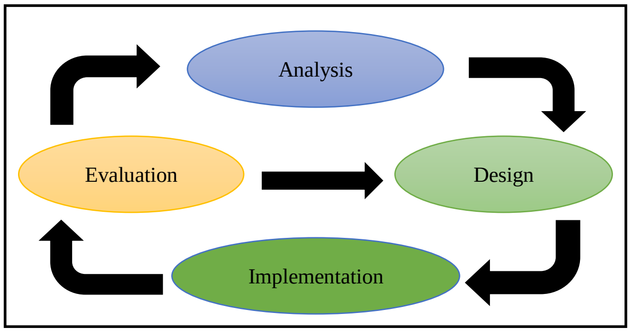


Fig. 1. User-Centered Design.

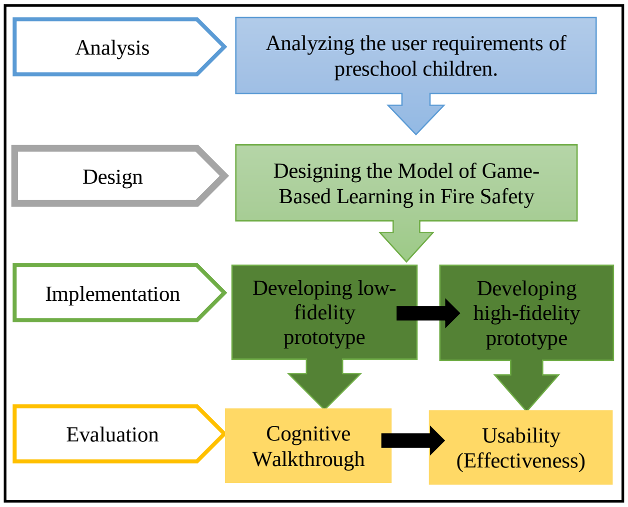

Analysis
Analyzing the user requirements of preschool children.
Design
Designing the Model of Game-Based Learning in Fire Safety
Implementation
Developing low-fidelity prototype
Developing high-fidelity prototype
Evaluation
Cognitive Walkthrough
Usability (Effectiveness)

Fig. 2. Processes of Phases.

Meanwhile, Fig. 2 showed the processes of every phase involved in this research.

By following the processes of UCD method, all the user requirements needed were used to develop the model of game-based learning in fire safety for preschool children. Knowing their limitations on using technologies should be focused to avoid difficulties during final testing conducted on them. Collecting the data was the crucial part to ensure that the end results achieved the objectives of the research. For the evaluation of cognitive walkthrough were tested on expert users through two techniques, observation and interviews. While, final evaluation of usability was tested on preschool children through observation and think aloud.

A. *Preliminary Study (Analysis Phase)*

The Model of Game-based Learning in Fire Safety was developed based on the user requirements obtained from the preliminary study conducted on the preschool children. By combining the results obtained from preliminary study and literature review, the model was developed specifically following the children's needs. This study involved six preschool children who were at the age of four to six years old from Tabika Kemas Kampung Berawan, Limbang Sarawak [24]. There were two types of existing fire safety games tested during the preliminary study, which, were shown in Table I. Based on the testing, some interaction styles issues in the existing games were analyzed to determine the research problems as well before conducting the experiment.

TABLE. I. THE PROBLEMS OF EXISTING FIRE SAFETY GAMES

Types of Games	Initial Investigation	Issues of Interaction (Users)
Help Mikey Make It Out	1. Limited user interaction. 2. Lack of information. 3. Use of language. 4. Less guidelines provided.	1. User only interacted with the limited button. 2. The interface was quiet confusing for the users. 3. Less instruction for the users on how to play. 4. Difficulty of understanding English.
Fire Safety Challenge	1. Lack of information 2. Element of sound or voice used. 3. Use of language.	1. No guidelines on how to play. 2. No instructions through voice provided. 3. Difficulty of understanding English.

The interaction styles issues were identified after the games tested on the preschool children. Following the experiment conducted, the data was collected through observation and think aloud. As shown in Table II, the children played the games with different skills level. It showed that different ages of children placed some constraints on the ability of cognitive, behaviour and psychomotor skills when being tested by the existing fire safety games.

TABLE. II. ANALYSIS OF PRELIMINARY STUDY

Age of Children	Help Mikey Make It Out	Fire Safety Challenge
4 years	1. Understand on how to play the game but needed more explanation because of the language used. 2. Response towards the game was slow because of the instructions given on how to play were not clear. 3. Able to point the answer using the buttons provided. 4. Good response on getting rewards.	1. Understand on how to play the games which showing positive response for getting correct answers. 2. Slow response because the interface was quiet confusing. 3. There were no proper guidelines to play where only music provided. 4. Difficult to point precisely at the objects.
5 years	1. Interested in visualisation and audio in the games. 2. Needed more explanations because of the language used was difficult to understand. 3. Good response on getting rewards.	1. Not really interested in playing because no voice instruction provided. 2. Able to point directly at the objects.
6 years	1. Able to complete the games after instruction given. 2. Slow response because of the language used in the game.	1. Enjoying the game when getting reward. 2. Able to control finger movements precisely. 3. Slow response because of no instructions provided.

TABLE. III. USER REQUIREMENTS

User Requirements	4 years	5 years	6 years
User Interaction	✓	✓	✓
Interface Design	✓	✓	✓
Psychomotor (Fine Motor)	✓	✓	✓
Cognitive (Knowledge)	✓	✓	✓
Behaviour	✓	✓	✓
Gaming Elements (Reward, Storyline, Player, Time)	✓	✓	✓
Multimedia Component (Animation and Audio	✓	✓	✓
Genre (Strategy)	✓	✓	✓
Malay Language	✓	✓	✓

Besides, observation technique defined on observing their capabilities of solving the games. Meanwhile, think aloud was a process of determining the children's opinions on engagement and enjoyment of playing the games. Thus, the data collected can be implemented to develop the Model of Game-Based Learning in Fire Safety as shown in Table II and Table III.

Based on Table III showed that the user requirements obtained after carrying out the experiment on the existing fire safety games. All the user requirements were focused on children's abilities, gaming elements and gaming factors [3], [6], [15], [19], [22], [25], [26].

Analyzing all the user requirements for developing the model such as user interaction, interface design, psychomotor skills, cognitive, behavior, gaming elements, multimedia components. However, during the experiment conducted on the preschool children, they were facing trouble to understand English because they were not using English as the main language at home and school. All of the preschool children participated well in the experiment which, showing the positive responses towards fire safety games. Some of them were facing difficulties on interacting with the buttons must be designed specifically for children to play with ease.

B. *Model of Game-based Learning In Fire Safety (Design Phase)*

The Model of Game-based Learning in Fire Safety was developed based on the combination of Game-Based Learning Model [22], Fire Safety Model [27] and the user requirements obtained from the preliminary study as shown in Fig. 3.

Game-based Learning Design Model:

1) User: Preschool children at the age of four to six years old are required to testify the effectiveness of fire safety game.

2) Device: Tablet is a tool used as the device to test the game which, showed the compatibility with the children's fine motor skills and cognitive.

3) APi Fire Safety Game: It consists of game elements such as rewards, player, storyline, feedback, time, multimedia components to improve motivation, enjoyment and interactivity. Besides, interface design is focusing on menu

Strategic Research Grant of KRA-2018-025, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia.

driven to guide the children in using fire safety game by providing animation and sound. The voiceover was using Malay language.

4) Prototype: It consists of low-fidelity prototype and high-fidelity prototype.

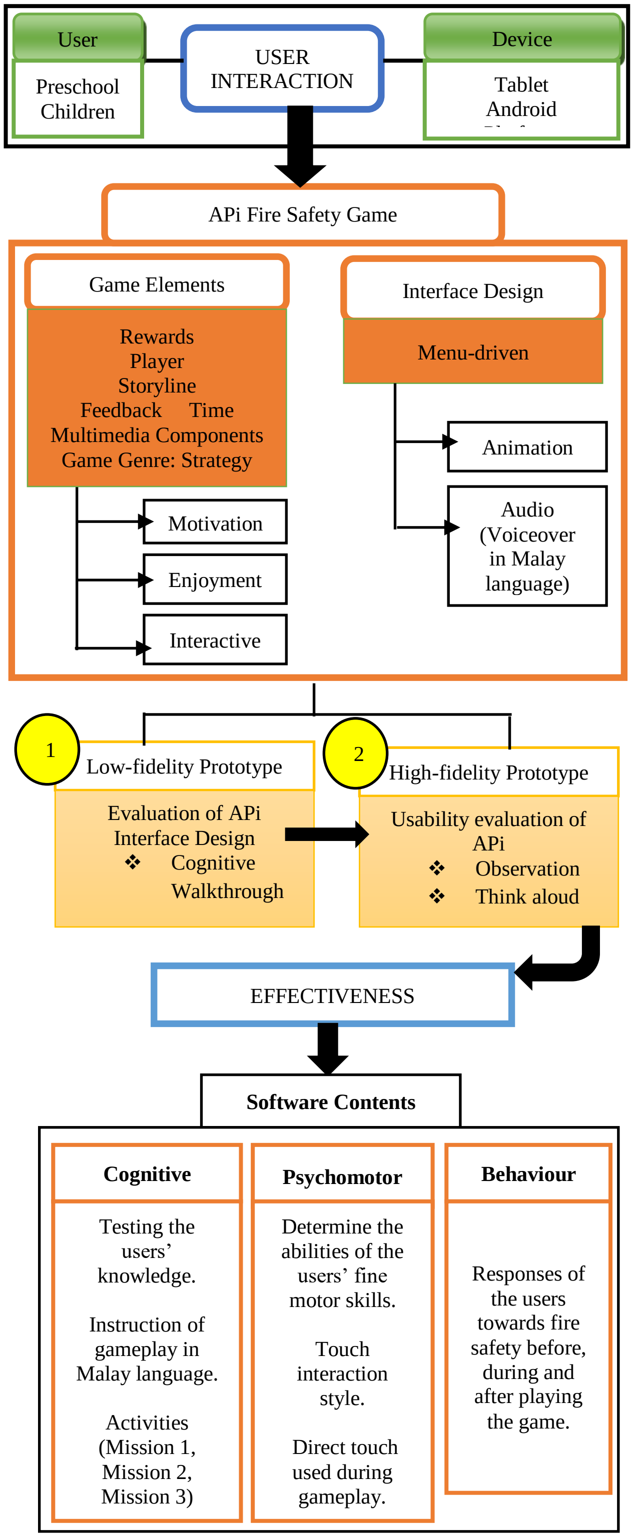


Fig. 3. Model of Game-Based Learning in Fire Safety for Preschool Children.

Software Contents: There are three aspects will be evaluated after testing the games that include behaviour, cognitive and psychomotor. Meanwhile, the effectiveness is the factor of usability used to validate the Model of Game-Based Learning in Fire Safety.

The model consisted of three missions that required the users to complete. Mission 1 needed the users to identify the causes of fire accidents such as inflammable substances while Mission 2 conveyed the information on how to use fire extinguisher and the way to escape from fire in a trap house. Meanwhile, Mission 3 required the users to identify whom they should rescue if fire accident happened. All these missions provided for the users to learn and know the importance of fire safety issues. Every missions tested on the users would affect their cognitive, behaviour and psychomotor skills. Thus, all the missions created based on the learning module of Tabika KEMAS.

All the game elements were obtained based on suitable elements needed for preschool children [8]. During preliminary study showed that audio and animation were the most important elements for them to engage while playing the existing fire safety games. These game elements improved the children's motivation, enjoyment and playing interactive game. This APi fire safety game was developed in 2D which categorized in strategy and offline game [26]. Besides, the interface design was based on menu-driven to ensure the users understand the game flow of playing.

In addition, every mission in this game was following the module of learning at Tabika Kemas Kampung Berawan, Limbang Sarawak. The missions developed were suitable for the preschool children in terms of the gameplay and easy to understand the contents. This was because the APi fire safety game was using Malay language as the medium of interaction between the users and game. All the instructions were given by using audio instead of text because the limitation of the preschool children in reading.

There were two types of prototypes developed based on the model. It consisted of low-fidelity prototype and high-fidelity prototype. In order to evaluate the low-fidelity of APi prototype, the cognitive walkthrough method was conducted on experts. This method used to improve the interface design of the prototype. While, to evaluate high-fidelity prototype were using observation and think aloud on the real users, the preschool children. This final evaluation used to evaluate the effectiveness of APi fire safety game tested on them.

The effectiveness of usability was evaluated based on cognitive, psychomotor and behavior of preschool children. Through the APi fire safety game, the children's abilities were tested in solving the missions with the use of tablet technology. The users' fine motor skills would be tested too to ensure that they could use touch interaction easily. With the use of APi fire safety game, their behaviors were evaluated before, during and after playing the game.

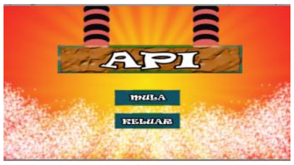
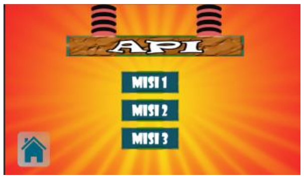
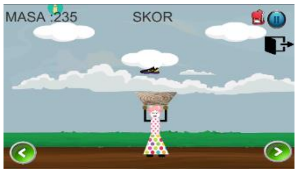
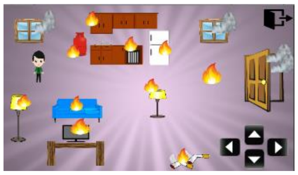
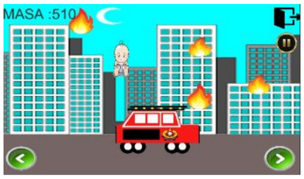
C. Developing APi Prototypes

In this section discussed on the prototype development which consisted of two processes, low-fidelity APi prototype and high-fidelity APi prototype. It was used to be named as

APi because of the fire theme that related with fire safety issues. These prototypes were developed based on the model shown on Fig. 2 for preschool children and suitable with their ages, skills and knowledge. Both prototypes were designed using Adobe Photoshop and Unity 2017. APi prototypes emphasized more on the interaction style issues that needed to solve the research problems. Table IV and V showed the processes on developing the APi prototypes.

A low-fidelity prototype was not a fully functioning system developed for the users. This process was showing the early stage of development to test the functionality and interface design that suitable for the real users. Table IV showed the interface design of low-fidelity APi prototype.

TABLE. IV. LOW-FIDELITY PROTOTYPE

APi INTERFACE DESIGN	**DESCRIPTION**
 APi Main Interface	The main APi interface which required the players to choose "MULA" or "KELUAR. MULA : Start the game KELUAR : Exit the game
 Missions' Screen	The player can choose the three missions which are Mission 1, Mission 2 and Mission 3.
 Mission 1: Gameplay	The instructions given to the players by using audio and animation on how to play. The players need to identify the flammable substances. The gameplay started by pressing the left and right button to catch the objects.
 Mission 2: Gameplay	The instructions given to the players by using audio on how to play. The players need to think on how to escape from the house with fire. There were four buttons provided for the players to move the character.
 Mission 3: Gameplay	The instructions given to the players by using audio on how to play. The players needed to identify whom they should rescue from fire. Pressing the right and left button to move the character in catching the objects targeted.

D. *Cognitive Walkthrough Method*

Extending to this low-fidelity APi prototype development, a Cognitive Walkthrough method was carried out to improve the weakness of the interface design [18], [28]. This process involved experts that consisted of two lecturers, two gamers and two graphic designers. Those experts needed to perform the tasks given to them in 20 minutes without discussing with other participants.

The experts were given the tasks to be completed during experiment along with the low-fidelity APi prototype. The tasks were shown below:

1) The participant needs to press the button "MULA" at the APi main interface to start the game.

2) There will be missions on the next screen which, consists of "MISI 1", "MISI 2" and "MISI 3". The participant needs to test the functionality of HOME button to go back to the main interface.

3) Next, the participant needs to choose "MISI 1" as the starting mission. The instruction will be given through audio before the mission has started. The participant needs to test the functionality of the button to proceed to the next screen.

4) The participant starts the "MISI 1" by pressing the "MULA" button and plays the game. Then, there are left and right button provided to move the character after the instructions given through audio.

5) After completing the "MISI 1", the participant can go to the mission screen by hitting the "OUT" button provided at the right side.

6) The participant will choose "MISI 2" for the next game. The instructions will be given through audio to guide the participant. There will be RIGHT and LEFT button provided to go to the next page.

7) The participant will be given instructions on how to play by using UP, DOWN, LEFT and RIGHT button provided.

8) After completing the "MISI 2", the participant can go to the mission screen by hitting the "OUT" button provided at the right side.

9) Next, the participant needs to choose "MISI 3" and the instruction will be given through audio before the mission has started. The participant needs to test the functionality of the button to proceed to the next screen.

*10)*The participant starts the "MISI 3" by pressing the "MULA" button and play the game. Then, there are RIGHT and LEFT button provided to move the character after the instructions given through audio.

*11)*After completing the "MISI 3", the participant can go to the mission screen by hitting the "OUT" button provided at the right side.

*12)*The participant will be rewarded by giving the badges after all the missions completed. Then, the participant needs to go back to the APi main screen by hitting the "OUT" button.

*13)*At the APi main screen, the participant can choose "KELUAR" to exit the game.

As shown in Table V was the analysis of Cognitive Walkthrough conducted on the expert users.

Through the Cognitive Walkthrough method tested on the experts, there were some interaction and interface designs needed to be improved in low-fidelity APi prototype. Thus, it involved with the controller buttons, font and size, animation of the buttons, background image, consistency of button's position and visibility of the buttons.

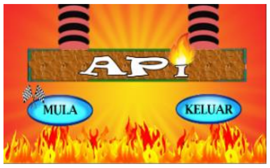
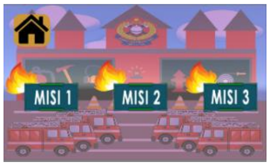
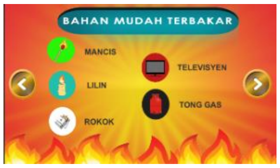
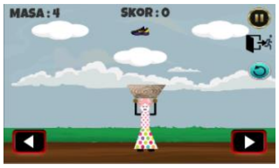
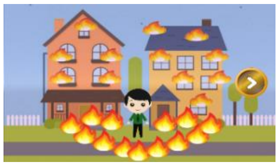
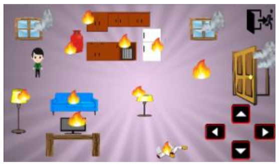
The analysis showed that APi prototype had the features that suitable for the preschool children. The missions provided for them can be easily played with minimal supervision where, the children did not need to read the instructions of the game due to their capabilities in reading. This was because every instruction given to them through the audio.

As shown in Table VI were the high fidelity of APi prototype after cognitive walkthrough method was conducted on six experts.

TABLE. V. ANALYSIS OF COGNITIVE WALKTHROUGH

CRITERIA	DESCRIPTION	AGREE	DISAGREE
Background Theme	1. The background colour. 2. The use of images as background.	50 50	50 50
Fonts	1. Font size 2. Font type	50 50	50 50
Buttons (*"MULA, KELUAR"*)	1. Size of button 2. Consistency of shape and size 3. Animation style of button	50 33 100	50 67 0
Controller Button	1. The use of button to move character 2. Consistency of shape and size. 3. Position of button	33 17 0	67 84 100
Icon	1. To deliver the information.	83	17
Character	1. Referring to the player 2. The use of colour 3. Size of characters	83 50 100	17 50 0
Game Goals	1. The way of playing and delivering information.	100	0
Menus	1. The position of menu 2. The use of colour 3. Animation style of button	100 83 100	0 17 0
Features	1. Missions 2. Reward 3. Time 4. Score	100 100 100 100	0 0 0 0

TABLE. VI. HIGH-FIDELITY PROTOTYPE

APi INTERFACE DESIGN	DESCRIPTION
 APi Main Interface	The main interface design added with the images of fire to let the users know that they are learning fire safety game.
 Mission 1: Gameplay	The background theme related with Fire and Rescue Department of Malaysia. The position of "MISI 1", "MISI 2", "MISI 3" changed along with the font type and size. MISI 1: Mission 1, MISI 2: Mission 2, MISI 3: Mission 3
 BAHAN MUDAH TERBAKAR MANCIS LILIN ROKOK TELEVISYEN TONG GAS Mission 2: Gameplay	The fire images added along with "Bahan Mudah Terbakar" to educate the users clearly. Besides, the position of Next and Previous buttons changed to avoid confusion among the users.
 Mission 3: Gameplay	The controller button changed on the design and position. For the design, the controller buttons were fixed to ensure the consistency in every mission. Meanwhile, there were differences between the positions of controller buttons and Next, Previous buttons to avoid confusion for the users.
 Mission 2: Instruction	The background theme changed based on the mission to escape from fire in house. The position of Next button was fixed too.
 Mission 2: Gameplay	The controller buttons added with UP and DOWN to control the player movements. The buttons were fixed to ensure the consistency in design and position.

IV. RESULTS

The development of the high-fidelity APi prototype was tested on the real users that were preschool children. At this evaluation of the APi prototype to evaluate the effectiveness of the Model Game-Based Learning in Fire Safety developed. The evaluation phase involved six participants of preschool children by using two techniques of observation and think aloud.

These techniques used to identify the reactions of the preschool children either they were able to complete all the missions in specific time given. The ways of responding to the game were the main part to evaluate the effectiveness of APi fire safety game tested on them.

As shown in Table VII, the results showed on the score obtained while playing the APi fire safety game.

Meanwhile, Fig. 4 showed the three main aspects investigated in this study, which were cognitive, psychomotor, and behaviour of the preschool children. The usability testing was conducted to validate the effectiveness of the fire safety game-based learning model for preschool children. The percentage of cognitive aspect showed only 50% of preschool children at the age of 4 could accomplish Mission 1 due to the ability of memorizing the inflammable substances was still at the low level. 100% obtained for both 5 and 6 years ol d preschool children where, they could solve the problems completely.

For psychomotor aspect showed that all of preschool children were able to control their finger movements while playing the game. They could fix the hand-eye coordination although 4 years old children were having difficulties in handling the speed of the objects. Due to the growth development of fine motor skills, the children were trying to play with the buttons provided.

Based on the percentage showed that 50% of the preschool children at the age of 4 could finish the missions. They were showing positive response and engagement during playing session. However, some of them was easily distracted due to the frustration of the incapable getting the objects correctly. It led the children to give up on playing the game. Other than that, preschool children at the age of 5 and 6 years old showed the engagement and motivation while playing the game. They kept on responding to seek for help when facing danger. Thus, they were able to understand the dangers of fire that led them to injuries. Their behaviours were observed before, during and after playing the game.

Therefore, Table VIII showed the analysis of the usability testing evaluation on preschool children at the age of 4 to 6 years old. Every mission aligns all different activities to test their capabilities on cognitive, psychomotor and behaviour of the children.

TABLE. VII. SCORES OF THE API FIRE SAFETY GAME

Mission	R1	R2	R3	R4	R5	R6
Mission 1	50	50	100	100	100	100
Mission 2	100	100	100	100	100	100
Mission 3	50	50	100	100	100	100

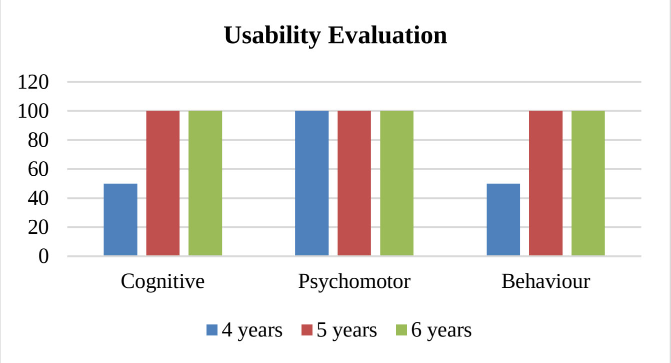


Fig. 4. Evaluation of usability Testing.

TABLE. VIII. HIGH-FIDELITY PROTOTYPE

Age of Children	Cognitive	Psychomotor	Behaviour
4 years	MISSION 1: Some of them were unable to memorize the inflammable substances easily. But still able to complete the mission to find inflammable substances. MISSION 2: Able to complete the mission to escape from the burning house. MISSION 3: Able to complete the mission to save lives of human and animal.	MISSION 1: A bit slower to control the game controller buttons. But still able to control the left and right buttons but facing difficulties with the speed of moving objects. MISSION 2: Able to control the left and right buttons easily. MISSION 3: A bit slower to control the game controller buttons. But still able to control the left and right buttons but facing difficulties with the speed of moving objects.	MISSION 1: Giving the positive response towards inflammable substances. MISSION 2: Shouting loudly to seek for help when facing danger. Happy to complete the game. MISSION 3: Giving the positive response to save lives and the engagement while playing the game.
5 years	MISSION 1: Able to complete the mission to find inflammable substances. MISSION 2: Easy to complete the mission to escape from the burning house. MISSION 3: Able to complete the mission to save lives of human and animal.	MISSION 1: Able to control the buttons easily without difficulties. MISSION 2: Able to control the buttons easily to move the characters. MISSION 3: Able to control the buttons easily with the fast speed on catching the objects.	MISSION 1: The engagement of the player to remember every inflammable substance. MISSION 2: Afraid to solve the game because of the shouting voice but still able to point out the escape door. MISSION 3: Showed the engagement of the player to save lives and happy to complete the game.
6 years	MISSION 1: Easy to complete the mission to find inflammable substances. MISSION 2: Easy to complete the mission to escape from the burning house. MISSION 3: Easy to complete the mission to save lives of human and animal.	MISSION 1: Facing no difficulties in handling touch interaction with the controlling buttons. MISSION 2: Easy to handle all the buttons precisely to move the character. MISSION 3: Facing no difficulties in handling touch interaction with the buttons.	MISSION 1: Know the functions of inflammable substances and showed the engagement while playing. MISSION 2: Shouting loudly to seek for help to escape from fire and completing the game easily. MISSION 3: Showed the engagement while playing to save lives from fire.

The results showed that the children at the age of four years old have difficulties in controlling the speed of moving objects and game controller buttons. But they still managed to play the game till the end. For Mission 1 and Mission 3, either they were able to use touch interaction by using only one finger or two fingers on both hands. These missions let the children to focus by using hand-eye coordination activities.

For Mission 2, most of them completed the mission with ease by escaping from the burning house. It required the children to think as fast as they could to find the escape way. Some of them were shouting very loudly to seek for help. For Mission 3, it showed that the engagement of playing the game improved their attention span.

They tried to solve the mission with limited time. Thus, by scoring points in the game kept them motivated to play more. The children were able to hear the instruction given clearly and followed the gameplay. All these missions provided helped to stimulate their brains in solving problems if they were facing the fire hazard situation. Through this APi fire safety game helped the preschool children to improve awareness on fire safety issue at the early age. So that they noticed on how life-threatening fire is. It may cause injuries and death. By conveying the basic knowledge of fire safety on them through game environment, it helped them to engage more in learning.

Therefore, it showed the effectiveness of the Model Game-Based Learning in Fire Safety developed. The preschool children understood the gameplay and learned the issue of fire safety with full of excitement. Based on the APi Game-Based Learning helped to improve their cognitive, psychomotor and behaviour towards fire safety issues. Learning fire safety issues through gaming environment improved their motivation too. Thus, game-based learning allowed better understanding of extremely dangerous effects of fire on children. In fact, educational game helped in promoting fire safety awareness.

V. Discussion

Time was a crucial moment in handling fire situation. In fact, causing the risks of injuries and death could be reduced by providing the fire training simulator. Virtual reality based fire training simulator provided the general public and firefighters to train them in making decisions and organized responses towards fire safety [5]. By providing activities to achieve the goals of effectiveness, the users needed to perform training and experienced fire environments. Apparently, evacuation and rescue activities of fire situation at road tunnels were evaluated on the users. Based on the proposed framework, the functions and real-time performance of the simulator were verified. Their behaviours on initiating right actions were observed specifically in handling real fire situations.

Therefore, by accomplishing the missions or activities of fire safety helped the users to think as fast as they could in order to rescue themselves. Conveying the information of fire safety without harming the users was important to let them aware and initiate effective response towards danger. There were a lot of technologies used to deliver new methods of learning and training the fire safety education such as CAVE and computers [10]. In addition, these technologies used should be compatible with the users' abilities and skills.

Based on BIM (building information modelling) that supported by virtual reality and serious games exposed the users on awareness of emergency training [14]. By providing real-time fire evacuation guidance let the users to understand evacuation process. Three main aspects were focused which were real-time two ways information updates, real-time evacuation route and real-time location of building. Through the activities provided for the users to accomplish, their behaviours were taken seriously towards the experiment. The research highlighted the effective way of actions for human behaviour in emergent situations. The Model of Game-Based Learning in Fire Safety focused on the real-time game where it provided users to accomplish the mission within the limited time given. Hence, by playing the game would improve the children's ways of thinking and right actions taken.

In this Model of Game-Based Learning in Fire Safety research highlighted the three main aspects of preschool children which were cognitive, psychomotor and behaviour towards fire hazard. Through gaming environment attracted them to engage well with the missions provided in APi prototype. Concerning the fact that children have limited capabilities compared to adults, all the missions created were suitable for their ages, knowledge and skills [20]. Thus, they could solve the missions in given time to test their skills in handling the technology of tablet. By using the touch interaction could identify their psychomotor of fine motor skills and capabilities. APi prototype provided learning through game environment to verify the model developed.

The usability evaluation showed the effectiveness of APi prototype tested on the preschool children to educate them in learning fire safety issues. The preschool children were able to complete the mission with minimal supervision. On the other hand, some of them were facing difficulties in controlling the buttons. It happened because of the speed of the moving objects. Overall, all the preschool children showed positive response towards fire hazard and knew the basic skills to escape from danger.

APi prototype was developed based on the user requirements of the Model of Game-Based Learning in Fire Safety. Therefore, all multimedia components used such as audio and animation in developing the APi prototype eventually helped the preschool children to stay focus while playing. Audio and animation aroused their attentions and engaged them to play longer. With the instructions given to them by using voice, the preschool children did not need to read the instructions because some of them were still having difficulties in reading progress. While, the use of technology tablet attracted the preschool children to play because of their ages to explore the surrounding.

Extending to this research, we hoped that the Model of Game-Based Learning in Fire Safety can be used in other disaster such as flood to guide the preschool children. They need to be exposed and trained well from the early age as the prevention from danger situation. This is because the children are highly-risk towards fire hazard that can cause damages and death. With this in mind, saving lives are important by doing the right actions.

VI. Conclusion

This study indicates the development of the Model of Game-Based Learning in Fire Safety for preschool children. All the user requirements needed should be discussed more in order to train them on fire safety issues. Due to the lack of awareness towards fire hazard, there is a need to educate them in the form of edutainment to lower the risk of injuries.

Therefore, the preschool children needed to be exposed well and alerted on how life-threatening fire is. Despite the design of the Model of Game-Based Learning in Fire Safety promoted on the fire safety awareness. Hence, it helped to foster the children's learning process in their daily lives. Further study should be directed towards addressing on how the model can be improved based on children's needs such as cognitive, psychomotor and behavior aspects.

Acknowledgment

This work was supported under the Strategic Research Grant of KRA-2018-025, Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia.

References

[1] Azman, I. & Mohd Ridwan, A. R., "Performance-based reward administration as an antecedent of job satisfaction: A case study of Malaysia ' s fire and rescue agencies," Malaysian Journal of Society and Space, vol 7, pp. 107-118, 2016.

[2] Marrion, C. E., "More effectively addressing fire/disaster challenges to protect our cultural heritage," Journal of Cultural Heritage, vol 20, pp. 746–749, July 2016.

[3] Noorhidawati, a., Ghalebandi, S. G., and Siti Hajar, R, "How Do Young Children Engage with Mobile Apps? Cognitive, Psychomotor, and Affective Perspective," Journal of Convergence Information Technology, vol 87, pp. 385-395, July. 2015.

[4] Wei, W.J., & Lee, L.C., "Interactive technology for creativity in early childhood education," Jurnal Teknologi, vol 75(3), pp 121-126, Nov. 2015.

[5] Morrongiello, B. a., Schwebel, D. C., Bell, M., Stewart, J., & Davis, A. L., "An evaluation of The Great Escape: Can an interactive computer game improve young children's fire safety knowledge and behaviors?," Health Psychology, vol 31, pp 496–502, Apr. 2012.

[6] He, Q., Hong, X., Zhao, G., & Huang, X., "An Immersive Fire Training System Using Kinect," in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, 2014, paper 14, p. 231–234.

[7] Chin, L. C., & Effandi Zakaria, "Development and Validation of the Game-Based Learning Module to Enhance Mathematics Achievement, Positive Learning Behaviours and Pro Social Behaviours," Journal of Science And Mathematics Letters, vol 2, pp. 23–31, Jan 2014.

[8] Tang, S., Hanneghan, M., & El Rhalibi, A., Introduction to games-based learning, Games Based Learning Advancements for Multi-Sensory Human Computer Interfaces, New York: IGI Global, 2009.

[9] Green, C. S., Kattner, F., Eichenbaum, A., Bediou, B., Adams, D. M., Mayer, R. E., & Bavelier, D., "Playing Some Video Games but Not Others Is Related to Cognitive Abilities: A Critique of Unsworth et al. (2015)," Psychological Science, vol 28, pp. 679–682, 2017.

[10] Williams-Bell, F. M., Kapralos, B., Hogue, A., Murphy, B. M., & Weckman, E. J., "Using Serious Games and Virtual Simulation for Training in the Fire Service: A Review," Fire Technology, vol 51, pp. 553–584, March 2015.

[11] Tsai, M. H., Wen, M. C., Chang, Y. L., & Kang, S. C., "Game-based education for disaster prevention," AI and Society, vol 30, pp. 463–475, Nov 2015.

[12] Noorhidawati, a., Ghalebandi, S. G., and Siti Hajar, R, "How Do Young Children Engage with Mobile Apps? Cognitive, Psychomotor, and Affective Perspective," Journal of Convergence Information Technology, vol 87, pp. 385-395, July. 2015.

[13] He, Q., Hong, X., Zhao, G., & Huang, X., "An Immersive Fire Training System Using Kinect," in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, 2014, paper 14, p. 231–234.

[14] Wang, B., Li, H., Rezgui, Y., Bradley, A., and Ong, H. N., "BIM based virtual environment for fire emergency situation,", The Scientific World Journal, pp 22, Aug. 2014.

[15] Abdul Jabbar, A. I., & Felicia, P., "Gameplay Engagement and Learning in Game-Based Learning: A Systematic Review," Review of Educational Research, vol 85, pp 740–779, 2015.

[16] Kamarudin, D., Hussain, Y., Applegate, E. B., & Yasin, M. H. M., "An Ethnographic Qualitative Study On The Malaysian Preschool And Special Needs Children's Home And School Reading Habits," International Journal of Pedagogy and Teacher Education (IJPTE), vol 2, pp. 224–234, April 2018.

[17] Towers, B., "Children ' s knowledge of bushfire emergency response," International Journal of Wildland Fire, vol 24, pp 179–189, March 2015.

[18] Zaini, N.A., Noor, S.F.M, Wook, T.S.M.T, "Evaluation of APi Interface Design By Applying Cognitive Walkthrough," International Journal of Advanced Computer Science and Applications, vol 10, 2019.

[19] Zainab, H, "Study of Touch Gesture Performance by Four and Five Year-Old Children: Point-and-Touch, Drag- and-Drop, Zoom-in and Zoom-out, and Rotate," Information Tech. thesis, Minnesota State University, Mankato, July 2014.

[20] Anthony, L., Brown, Q., Nias, J., Tate, B., Mohan and S., "Interaction and recognition challenges in interpreting children's touch and gesture input on mobile devices," in Proceeding of the 2012 ACM International Conference on Interactive Tabletops and Surfaces – ITS'12, 2012, p. 225.

[21] Sung, H.-Y., & Hwang, G.-J., "A Collaborative Game-based Learning Approach to Improving Students' Learning Performance in Science Courses," Computers & Education, vol 63, pp. 43–51, Nov 2012.

[22] Shi, Y.-R., & Shih, J.-L., "Game Factors and Game-Based Learning Design Model," International Journal of Computer Games Technology, pp 1–11, Aug 2015.

[23] Preece, J., Sharp, H., Rogers, Y., Interaction Design: Beyond Human-Computer Interaction, vol 4, United Kingdom: Wiley, 2015.

[24] Zaini, N.A., Noor, S.F.M, Wook, T.S.M.T, "The User Requirements of Game-Based Learning in Fire Safety for Preschool Children," Journal of Advanced Science Letters, vol 24, pp. 7795-7799, Oct 2018.

[25] Singh, D. K. A., Ab Rahman, N. N. A. A., Rajikan, R., Zainudin, A., Mohd Nordin, N. A., Karim, Z. A., & Yee, Y. H., "Balance and motor skills among preschool children aged 3 to 4 years old," Malaysian Journal of Medicine and Health Sciences, vol 11, pp. 63–68, Jan 2015.

[26] Oh, S. J., Fritz, M., & Schiele, B., "Adversarial Image Perturbation for Privacy Protection A Game Theory Perspective," in Proceedings of the IEEE International Conference on Computer Vision, 2017, p. 1491–1500.

[27] Wang, K.-C. ., Shih, S.-Y. ., Chan, W.-S. ., Wang, W.-C. ., Wang, S.-H., Gansonre, A.-A., Yeh, M.-F., "Application of building information modeling in designing fire evacuation-a case study," in 31st International Symposium on Automation and Robotics in Construction and Mining, ISARC 2014 - Proceedings, (Isarc), 2014, p. 593–601.

[28] Wook, T. S. M. T., Mohamed, H., Judi, H. M., and Ashaari, N. S., "Applying cognitive walkthrough to evaluate the design of SPIN interface," Journal of Convergence Information Technology, vol 7, pp. 106-115, March 2012.

A Defeasible Logic-based Framework for Contextualizing Deployed Applications

Noor Sami Al-Anbaki[1], Nadim Obeid[2], Khair Eddin Sabri[3]
King Abdullah II School for Information Technology
University of Jordan, Amman, Jordan

***Abstract*—In human to human communication, context increases the ability to convey ideas. However, in human to application and application to application communication, this property is difficult to attain. Context-awareness becomes an emergent need to achieve the goal of delivering more user-centric personalized services, especially in ubiquitous environments. However, there is no agreed-upon generic framework that can be reused by deployed applications to support context-awareness. In this paper, a defeasible logic-based framework for context-awareness is proposed that can enhance the functionality of any deployed application. The nonmonotonic nature of defeasible logic has the capability of attaining justifiable decisions in dynamic environments. Classical defeasible logic is extended by meta-rules to increase its expressiveness power, facilitate its representation of complex multi-context systems, and permit distributed reasoning. The framework is able to produce justified decisions depending on both the basic functionality of the system that is itself promoted by contextual knowledge and any cross-cutting concerns that might be added by different authorities or due to further improvements to the system. Active concerns that are triggered at certain contexts are encapsulated in separate defeasible theories. A proof theory is defined along with a study of its formal properties. The framework is applied to a motivating scenario to approve its feasibility and the conclusions are analyzed using argumentation as an approach of reasoning.**

Keywords—Context-awareness; nonmonotonicity; defeasible logic; distributed reasoning; argumentation

I. Introduction

It is fair to say that the ubiquitous computing paradigm revolutionized our understanding of computing and what it can deliver. It merges computer devices and sensors in an integrated environment, to provide better communication and enhanced accessibility to information sources. The final objective is to provide users with services available whenever, however, and wherever needed [1]. Applications should be intelligent enough to handle the mobility of users and resources themselves as well as the ever-changing context in a seamless manner with minimum human intervention. In other words, applications should be context-aware.

The term Context-Aware Computing was first introduced in 1994 [2], this study focused on the communication aspects related to broadcasting information from a server to its clients. Context was considered to be the information related to the location of users and other objects in the system and how this information changes over time, in addition to the communication overload. In [3], context awareness role in mobile computing was discussed, the study considered context to be the identity of the user, nearby users, location, time and season. Other studies that discussed what context could be can be found in [4] [5] [6].

In 2001, Dey [7] introduced the most well-known definition of context: "Any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves". This definition was a milestone in the growth of the notion of context as it is generic, operational and exceeded the boundary of (time, location and user's identity) where context was always defined accordingly. On the other hand, Context-Awareness is considered to be the ability of the system to sense (gather information) about its surrounding physical and operational environment at any given time, perceive and adapt behavior accordingly [8].

A context-aware system should support mechanisms for collecting contextual information, representation, reasoning and application [9]. Contextual information is domain-dependent, it can be any piece of information that describes the entity involved in the interaction, it could be time, location, task, identity, etc. or a group of them. The acquisition of this information is beyond the scope of this work, it is achieved using different technologies. The emphasis of this work is on the two most important phases in any framework that supports context awareness: representation and reasoning.

In this paper, a generic framework is present that can guide the contextualizing process of deployed applications. The framework provides a powerful mechanism to represent multi-context distributed systems and permits distributed reasoning. An extension to defeasible logic theory was proposed by adding the notion of meta-rules that are able to reason over theories; this enhancement would open the door of new usage of DL in the representation and reasoning of complex systems.

The significance of the study lies in its conceptual analysis of context by considering it to be both, information that can characterize entities and information that has the ability to characterize a whole new behavior of the system.

Another advancement of the framework is that it permits distributed reasoning which is a challenging area in AI, as there is no central authority to control the context flow in the overall system, but rather each component in the system is allowed to add its own view of manipulating contextual knowledge. This is achieved using a separation of concerns principle and can highly increase users' and administrators' satisfaction.

The work is of both theoretical and empirical significance to the research in context awareness and contextual reasoning. The theoretical importance lies in the proposed extension to the defeasible theory that permits the representation of complex multi-context systems and facilitates distributed reasoning, while empirical significance lies in the ability to employ the framework to contextualize any kind of application. It allows the developers of context-aware applications to easily represent and manage different behaviors of the application in different contexts.

This paper is organized as follows: Section 2 highlights some issues in contextual reasoning. Section 3 presents related work. Section 4 presents the defeasible logic. An illustrative scenario is presented in Section 5. Section 6 discusses our interpretation of context and context-awareness. Section 7 presents the proposed framework of context-awareness. Section 8 defines the formal proofs of the framework. An implementation of the illustrative case study in the proposed framework is presented in Section 9 along with its analysis. A brief discussion is presented in Section 10 and finally, Section 11 covers the conclusions and future work.

II. Some Issues in Contextual Reasoning

There are many alternatives in the literature that deal with knowledge representation and reasoning issues [10] [11] [12] [13] [15] [14] [16] [17] [18]. However, when this knowledge is characterized as contextual knowledge (i.e. "as information that can be used to characterize the situation of an entity"), there are extra properties that need special treatment.

- First of all, context is domain-dependent (e.g. the identity of a user plays a subtle role in an access control system, but it is not important in a supermarket billing system). This is considered an appealing property that helps to develop personalized services.

- Second, context is a conflict-sensitive concept, i.e. multiple sources of contextual information might lead to infer conflicting decisions. This happens due to multiple sources of contextual information which lead to ambiguity. The study in [19] highlighted other problems related to contextual information in that they might be unknown, imprecise, and erroneous.

- Third, when reasoning is employed, context becomes nested. In complex systems, the context of an entity is not merely restricted to basic contextual attributes that are collected directly from sensors (e.g. the temperature of a room) but rather, it refers to complex contextual attributes that are inferred from basic contextual attributes. For example, if the temperature of the room is between (72 F and 76 F), the room warmth is comfortable), in this way, a room with a temperature degree (74 F) is characterized by two contextual attributes, its temperature is (74 F) and its warmth is comfortable. This different level of abstraction gives context an operational power, such that a basic contextual attribute may lead to a whole new behavior and direct the characterization of many other aspects in the system e.g. a room's temperature may affect not only the degree of relief in the room but rather may play a role in deciding the placement of certain assets in the room e.g. a server, or turning on the air conditioning which is, in turn, affects the energy consumption, and so on.

These characteristics lead to challenges that cannot be avoided especially in complex systems that operate in ubiquitous environments where the system contains multiple entities and the process integration spans organizations where interactive entities in the system may belong to different authorities and each works under different regulations. The system should be able to reason and reach justifiable decisions regardless of these complications.

To handle these issues, a solid representation mechanism should be employed that can deal with ambiguity and a concrete conflict resolution mechanism that enables inferring justifiable non-conflicted decisions. McCarthy [20] was one of the first scientists that point out the issue of contextual reasoning. He suggested that the combination of nonmonotonic reasoning and context formalism would constitute an adequate solution to overcome the problems associated with including contextual information in the decision-making process. Nonmontonicity provides mechanisms that allow the system to reason and reach justifiable decisions by retracting conclusions that turned out to be incorrect and derive new, better-justified conclusions instead [21]. This makes it very suitable to tackle the reasoning process in dynamic situations with incomplete/changing information.

Defeasible logic (DL) [22] is a well-known skeptical nonmonotonic logic that can be used in dynamic environments due to its characteristics: it is expressive, natural, not ambiguous and programmable. It has attracted many researchers to incorporate it in different application domains such as modeling of contracts [23], legal reasoning [24], modeling social agents [25], modeling social commitments [15] [17] [18], etc. The most significant feature of DL is that it preserves the consistency of the system regardless of conflicts because it does not produce contradictory conclusions. When a conflict occurs, conflicting rules do not arouse. It supports the use of priorities to resolve these conflicts to allow the system inferring with incomplete/partial information.

III. Related Studies

There are many attempts in the literature to formalize context in order to be able to reason based on its attributes along with its accompanied obstacles that might lead to conflicts in the decision-making process.

As the issue of context sensing and integration in highly connected to the technical infrastructure of the system, most of the researches that aimed to define generic frameworks for context awareness, pointed out the architecture aspects of the framework, e.g. the authors in [26] proposed a context management framework that enables the collaboration of multiple domains by exchanging contextual information. Their framework highlighted the architectural issues; it is based on a peer-to-peer architecture. The framework imposes a hierarchic ordering of context sources and multiple reasoning tools. This facilitates adaptability as new context sources and reasoning techniques can be added. The most important parts of the framework are the uniform interface where all of the context-provides are attached to a reasoner where all the reasoning methods can be employed.

Other studies presented techniques to deal with contextual information, e.g. [27] defined a Context Toolkit that provided an infrastructure for prototyping context-aware applications. However, it didn't provide a mechanism to reason about contexts. There is no formal tool to write reasoning rules for contexts or to infer higher-level contexts decisions.

Formal representation of context can be found in [28], where an architecture and programming framework for triggering application adaptation to changes in context was proposed. It employed basic (if-then rules) to formalize the behavior of an application in different contexts. In [29], first-order logic was used to describe contextual information and reasoning was done using Boolean operators and existential and universal quantifiers.

Recent trends in context-awareness pointed out the significance of generic frameworks in manipulating context flow in smart environments.

A formal representation of context can be found in [30] where the authors used ontologies to model information gathered from IoT devices in a smart home environment and used Description logic [31] to deduce activities depending on the gathered contextual attributes from the devices.

Another study [32] proposed a context-aware framework for multi-agent environment. Agents in their framework extract contextual information from ontologies; in fact, an agent can extract its rules and facts from one or more ontologies. Each agent performs reasoning based on the collected information and communicates with other agent(s) using bridge rules; the concluded decision is used to adapt the system behavior. The framework is used to generate preference sets for users, which is a set of active rules for each user.

Defeasible Logic DL [16] [22] had approved to be one of the famous logic tools that are successful to characterize contextual reasoning; it has a nonmonotonic relation between the premises and their consequences which is an effective way of formalizing the dynamic nature of ubiquities computing. Several studies succeeded to build models that could reason in the shade of contextual information based on DL [19] [33] [34]. However, these studies handle context in an environment of operating agents, they consider context to be whatever local knowledge the agent has. This view is correct and it serves the goal of showing how collaborating agents can cooperate to achieve a specific goal regardless of the challenges caused by the imperfect nature of context.

These approaches can be viewed as enhanced versions of previous approaches that aim at solving the partial knowledge issues of autonomous agents by collaboration. This is achieved using bridge rules [34] and mapping rules [33]. None of these studies investigated the effect of context on the decision made by each agent/entity and how contextual information can affect the overall behavior of the system.

The proposed framework discusses how to enhance deployed applications using context. Rather than considering it to be raw agent's knowledge received from contributed sensing devices, a conceptual view of context is adopted, it considered it as a concern/goal that needs to be achieved, it is different than the models in the literature as it defines the boundary between what local knowledge the system is already designed to manipulate (i.e. what is the input information that system rules make decisions accordingly) and what is contextual knowledge that could be used to enhance the system operation. We argue that the integration of contextual information in the reasoning process of a system that is driven by many concerns can not only be achieved by adding additional attributes/predicates that describe contextual information and additional rules that manipulate them. The projection of contextual knowledge on the system affects both the nature of its base functionality (base concern) and the way it handles cross-cutting objectives, concerns or exceptions. This simulates how humans think. Humans' decisions are never static; they are always changing based on upcoming knowledge, i.e. current context. For example, a student might choose an academic major based on his/her interest (a basic aspect), in addition to the GPA, budget, family opinion, the need for the labor market at that time (a contextual aspect).

The framework is implemented using Defeasible Logic DL, it benefits from both the expressiveness power of logic in representing knowledge and the nonmonotonic feature of the defeasible theory that facilitates a smooth reasoning process in a dynamic environment.

Based on this representation of context, the framework can be viewed as a platform that can be used to augment ubiquitous applications with context awareness by employing a conceptual view of context that is able to infer high-level decisions. The framework allows easy integration of different modes of operations triggered by different contexts and at the same time preserves the consistency of the decisions made by the system.

IV. Defeasible Logic

Defeasible logic (DL) was proposed by Nute in 2001 [22], unlike monotonic reasoning, it has a nonmonotonic relation between the premises and their logical consequences which made it suitable for reasoning in dynamic environments. In order to illustrate the nonmonotonic reasoning power, assume the situation of the following example that resembles the monotonic kind of reasoning.

Example 1: *Bob is often invited to social events by his friends. He usually attends these events; however, he has the following two preferences about going to a party.*

> *P_1: If the inviting person is one of his closest friends, he would go.*
>
> *P_2: He prefers not to go if Adam is invited.*
>
> *Bob was invited by his best friend, Julie, and she told him that Adam is invited as well.*

In a monotonic kind of reasoning, the two rules are applied and both of their consequences are valid (*go* and *don't go*) which leads to inconsistency, it is the system developer's responsibility to design rules that avoid such conflicts. Monotonic reasoning needs a lot of administrative effort and it neither scales well nor can be used in an environment with multiple administrative authorities. On the other hand, a nonmonotonic reasoning approach is founded on the ability to infer tentative conclusions that can be retracted based on new evidence [14].

Formally, DL can be seen as an extension to first-order predicate calculus FOPC [35], with the addition of the defeasible implication (⇒) that is used to infer the tentative conclusions, and the ambiguity-blocking priority relation (>) that is used to preserve the consistency of the system and infer justifiable conclusions in both static and dynamic domains.

Basically, a defeasible theory D (also called a knowledge base in DL) is a triple *(F, R, >)*, it consists of three main components:

1. **Facts (F)**: is a finite set of literals that represent indisputable statements.

2. **Rules (R)**: is a finite set of three types of rules ($R = R_s \cup R_d \cup R_f$) each rule comes in the form,

$$R: Ant(R) \bullet\!\!\rightarrow Conseq(R)$$

Where, (*R*) is a unique label, *(Ant(R))* is an antecedent, ($\bullet\!\!\rightarrow$) is a set of one-direction arrows that identify three types of implications (→ to denote strict rules, ⇒ to denote defeasible rules and ⇝ to denoted defeaters) and a *(Conseq(R))* is the head/consequence which is the conclusion of the rule. R[B] means the set of rules in R with consequence B.

A) *Strict rules R_s*: is a set of rules that cannot be defeated, e.g. " if a country is on the equator, it would be very hot during summer",

R_1: equatorial(X) ∧ during_summer(X) → very_hot(X),

B) *Defeasible rules R_d*: is a set of rules that can be defeated by contrary evidence, e.g. " if a country is on the coast, it is usually very hot during summer",

R_2: coastal(X) ∧ during_summer(X) ⇒ very_hot (X),

Rule R_2 indicates that during summer, coastal counties weather is very hot unless there is other evidence suggesting a contrary result, such as (R_3) which states that "if it rains in summer, the weather would not be very hot"

R_3: raining_at(X) ∧ during_summer(X) ⇒ ¬very_hot(X)

C) *Defeaters R_f*: rules presented by (⇝), they do not use to conclude but rather to prevent deriving conclusions of some defeasible rules by producing evidence to the contrary e.g.

R_2': coastal(X) ∧ during_spring(X) ⇝ ¬very_hot (X),

3. **The superiority relation** >: is a binary relation over the set of defeasible rules R_d i.e. ($> \subseteq R_d \times R_d$). It is defined externally and statically to resolve conflicts. For example, given that defeasible rules R_2 and R_3 are both approved, no conclusive decision can be made about whether the weather is very hot or not. But, if the superiority relationship ($R_3 > R_2$) is introduced, then R_3 overrides R_2 and it can be concluded that the weather is not very hot while it is raining even during the summer season. The superiority relation > is acyclic, that is, the transitive closure of > is irreflexive.

The interaction of these three components permits the conclusion of justifiable decisions. This is referred to in Fig. 1.

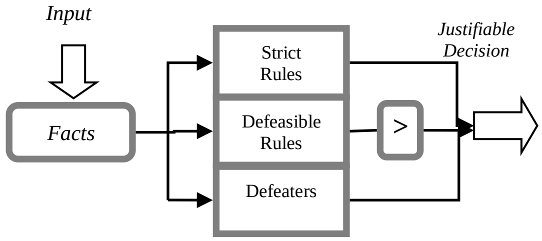


Fig. 1. Classical Defeasible Logic.

V. Illustrative Scenario

In this section, a motivating scenario from a ubiquitous environment is presented in order to illustrate the challenges of the domain and the capabilities of the proposed framework in reasoning in such environments.

Assume a situation where a smart application for lecturers' and employees' phone calls management inside a university campus was installed on all lecturers' and employees' mobile phones. This application manages the calls during the time when the lecturer is giving an online course.

The system was originally designed after an anti-disturbance base concern; it filters out calls based on the *identity* and *location* of the caller. It contains three rules that reason about two contextual attributes: (1) the identity of the caller that is identified by either: a) the caller being in the urgent list e.g. the dean's secretary b) the identity is unknown i.e. the number cannot be mapped to any of the names in the phone database, and (2) the location where the call is issued, it can be either a local or international call. The system makes its decision according to the following three rules:

- If the call is issued by a person on the urgent list, the phone rings.
- If the call is an international call, the phone rings.
- If the caller was unknown, the phone wouldn't ring.

To resolve the conflicts that might occur due to characterizing the caller as (being in the urgent list, international, and unknown); the user has to set priorities to decide which argument to support if more than one attribute hold. Sami set that if the call is international, the phone would ring even if the caller is unknown.

To further enhance the capabilities of the system using context awareness, the users of the application opted to personalize the service by formulating their own preferences. The application was attached to three different context providers that their knowledge can be used to better personalize the functionality of the system: a location detection service of the lecturer, schedule and the number of students engaged with the professor in the online session. A system user can set his/her own preferences based on the three available contextual attributes (location, schedule, and status that could be either busy or not busy depending on the number of students that are active during the online session). Preferences are activated upon turning on a flag of interest on the user's mobile phone. For example, Professor Sami has the following rules:

- If he is located inside Samsung-Lab, the phone wouldn't ring.
- If there is a scheduled lecture, the phone wouldn't ring.
- If he is engaged with less than five students in an online session, he is not busy and the phone could ring. This rule overrules the first two rules.

Suppose the situation when the dean asked the secretary Linda to call Professor Sami. Linda's number is in the urgent list; according to the anti-disturbance system rules, the phone should ring. However, Sami is inside Samsung Lab and is in an active session with five students, the phone should not ring.

From Linda's point of view, the phone should ring. She is sure that her number is already listed in the urgent list; however, she is not aware of Sami's preferences. The system is not able to decide which argument to support, the anti-disturbance concern argument or the users' preferences argument. Thus, an inter-concern conflict resolution mechanism is used to regulate the decision-making process.

As the end goal is to deliver personalized context-aware service, the designer sets that the decision inferred by users' preferences overrules the base system decision. In this arrangement, Sami won't be informed about the call.

One of the stakeholders, namely, the dean, was not satisfied with the services provided by this system, as his secretary uses the schedule of all professors and the administrative staff to determine the time of urgent meetings and she calls them based on this knowledge. However, according to the above settings, even though the user has no scheduled lecture at the time of the call but is inside the lab giving advice to some students on an online session, he/she was not informed of urgent meetings.

To resolve this issue, the system should address stakeholders' concerns such as urgent invitations. The system is connected to a meeting database that is controlled by several stakeholders, it saves the time, location, invitees of meetings, some of them are saved in prior e.g. a workshop and some of them are set up in an ad-hoc manner e.g. urgent meeting to discuss exams results. This concern manages the system as follows:

- If a person is invited, he/she should be informed.
- However, if the invitee has a scheduled lecture, he/she should not be informed.

The inter-concern conflict resolution mechanism should be carefully designed to represent the directions of the stakeholders as they represent a higher administrative authority. Such that the decision made according to the urgent invitations concern would be supported.

This simple scenario clarifies the challenges of using contextual knowledge in the decision-making process. In addition to the challenges of distributed reasoning in systems that encompass multiple authorities, each has its own preferences/regulations and its own interpretation of internal contextual knowledge. Each authority aims at making the decision referring to its own rules. This motivates the need to employ a distributed reasoning mechanism that can handle the production of justified and solid decisions in multi-context ubiquitous environments.

VI. Context and Context Awareness

Due to the enormous improvement of how computers, diffused sensors, and other devices collect situational/ contextual information, a lot of researchers tried to define context in several manners. Basically, context is identified by its attributes i.e. contextual information/variables that: (1) describe the user in an interaction with an application, the application/process, the environment and the interaction itself, (2) can be used to deliver more user-centric personalized services. The range of this information is quite vast and it depends on the domain itself, it could be time, location, number of users, the identity of the user, user's emotional states, the focus of attention, etc. [8] [9] [36] [37].

In order to build a framework that is able to enhance the operation of any application using contextual knowledge, it is very important to define the system's manipulated knowledge and enhancing contextual knowledge. Thus, throughout this work Dey's definition of context is extended to best suit this purpose: "For a deployed application, context is any information used to characterize the situation of an entity and can be sensed, collected and represented. This information is not part of the group of information that already describes that entity in the deployed application. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves".

In this work, the set of contextual information that represents domain knowledge C in an environment would be classified according to its presence in the system, as shown in Fig. 2:

1) Information that is collected from the environment in the digital form or can be presented digitally, collected context, ($C^o \subseteq C$) e.g. identity of the user, light, sound, location, size, etc.

2) Information that the system is designed to manipulate ($C^u \subseteq C^o$) e.g. in an access control model, the identity of the user and his role is used to determine what object(s) he/she can access.

3) Contextual information that can be added to enhance the functionality of the system ($C^h \subseteq C^o$) e.g. in an access control model in a dynamic environment, the time and location of the user requesting access is of major importance.

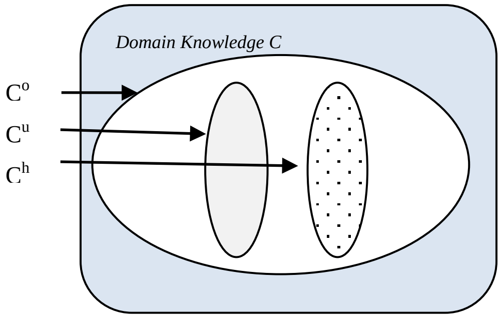


Fig. 2. Spectrum of Context.

A context of an entity is the net of all contextual attributes that describe that entity, the attributes might be physically collected by sensors, e.g. a GPS service can return the location of a person, LocatedIn(sami, home) to state that Sami is located at home (different schemes can be used to describe the location of an entity, e.g. XY coordinates) or logically constructed.

The proposed framework is built on defeasible logic; first-order literals are used to represent contextual attributes. Each literal represents a property that an entity holds or can be characterized by, it contains a name and a value e.g. location("university", near), role("mary", manger), connected_users(5) and so on. A literal is an atomic formula or its negation if α is an atomic formula; $\neg\alpha$ is its complement [38]. It should be defined in advance what is the meaning of the name and what is the range of values the literal may hold.

In the proposed framework, enhancing contextual attributes are referenced in four different ways:

1) According to the way they are gathered:

- Basic contextual attributes, to refer to the attributes that are collected directly from sensors, e.g. Humidity(basement, 40%), Temperature(basement, 65 F).
- Complex contextual attributes, to refer to the attributes that can be assembled as a result of logical operations on basic or other complex attributes, e.g. Humidity(basement, 40%) $\wedge$ Temperature(basement, 65 F) $\Rightarrow$ Comfortable(basement).
- Due to the multiple sources of contextual information, the decidability of complex attributes needs extra care, e.g. Noise(basement, 120 dB) $\Rightarrow$ ~Comfortable(basement).

In this case, any reasoning mechanism should uses priorities to resolve this issue.

2) According to the way they are manipulated:

- Internal contextual attributes, to refer to the attributes that are manipulated locally by the entity/administrative domain.
- External contextual attributes, to refer to the attributes that are manipulated outside the entity/administrative domain.

A context of a system that contains multiple collaborative sub-systems or components is a snapshot of the system's situation at a given point/interval of time. This encapsulates the external contextual attributes, internal contextual attributes of each sub-system and component in addition to the relationship between those subsystems and components.

A context-aware system is a system that is able to make a solid justified decision for every upcoming context. A contextualized deployed application is a system that is able to adapt its behavior in the shade of collected enhancing contextual attributes. In the proposed framework, the defeasible logic machinery would be employed for knowledge representation and reasoning and a concern-based model for context integration.

VII. Defeasible Logic Framework for Context Awareness

A deployed application, a base system, can be seen as a domain of knowledge that is governed by static rules that manipulate its local knowledge, i.e. (C^u); it is designed to serve a certain purpose. Augmenting such a system with context awareness can considerably improve its functionality by making it adaptable to the processing environment in order to provide a better experience to the user, better utilization of resources, etc. Contextual knowledge that is integrated into the system (C^h) can be embedded either implicitly or explicitly. In other words, it can be used to contextualize the base system's rules, if it is added by the same administrative authority and it serves the same purpose/concern of the base system, or it can be used to characterize other entities concerns that could be compatible or crosscutting to the base system's concern, they are triggered at some exceptional situations, normally this knowledge is perceived by other participating authorities or components in the system and it indicates their concerns from their own viewpoint. A context-aware system should be carefully designed to permit distributed reasoning and at the same time make justifiable decisions in spite of the fact that the distributed entities might have conflicting concerns.

The proposed context-aware framework based on defeasible logic is a theory L<G, β, D, λ>, that consists of the following components.

A. Triggers G

Triggers is a finite set of positive and negative ground literals that represents external basic contextual attributes acquired from the application domain. Triggers are imported from the system's global knowledge that is not necessarily known by the participating entities/users. They have a certain property is that they are issued by/collected from multiple participating sub-domains or different authorities in the application domain. It should be noted that not all contextual attributes can be used as triggers, a trigger's impact extends far beyond changing a single rule, yet, it can add/remove/change different rules and regulations in different components of the system e.g. an emergency situation that leads to a break glass procedure.

Formally, each trigger is an atomic formula. A valid framework can have no two complementary triggers i.e. an atomic formula and its negation. Triggers activate concerns using meta-rules. Meta-rules are rules that consequences are rules; they have been used in the literature as a powerful machine that facilitates reasoning about rules for contextualizing the provability of goals [34]. In the proposed framework their use would be extended; meta-rules are rules that consequences are defeasible theories.

Each trigger activates one concern using a defeasible meta-rule such that

G $=\{g_1,g_2,\ldots,g_n\}$ where, $n\geq 0$ is the number crosscutting concerns in the system

$M_i : g_i \Rightarrow D_i$

When a meta-rule contains an empty bode i.e. no antecedent, it denotes the activation of the base system,

$M_0 :\Rightarrow \beta$

For the illustrative scenario, the trigger that activates the preferences concern is the flag on the user's mobile phone.

B. *The Base System β*

The base system is the actual deployed application that is governed by rules that reflect obligations; these rules are put at the design phase to achieve a certain purpose or goal. In this framework β is represented using defeasible theory, it contains rules that reason about local attributes of the system (C^u) in order to serve a certain goal. When the need arises to integrate a new contextual knowledge in the decision-making process, the designer has two options, (1) If the newly added contextual knowledge is a simple attribute that does not crosscut the base concern of the system and is issued by the same administrative authority, it can be added implicitly to the base system, either as a new rule or as a predicate in an existing rule. (2) However, if the newly added contextual knowledge serves a concern that crosscuts the base system or is issued by a different administrative authority, it will be encapsulated as a distinct concern that is formalized.

Formally, the base system is a defeasible theory denoted as $\beta(F^{\beta}, R^{\beta}, >^{\beta})$, The formal definition of β flows naturally from the definition of classical defeasible theory, however, the components of the base system theory would be superscripted with the base system name β.

C. *Distributed Contextual Concerns Theories D*

Based on the separation of concerns principle, when the collected contextual knowledge refers to a cross-cutting concern or is issued by a different administrative authority, it will be encapsulated in a distinct theory(s). This would considerably enhance the development, maintenance, and security of the overall system and can enable reaching justifiable decisions even if only partial knowledge is available.

A concern refers to the context of participating entities/authorities regarding the service provided by the base system; it reflects their interpretation of the service based on their own manipulating of internal contextual knowledge that they can access. For example, suppose an energy-saving software to control an air conditioning system in a building, its base/main concern is to manage energy consumption; it turns ACs off for uninhabited areas when the energy level exceeds a certain threshold. At the same time, the system is affected by an asset safety concern, the IT department controls the air conditioning system operation regarding the safety of certain assets in the building e.g. servers. On the other hand, the operation of the system is further influenced by the maintenance department rules that turn off ACs in case of any problems related to the hardware parts of the AC, etc.

Concerns are used to alter the behavior of the base system by applying their conclusion. It should be noted that concerns do not only affect the base system, but rather affect other concerns of the system; for example, the user's preferences in the illustrative scenario.

Concerns are represented as a set of distributed defeasible theories D. Each theory has a unique name. System components are referred to as,

Sys-c $= \{\beta\} \cup D$, where $D = \{D_1,\ldots, D_n\}$, n is the number of concerns in L

The formal definition of each theory in D flows naturally from the definition of classical defeasible theory, however, the components of each theory would be superscripted with the concern name, e.g. concern theory D_i is a tuple (F^{D_i}, R^{D_i}, $>^{D_i}$). Each concern is activated by one trigger using a meta-rule.

It should be mentioned that throughout the work of this paper, the decision inferred by *β* is called *a base conclusion*, while the decision inferred by any concern theory is superscripted with the name of the concern, e.g. $Pass^{D1}(X)$, means that according to concern D_1, the conclusion Pass(X) is inferred.

D. *Inter-Concerns Conflict Resolution λ*

Basically, concerns conclusions overrule the base system conclusion. In other words, when a query is issued for a service provided by a system that includes multiple concerns, if any of these concerns concluded a decision that contradicts the conclusion concluded by the base system, the concerns conclusion would be preferred; this is exactly where the effect of context in changing the behavior of the system, is captured.

However, in certain contexts several concerns can be activated; this might lead to conflicts in the decision-making process. This case happens when the conclusion inferred from one concern i.e. defeasible theory contradicts the conclusion inferred from another concern(s). In this case, the system would use λ, a conflict resolution mechanism that follows a prioritized ordering scheme to resolve inter-concern conflicts.

$\lambda = \{(D_i, D_j) \in \text{Sys-c}^2 \mid (D_i \sqsupset D_j)\ D_i,D_j \in D$ and

$(D_k \sqsupset \beta)\ \forall k\ D_k \in D\}$

λ is a total ordering relation that is defined over system components, it uses the operator ⊐ to denote priority, such that $D_i \sqsupset D_j$ states that the conclusion of D_i is preferred over the conclusion of D_j, and so on. However, it has another property; the definition also implies that the conclusion of any concern is preferred over the base conclusion.

It is important to notice that a total ordering relation is used instead of a partial ordering relation to prioritized concerns. Whenever a new concern is added, λ should be re-evaluated;

and the relation between the newly added concern and all other system components should be set in a proper way. It is the designer's responsibility to decide how to prioritize concerns based on the criticality level in the decision-making process.

VIII. Formal Proofs

The provability of the framework would be discussed according to the concern-level local distributed theory and the system level theory.

A. *Concern Level Proof*

Each concern is represented as a defeasible theory D, the probability of a defeasible logic is based on the concept of a derivation (or proof) from the theory [22]. A derivation is a finite sequence Pn=(P(1), . . ., P(n)) of tagged literals satisfying the following four conditions (i.e. the inference rules for each of the four kinds of conclusion).

Let P(1..i) denote the initial part of the sequence P_n of length i where i ≤ n. Then a conclusion, proved subsequently [16], could be either:

(1) Definitely provable in D.

+Δ: If P(i+1) = +ΔB then

(1) B ∈ F or

(2) (∃R1 ∈ Rs[B])(∀A∈Ant(R1): +ΔA ∈P(1..i)).

(2) Not definitely provable in D.

-Δ: If P(i+1) = -ΔB then

(∀R1 ∈ Rs[B]) (∃A∈Ant(R1): -ΔA ∈P(1..i)).

(3) Defeasibly provable in D.

+*δ*: If P(i + 1)= +*δ* B then

(1) +ΔB ∈ P (1..i) or

(2) (2.1) (∃R1∈ R[B])(∀A ∈ Ant(R1): +*δ*A ∈ P(1..i)

and (2.2) (−Δ¬B ∈ P (1..i))

and (2.3) (∀R1 ∈ R[¬B]) either

(2.3.1) ((∃A ∈ Ant(R1): −*δ*A ∈ P(1..i)) or

(2.3.2) ((∃R2 ∈ R[B]) such that (R2>R1) and

(∀A ∈Ant(R2): + *δ*A ∈P(1..i))

B is defeasibly provable from D, if either: (1) B is definitely provable or (2) use the defeasible part of D which requires: (2.1) finding a strict or defeasible rule with consequent B which can be applied, **and** (2.2) showing that ¬ B is not definitely provable **and** (2.3) counterattacking each rule that attacks the conclusion B by **either** (2.3.1) proving that the attacking rule is not defeasible proved **or** (2.3.2) finding a stronger rule that defeasible prove B.

Not defeasibly provable in D [38].

-δ: If P(i + 1)= -δB then

(1) -ΔB ∈ P (1..i) or

(2) (2.1) (∀R_1∈R[B]) (∃A ∈ Ant(R1): -*δ*A ∈ P(1..i)

or (2.2) (+Δ¬B ∈ P (1..i)) or

(2.3) (∃R2 ∈ R[¬B]) such that

(2.3.1) ((∀A ∈ Ant(R2): +*δ*A ∈ P(1..i)) and

(2.3.2) ((∀R3 ∈ R[B]) either (R2>R3) or

(∃A ∈Ant(R3) : - *δ* A ∈P(1..i)) ■

B. *System-Level Proof*

For a conclusion to be inferred from the framework it should be either strictly or defeasibly approved by the base system (when no concerns are addressed) or by a higher priority concern. Two types of tagged literals are introduced to approve/not approve a conclusion:

- +θ B, globally approved in system L, which means that there is a reasoning chain that strictly or defeasibly approves B in concern Di that is not defeated by any applicable reasoning chain of a higher priority concern Dj, where both Di and Dj ∈ Sys-c.
- -θ B, not globally approved in system L, which means that every reasoning chain that strictly or defeasibly approves B in concern Di is defeated by an applicable reasoning chain of a higher priority concern Dj, where both D_i and D_j ∈ Sys-c.

The tagged literals can be formally defined by the following proof conditions:

- Globally defeasibly provable in L:

+θ: If P(i + 1)= +θ B then

(1) (($+\Delta B^{\beta}$) or ($+\delta B^{\beta}$)) and (D = {})) or

(2) (∃M_i∈ M[+δ D_i] (∀g_i ∈ Ant(M_i): +Δg_i ∈ P(1..i)) and

(∃M_j∈ M[+δ D_j] (∀g_j ∈ Ant(M_j): +Δg_j ∈ P(1..i)) and

(($+\Delta B^{Di}$) or ($+ \delta B^{Di}$)) and either

(2.1) (∀D_j ∈ D) (($+\Delta\neg B^{Dj}$) or ($+\delta\neg B^{Dj}$)) and ($D_i \sqsupset D_j$) or (2.2) (∀D_j ∈ D) (($-\Delta\neg B^{Dj}$) or ($-\delta\neg B^{Dj}$)) and ($D_j \sqsupset D_i$)

- Globally not defeasibly provable in L

-θ: If P(i + 1)= -θ B then

(1) (($+\Delta\neg B^{\beta}$) or ($+\delta\neg B^{\beta}$)) and (D = {})) or

(2) (∃M_i∈ M[+δ D_i] (∀g_i ∈ Ant(M_i): +Δg_i ∈ P(1..i)) and

(∃M_j∈ M[+δ D_j] (∀g_j ∈ Ant(M_j): +Δg_j ∈ P(1..i)) and

(($+\Delta\neg B^{Di}$) or ($+\delta\neg B^{Di}$)) and either

(2.1) (∀D_j ∈ D) (($+\Delta B^{Dj}$) or ($+\delta B^{Dj}$)) and ($D_i \sqsupset D_j$) or

(2.2) (∀D_j ∈ D) (($-\Delta B^{Dj}$) or ($-\delta B^{Dj}$)) and ($D_j \sqsupset D_i$)■

IX. Case Study and Analysis

In this section, a formalization of the illustrative scenario would be presented and analyzed.

A. *Case Study*

The illustrative scenario's base system is represented in the proposed framework as follows:

$\beta = (F^{\beta}, R^{\beta}, >^{\beta})$,

F^{β} = {calling(X,Y), unknown(X), international(X)}

R_1^{β} :calling(X,Y) $\wedge$ in-urgent(X) $\Rightarrow$ ring(Y)

R_2^{β} :calling(X,Y) $\wedge$ international(X)$\Rightarrow$ ring(Y)

R_3^{β} :calling(X,Y) $\wedge$ unknown $\Rightarrow$ $\neg$ring(Y)

$>^{\beta} = \{(R_1^{\beta} > R_3^{\beta}), (R_2^{\beta} > R_3^{\beta}) \}$

However, Concern D_1 encodes the lecturer's preferences regarding call management. The lecturer makes his decision based on three contextual attributes, his location, schedule and his status that could be either busy or not based on the number of students that are active during the online session. Professor Sami's preferences are formalized by a contextual concern theory D_1 which is activated using meta-rule M_1 due to a flag on the user's phone.

M_1: flag(on) $\Rightarrow$ D_1

F^{D1} = {calling(X,Y), samsung-lab(Y), lecture-time(Y), busy(Y), nStudents(Y)}

R_1^{D1} :calling(X,Y) $\wedge$ samsung-lab(Y) $\Rightarrow$ $\neg$ring(Y)

R_2^{D1} :calling(X,Y) $\wedge$ lecture-time(Y) $\Rightarrow$ $\neg$ring(Y)

R_3^{D1} : nStudents(Y) <5$\Rightarrow$ $\neg$busy(Y)

R_4^{D1} : $\neg$busy(Y) $\Rightarrow$ ring(Y)

$>^{D1} = \{(R_4^{D1} > R_2^{D1}), (R_4^{D1} > R_1^{D1}) \}$

Stakeholders concern for urgent meetings is formalized by the following context theory D_2:

$D_2 = (F^{D2}, R^{D2}, >^{D2})$,

F^{D2} = {calling(X,Y), lecture-time(Y), invited(Y)}

R_1^{D2} :calling(X,Y) $\wedge$ lecture-time(Y) $\Rightarrow$ $\neg$ring(Y)

R_2^{D2} :calling(X,Y) $\wedge$ invited(Y) $\Rightarrow$ ring(Y)

$>^{D2} = \{(R_1^{D2} > R_2^{D2})\}$

When an urgent meeting is set up, the stakeholder activates an immediate indication. The meta-rule that activates this concern is M_2,

M_2: urgent-meeting $\Rightarrow$ D_2

As this concern is added by a higher authority, the dean, the designer decided to set λ as,

$\lambda = \{(D_2 \sqsupset D_1), (D_2 \sqsupset \beta), (D_1 \sqsupset \beta)\}$

B. Analysis

In this framework, Argumentation is used to analyze the conclusions of the contextual distributed theories. Argumentation is a mechanism used for tracing the reasoning process over a knowledge base that contains possibly partial and/or conflicting knowledge [39] [40]. It can be used to obtain useful conclusions from the defeasible logic theory.

Definition: Suppose D (F, R, >), is a defeasible logic theory and B is a ground literal. Arg is said to be an argument that supports the conclusion B from D, denoted by <Arg, B>, if Arg is a minimal set of defeasible rules (Arg $\subseteq$ Rd), such that:

1) B can be defeasibly derivate from (Arg $\cup$ F $\cup R_s$),

2) No pair of contradictory literals can be defeasibly derived from (Arg $\cup$ F $\cup R_s$), and

3) Arg contains no rule that contains an antecedent that is complementary to an antecedent of another rule in Arg.

With respect to analyzing the behavior of the theory in the case study using argumentation, suppose the situation when the dean asked the secretary Linda to call Professor Sami. Professor Sami has the preferences flag set on. Linda's number is in the urgent list; Sami is inside Samsung Lab and is in an active session with five students.

There are two arguments that support conflicting conclusions:

< $Arg_{1,\beta}$, $ring^{\beta}$> = ({calling(linda,sami) $\wedge$ in-urgent(linda)}, ring(sami))

< $Arg_{1,D1}$, $\neg ring^{D1}$> = ({calling(linda,sami) $\wedge$ samsung-lab(sami)}, $\neg$ring(sami))

Although argument $A_{1,\beta}$ supports the conclusion ringβ, argument $A_{1,D1}$ counterarguments $A_{1,\beta}$ i.e. it attacks its conclusion and vice versa. The global conflict resolution mechanism λ is used to support the conclusion of the higher priority theory. In this case $\lambda = \{(D1 \sqsupset \beta)\}$ and ($\neg ring^{D1}$) is globally approved.

Now, suppose the preferences flag is on and the urgent meeting indication is active, meta-rule M_1 will activate concern D_1 and meta-rule M_2 will activate concern D_2. Linda called Professor Sami. Sami is invited and he is in the Samsung-lab but he has no lecture at this time, he is giving advice to 5 students. The argumentation process would go as follows:

<$Arg_{1,\beta}$, $ring^{\beta}$> = ({calling(linda,sami) $\wedge$ in-urgent(linda)}, ring(sami))

<$Arg_{1,D1}$,$\neg ring^{D1}$> = ({calling(linda,sami) $\wedge$ samsung-lab(sami)}, $\neg$ring(sami))

<$Arg_{1,D2}$, $ring^{D2}$> = ({calling(linda,sami) $\wedge$ invited(sami)}, ring(sami))

According to D_1, (ring) cannot be inferred as it is defeasibly approved that Sami is not busy which blocks the conclusion of (ring), add to that ($\neg$ring) is defeasibly approved by R_1^{D1}. However, according to λ, $Arg_{1,D1}$ is attacked by a higher priority argument $Arg_{1,D2}$ that supports the conclusion ($ring^{D2}$). As long as the external contextual attribute (urgent-meeting) is active, ($ring^{D2}$) is globally approved.

X. Discussion

The proposed framework differs from all the approaches in literature in that it captures the effect of the different conceptual aspects of context using defeasible logic. It provides a powerful mechanism to manipulate multi-context distributed systems. It complies with the main characteristics of ubiquitous and distributed systems in providing transparency, reliability, and scalability. At the same time, it enables the evaluation of distributed decisions and produces globally justifiable conclusions. This is achieved using triggers and the relation between active concerns, unlike the bridge rules and mapping rules that were used in literature.

The consistency of the system is attained by the consistency of defeasible logic itself as for any statement; there is a proof/reasoning chain that can determine whether or not that statement holds and inconsistencies can be detected using the proof theory.

XI. Conclusions and Future Work

In this paper, we proposed a novel framework for context-awareness that can contextualize any deployed application. The framework is based on a conceptual analysis of context; it captures the behavior of contextual knowledge as it penetrates into deployed applications. It fairly simulates how the human being perceives context either as plain attributes or as concerns that need to be considered in order to make better decisions, change behavior and personalize services.

The framework is efficiently mapped to defeasible theory. It is generic, flexible and scalable. It allows the system to make justifiable decisions regardless of the number of available contextual attributes, concerns, or the number of administrative authorities that control the decision-making process. Its main strength lies in its distributed approach of reasoning and its ability to represent concerns in defeasible theories.

The analysis showed that the framework is able to capture both the contextual aspects and the concerns of different authorities in the system. The consistency in the system is attained by two levels of conflict resolution mechanisms, concern level, and system level.

The proposed extension of the defeasible theory using meta-rules improved the expressiveness power of the logic through enabling nonmonotonic reasoning over sets of defeasible theories rather than defeasible rules.

We have investigated the capabilities of the system in reasoning in environments with multiple entities that have cross-cutting concerns. Future work may exploit the flexibility of the proposed framework and its augmented power of expressing complex systems in providing personalized services i.e. entities/users that share the same concern but each one of them preserves its own right of manipulating contextual knowledge in its own way. For example, according to the scenario, more than one user has the same concern (e.g. preferences) but each of them has its own setting of preferences.

Further work would also consider investigating the capabilities of this framework by implementing it on real-world ubiquitous systems where context plays an important role.

Another aspect to be considered in contextual reasoning is the effect of context on the manipulation of the prioritizing scheme of both the classical defeasible logic and the proposed framework. We believe the management of this issue can present a magnificent step in the field of context-awareness.

References

[1] Park, I. S., Kim, W. T., & Park, Y. J. (2004, February). A ubiquitous streaming framework for multimedia broadcasting services with QoS based mobility support. In International Conference on Information Networking (pp. 65-74). Springer Berlin Heidelberg.

[2] Schilit, B. N., & Theimer, M. M. (1994). Disseminating active map information to mobile hosts. IEEE network, 8(5), 22-32.

[3] Brown, P. J., Bovey, J. D., & Chen, X. (1997). Context-aware applications: from the laboratory to the marketplace. IEEE personal communications, 4(5), 58-64.

[4] Capurso, N., Mei, B., Song, T., Cheng, X., & Yu, J. (2018). A survey on key fields of context awareness for mobile devices. Journal of Network and Computer Applications, 118, 44-60.

[5] Schmidt, A., Beigl, M., & Gellersen, H. W. (1999). There is more to context than location. Computers & Graphics, 23(6), 893-901.

[6] Gruber, T. R., Brigham, C. D., Keen, D. S., Novick, G., & Phipps, B. S. (2018). U.S. Patent No. 9,858,925. Washington, DC: U.S. Patent and Trademark Office.

[7] Dey, A. K. (2001). Understanding and using context. Personal and ubiquitous computing, 5(1), 4-7.

[8] Fischer, G. (2012). Context-aware systems: the 'right' information, at the 'right' time, in the 'right' place, in the 'right' way, to the 'right' person. In Proceedings of the International Working Conference on Advanced Visual Interfaces (pp. 287-294). ACM.

[9] Ryan, N., Pascoe, J., & Morse, D. (1999). Enhanced reality fieldwork: the context aware archaeological assistant. Bar International Series, 750, 269-274.

[10] Pollock, J. L. (1996). OSCAR: A general-purpose defeasible reasoner. Journal of applied non-classical logics, 6(1), 89-113.

[11] Moubaiddin, A., & Obeid, N. (2009). Partial information basis for agent-based collaborative dialogue. Applied Intelligence, 30(2), 142-167.

[12] Obeid, N., & Moubaiddin, A. (2010). Towards a formal model of knowledge sharing in complex systems. In Smart Information and Knowledge Management (pp. 53-82). Springer, Berlin, Heidelberg.

[13] Obeid, N., & Rao, R. B. (2010). On integrating event definition and event detection. Knowledge and information systems, 22(2), 129-158.

[14] Obeid, N. (2012). Three-Values Logic and Non-Monotonic Reasoning. COMPUTING AND INFORMATICS, 15(6), 509-530.

[15] Moubaiddin, A., & Obeid, N. (2013). On formalizing social commitments in dialogue and argumentation models using temporal defeasible logic. Knowledge and information systems, 37(2), 417-452.

[16] Sabri, K. E., & Obeid, N. (2016). A temporal defeasible logic for handling access control policies. Applied Intelligence, 44(1), 30-42.

[17] Moubaiddin, A., Salah, I., & Obeid, N. (2018). A temporal modal defeasible logic for formalizing social commitments in dialogue and argumentation models. Applied Intelligence, 48(3), 608-627.

[18] Mobaiddin, A., & Obeid, N. (2018, June). On Commitments Creation, Compliance and Violation. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 465-476). Springer, Cham.

[19] Bikakis, A. and Antoniou, G. (2010). Rule-based contextual reasoning in ambient intelligence. In International Workshop on Rules and Rule Markup Languages for the Semantic Web (pp. 74-88). Springer Berlin Heidelberg.

[20] McCarthy, J. (1987). Generality in Artificial Intelligence. Communications of the ACM, 30(12), 1030-1035.

[21] Antoniou, G., & Williams, M. A. (1997). Nonmonotonic reasoning. Mit Press.

[22] Nute, D. (2001, October). Defeasible logic. In International Conference on Applications of Prolog (pp. 151-169). Springer Berlin Heidelberg.

[23] Governatori G (2005) Representing business contracts in RuleML. International Journal of Cooperative Information Systems 14(2-3):181–216.

[24] Governatori, G., Olivieri, F., Scannapieco, S., & Cristani, M. (2012). Revision of defeasible logic preferences. arXiv preprint arXiv:1206.5833.

[25] Governatori G, Rotolo A, Padmanabhan V (2006) The cost of social agents. In: Proceedings of the AAMAS 2006, pp 513–520.

[26] Van Kranenburg, H., Bargh, M. S., Iacob, S., & Peddemors, A. (2006). A context management framework for supporting context-aware distributed applications. IEEE Communications Magazine, 44(8), 67-74.

[27] Dey AK et al. (2001) A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. Cont Aw Comput-HCI J 16:97–116.

[28] Schilit WN (1995). A context-aware system architecture for mobile distributed computing. Dissertation, Columbia University.

[29] Ranganathan, A., & Campbell, R. H. (2003). An infrastructure for context-awareness based on first order logic. Personal and Ubiquitous Computing, 7(6), 353-364.

[30] Alirezaie, M., Renoux, J., Köckemann, U., Kristoffersson, A., Karlsson, L., Blomqvist, E., ... & Loutfi, A. (2017). An ontology-based context-aware system for smart homes: E-care@ home. Sensors, 17(7), 1586.

[31] Obeid, M., Obeid, Z., Moubaiddin, A., & Obeid, N. (2019, July). Using Description Logic and Abox Abduction to Capture Medical Diagnosis. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 376-388). Springer, Cham.

[32] Uddin, I., Rakib, A., Haque, H. M. U., & Vinh, P. C. (2018). Modeling and reasoning about preference-based context-aware agents over heterogeneous knowledge sources. Mobile Networks and Applications, 23(1), 13-26.

[33] Antoniou, G., Bikakis, A., Karamolegou, A., & Papachristodoulou, N. (2006). A Context-Aware Meeting Alert Using Semantic Web and Rule Technology-Preliminary Report. Semantic Web Technology For Ubiquitous & Mobile Applications (SWUMA'06), 23.

[34] Dastani, M., Governatori, G., Rotolo, A., Song, I., & Van Der Torre, L. (2007, November). Contextual agent deliberation in defeasible logic. In Pacific Rim International Conference on Multi-Agents (pp. 98-109). Springer, Berlin, Heidelberg.

[35] Harel, D. (1979). First-order dynamic logic (Vol. 68). Berlin: Springer.

[36] Al-Zyoud, M., Salah, I., & Obeid, N. (2012, November). Towards a model of context awareness using web services. In International Conference on Computational Collective Intelligence (pp. 121-131). Springer, Berlin, Heidelberg.

[37] Musumba, G. W. and Nyongesa, H. O. (2013). Context awareness in mobile computing: A review. International Journal of Machine Learning and Applications, 2(1), 5-pages.

[38] Antoniou, G., Maher, M. J., & Billington, D. (2000). Defeasible logic versus logic programming without negation as failure. The Journal of Logic Programming, 42(1), 47-57.

[39] Moubaiddin, A., & Obeid, N. (2008). Dialogue and argumentation in multi-agent diagnosis. In New Challenges in Applied Intelligence Technologies (pp. 13-22). Springer, Berlin, Heidelberg.

[40] García, A. J., & Simari, G. R. (2014). Defeasible logic programming: Delp-servers, contextual queries, and explanations for answers. Argument & Computation, 5(1), 63-88.

Investigation of Pitch and Duration Range in Speech of Sindhi Adults for Prosody Generation Module

Shahid Ali Mahar[1], Mumtaz Hussain Mahar[2], Shahid Hussain Danwar[3], Javed Ahmed Mahar[4]
Department of Computer Science, Shah Abdul Latif University, Khairpur, Pakistan

***Abstract*—Prosody refers to structure of sound and rhythm and both are essential parts of speech processing applications. It comprises of tone, stress, intonation and rhythm. Pitch and duration are the core elements of acoustic and that information can make easy to design and development for application module. Through these two peculiarities, the prosody module can be validated. These two factors have been investigated using the sounds of Sindhi adults and presented in this paper. For the experiment and analysis, 245 male and female undergraduate students were selected as speakers belonging from five different districts of upper Sindh and categorized into groups according to their age. Particular sentences were given and recorded individually from the speakers. Afterward, these sentences segmented into words and stored in a database consisting of 1960 sounds. Thus, distance of the frequency in pitch was measured via Standard Deviation (SD). The lowest Mean SD accompanied 0.25Hz and 0.28Hz received from male and female group of district Sukkur. The highest Mean SD has measured with male and female group of district Ghotki along 0.42Hz and 0.49Hz. Generally, the pitch of female's speakers was found high in contrast to male's speaker by 0.072Hz variation.**

***Keywords*—Prosody generation; speech analysis; pitch; duration; Sindhi sounds**

I. Introduction

Sindhi Language is being spoken with various accents across the Sindh. Linguists generally divided this language in six dialects [1]. People speak the language in different accent in the same region because either they have migrated from other region or they are living in the districts adjacent to Punjab and Balochistan provinces.

Phonologically, Sindhi language is rich and has sufficient sound inventory [2]. The complex variation in accent is major cause for the less accuracy in Sindhi speech processing software applications specifically prosody generation module. To reach the maximum accuracy in software applications it is mandatory to measure the fundamental frequency of Sindhi sounds.

While speaking the Sindhi language, the variations in sound duration and pitch are normally observed with routine sounds of words but surprisingly these variations are also observed when the homographic words are spoken with the different diacritic symbols even spoken by the same adult. In Sindhi, huge number of homographic words is available which are commonly used and spoken. For instance سُرُ andسُرَ are the homographic words used as singular and plural respectively having difference in pitch frequency of 146.89Hz and 150.61Hz and in sound duration of 0.721ms and 0.566ms seconds respectively can be seen in "Fig. 1" and "Fig. 2".

During the literature review it is observed that the pitch and duration for Sindhi sounds have not been digitally analyzed at acceptable level and statistically measurement also have not been done for Sindhi sounds parameters like pitch and duration whereas deep analysis is mandatory for various speech processing application specifically prosody generation. Therefore pitch and duration is statistically measured and presented in this paper.

The development of Prosody Generation Module is the main objective of this research for which the pitch and duration ranges are the prerequisites parameters. In solitude condition of prosody, the unusual effects of prosody are complicated to reproduce and also the analysing of prosody is difficult due to the function multiplicity [3]. The prosody generation module is the mandatory components of various speech processing software applications specifically business-related Text-to-Speech systems today make use of rather unsophisticated systems, characteristically conveying a defaulting sentence accent based on the function word distinction [4].

In this study, various male and female adult inhabitants of five districts: Khairpur (K), Sukkur (S), Ghotki (G), Shikarpur (Sh) and Larkana (L) are chosen for recording the sounds to evaluate the fundamental frequencies particularly pitch and duration through which prosodic information pertaining to recorded sounds can be depicted and analyzed for further processing and development of speech based software applications [5] [6].

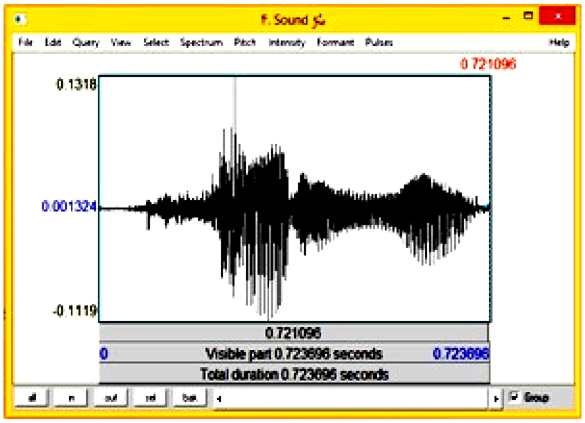

F. Sound سُرُ
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses Help
0.721096
0.1318
0.001324
-0.1119
0.721096
0
Visible part 0.723696 seconds
0.723696
Total duration 0.723696 seconds
all in out sel bak
☑ Group

Fig. 1. Sound Information of Word سُرُ using PRAAT.

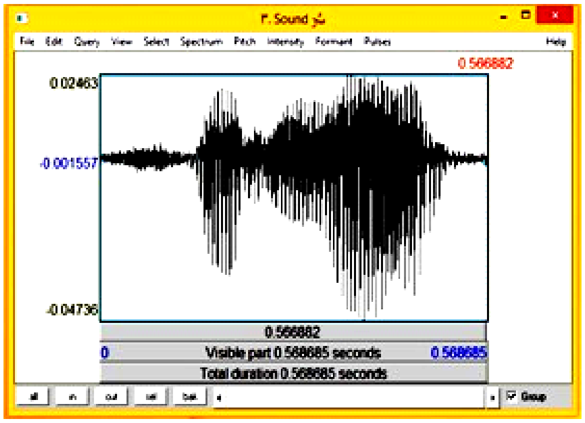

F. Sound سُرَ
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses Help
0.566882
0.02463
-0.001557
-0.04736
0.566882
0
Visible part 0.568685 seconds
0.568685
Total duration 0.568685 seconds
all in out sel bak
☑ Group

Fig. 2. Sound Information of Word سُرَ using PRAAT.

II. Related Work

The research efforts have been taken for the investigation of Sindhi phonemes. The phonological problem of Sindhi is addressed by Shaikh [7] and the observation reveals that the melody and intonation of Sindhi language in six dialects are different from each other. The accent of peoples of different dialects are compared based on the waveform visualization of image for resolving the detected problems.The Sindhi phonology is also investigated by Mahar [2] [8] along with the letter to sound conversions. The research is specifically focused to demonstrate the f0 peak of different classes of syllables in which short and long vowels are used at different positions in the words.

The research contributions pertaining to the analysis of fundamental frequency of Sindhi language are published by Abbasi [9] [10]. The role of pitch between intonation and stress is investigated. The pitch accent is analysed on the recorded sounds of 69 words and the digital experiments are performed on the recorded words having different number of syllables. The obtained results proved that stress is entirely orthogonal to F0 contours.

The acoustic analysis of Sindhi language is presented by Keerio [11] focusing the consonant sounds. The experiments were performed on the collected sounds having the VCV formats. The research is based on the liquid class of consonants and emphasis is given on the difference in the trill and lateral consonants. Moreover, the vocalic variation in vowels is analysed by Mahwish [12] considering differences among the languages spoken in Pakistan. In their research, the phonology of Sindhi is also discussed and variation is found in vowels with reference to the spoken languages of Pakistan using the PRAAT speech analyzer. Furthermore, the basic idea of digital analysis and statistical measurements of the Sindhi sounds is taken from the available literature mentioned above which will help to develop prosody generation module of Sindhi language.

III. Research Methodology

The research methodology of this work is mainly based on seven different phases. The first step is the assortment of the speaker because they are needed for recording the sounds. Therefore, 245 undergraduate students of Department of Computer Science were selected. The students are belongs to five Districts of Upper Sindh. After the selection of appropriate speakers, the sounds were recorded at the radio station, Khairpur.

The next step is the development of speech database so that 65 descriptive sentences were composed and 8 sentences were randomly given to the male and female speakers. The speech corpus is made to collect the prosodic information available in the recorded sounds. For digital investigation of the recorded sound, PRAAT speech analyzer is used to measure the pitch and duration of recorded Sindhi sounds.

The obtained results are classified into male and female duration in ms and pitch in Hz of recorded sounds. The Mean and SD of durations and pitches are also calculated to quantify the amount of variation. After that obtained results are compared and presented according to age groups and districts.

IV. Selected Speakers

Generally in experimental research, the speakers are always required for recording of voices to analyze the pitch and duration [13]. In this research, several undergraduate students from the Department of Computer Science Shah Abdul Latif University Khairpur were randomly selected for sounds recording. Our research is particularly focused on the analysis of the sounds of the people living in five districts of upper Sindh and mostly belongs to dialect Siroli. The speakers are classified in terms of gender and age groups.

Table I presents the number of speakers of a particular District with age groups. 61 speakers from district Khairpur, 53 from district Sukkur, 52 from Ghotki, 45 from Shikarpur, and 34 from Larkana are selected respectively. The selection process was entirely based on the availability of the students and the willingness of the speakers.

TABLE. I. Selected Speakers with Age Groups

Districts	**Age Group (Years)**	**No. of Speakers**
Khairpur	19	09
	20	14
	21	22
	22	16
Sukkur	19	11
	20	12
	21	15
	22	15
Ghotki	19	06
	20	09
	21	17
	22	20
Shikarpur	19	06
	20	14
	21	15
	22	10
Larkana	19	13
	20	03
	21	08
	22	10

V. Recording Procedure

The sound files are recorded at the radio station, Khairpur with a well-connected system of four components out of which three are hardware and only one is software product. The four components include a microphone, an audio console, a computer system and adobe auditions. The first component; microphone is a high frequency maintaining equipment that has the capacity of receiving sounds up to 100 decibels properly, the second component; audio console owns the capacity of maintaining, modifying and delivering clear audio to the computer system within the range of 1 to 100 decibels, the third component; computer system is a normal computer with the minimum configuration of Core2duo and the last and most important component is adobe auditions, that plays, records and saves the audio files on the computer hard disk.

VI. Speech Corpus

Actually, the speech corpus is made to gather the prosodic information residing in recorded Sindhi speech with a 16-Khz sampling rate and 16-bit encoding. For this, we have composed 65 descriptive sentences among them 8 sentences were randomly chosen and given to the male and female speakers with some prosodic boundaries for spoken at different timings. The total number of speakers are 245 among them 197 are male and 48 are females speakers. The words spoken by male speakers are 197x8 = 1576 and the words spoken by females are 48x8 = 384 so that the total number of spoken words is 1960.

In our composed sentences, the minimum length of words is one letter and the maximum length is six letters hence, for experiment and results analysis, the recorded sounds of words are classified according to number of letters used to compose the word. The sample of speech sounds in wave form is depicted in “Fig. 3”. The recorded speech was segmented and labeled with prominence values using speech analysis software tools. The segmented sounds along with letters based word classification were separately stored in a speech database.

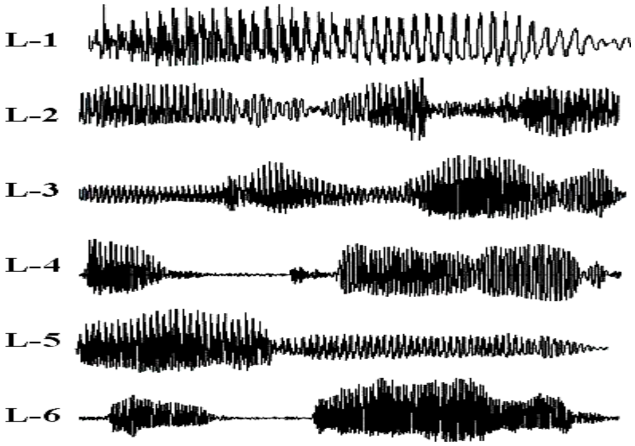


Fig. 3. Sample Wave Forms of Recorded Sounds.

VII. Speech Analyzer

Various speech analysis tools are available for investigating, analysing and restoration the audio speech signals. The PRAAT speech analyzer has great features like spectrographic, intensity measurement and formant analysis and it is freeware software. Due to these characteristics most of the researchers used this software [14]. Hence, PRAAT speech analyzer is selected for experiments and sound analysis of Sindhi adults.

The recorded sound for testing purpose is initially saved in wav format which is later loaded for evaluation. The pitch of the sound file, spectrogram and signal waves of recorded data when uploaded in the PRAAT software it has some default setting for pitch and signal waves. Duration of the word is shown in seconds in a duration bar which appeared after selecting a particular word from a sound file. In the same way to find out the pitch measurement in term of Mean pitch in Hz of any specific word.

VIII. Results

The obtained results using PRAAT are further synthesized and classified into male and female duration and pitch of recorded sounds. Due to the large number of calculations, the calculated sound durations in ms and pitches in Hz are summarized by calculating Mean value. Furthermore, set of received values were measured by the researchers with Standard Deviation (SD) to quantify the amount of variation specifically in pitch using the formula given below [15]. The same process is adopted for presenting the obtained pitch results.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bar{X}-x_i)^2}$$

The recorded sounds of sentences are separated as 1 to 6 letter(s) words to evaluate the duration and pitch of the speakers selected from the five Districts of Upper Sindh. The results are separately presented in four columns: Mean Duration Male (MDM), Mean Duration Female (MDF), Mean Pitch Male (MPM) with SD and Mean Pitch Female (MPF) with calculated SD.

Table II presents the calculated results with 1 letter words. The inferior MDM of 0.1581ms and 0.1426ms and maximum MDM of 0.1924ms and 0.1799ms are received from the male and female speakers. The lowest MPM of 156.43Hz and 158.07Hz are received. The highest MPM of 159.78HZ and 161.14Hz are obtained from the recorded sounds of male and female speakers. We have also calculated Mean SD of each District to measure the variation. The Mean SD of male and female pitches with 1 letter words is depicted in “Fig. 4”. The lowest Mean SD is received with the speakers belongs to district Sukkur and the highest Mean SD is calculated with the speakers of Ghotki.

Table III presents the calculated results of words having 2 letters. The minimum MDM of 0.2160ms and the maximum MDM of 0.2403ms are calculated. The minimum MDF of 0.1715ms and maximum MDF of 0.1906ms are received from the male and female sounds. The lowest MPM of 136.51Hz and highest MPM of 137.87Hz are measured with male sounds. The low MPF of 254.01Hz and the maximum MPF of 255.39Hz are obtained with female sounds. The Mean SD is graphically represented in “Fig. 5”. The lowest SD is obtained from the speakers of Shikarpur District and the highest SD is calculated with the speakers of Ghotki.

The pitch and duration is also calculated with the recorded sounds of male and female speakers. The calculated results of spoken words having 3 letters are presented in Table IV. The lowest MDM of 0.2974ms and 0.2309ms are calculated and greatest MDM of 0.2994ms and 0.2392ms are received.

The lowest MPM of 142.54Hz and the highest MPM of 144.83Hz are obtained from male sounds and the lowest MPF of 268.29Hz and the maximum MPF of 269.45Hz are received from female sounds. For the analysis, Mean SD of male and female pitch is calculated and presented in “Fig. 6”. The surprising results are received from this analysis because the lowest SD is calculated with the speakers of Larkana and the highest SD is calculated with the speakers of Khairpur. However, male and female speakers of district Ghotki have high pitch.

Table V presents the calculated results of 4 letter words. The lowest MDM of 0.3828ms and the highest MDM of 0.4006ms are calculated. From the female speakers, the minimum MDF of 0.3902ms and the maximum MDF of 0.4129ms are calculated from the recorded sounds. The lowest MPM of 147.46Hz and the highest MPM of 147.98Hz can be seen in the table.

The low MPF of 265.33Hz and the maximum MPF of 266.22Hz are received. “Fig. 7” shows the calculated Mean SD of the pitch of male and female speakers. The lowest SD is calculated with the speakers of Shikarpur and Larkana and the highest SD is calculated with the speakers of Khairpur. The variability in the male and female speakers is observed at low level with the speakers of Shikarpur and Larkana but similar cumulative SD is received from the speakers of both districts.

Table VI presents the calculated results of 5 letter words. The lowest MDM is 0.3952ms and the highest MDM is 0.3989ms. The minimum MDF is 0.3636ms and the maximum MDF is 0.3695ms. The lowest MPM is 127.76Hz and the highest MPM is 128.17Hz. The low MPF is 235.53Hz and the maximum MPF is 235.89Hz. The Mean SD of all calculated pitch of male and female sounds is depicted in “Fig. 8”. The lowest Mean SD is calculated with the speakers of Larkana. The Mean SD of male speakers of Khairpur is high but averagely Mean SD of male and female speakers of Sukkur are same and high. It is natural because there is small distance between both cities and peoples are usually shifted and interconnected with same business.

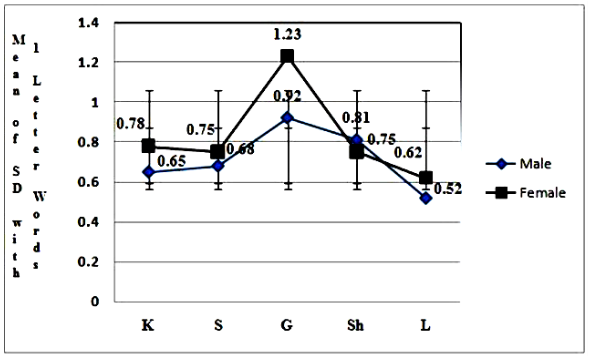


Fig. 4. Mean SD of Male and Female Pitch with 1 Letter Words.

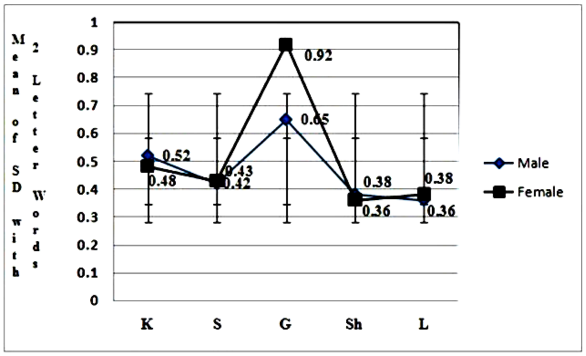


Fig. 5. Mean SD of Male and Female Pitch with 2 Letter Words.

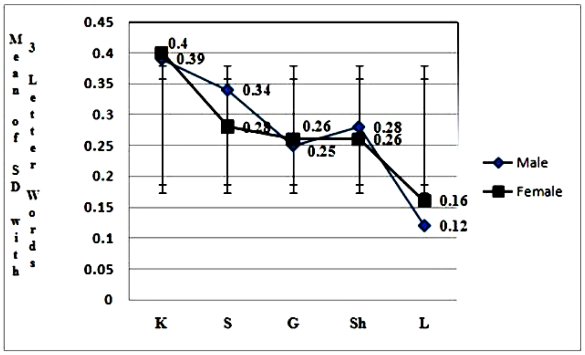


Fig. 6. Mean SD of Male and Female Pitch with 3 Letter Words.

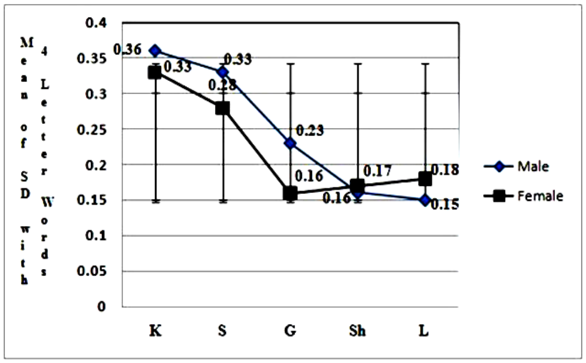


Fig. 7. Mean SD of Male and Female Pitch with 4 Letter Words.

TABLE. II. INVESTIGATED MEAN PITCH AND DURATION OF RECORDED SOUNDS OF 1 LETTER WORDS

District and	n	μDuration in ms	μ Duration in ms	μ Pitch in Hz	μ Pitch in Hz
Age Group		(Male)	(Female)	and σ (Male)	and σ(Female)
Khairpur					
19 years	09	0.1552	0.1427	156.43 (0.67345)	158.07 (0.96406)
20 years	14	0.1554	0.1428	156.49 (0.50542)	158.33 (0.92923)
21 years	22	0.1554	0.1429	156.68 (0.89033)	158.35 (0.62593)
22 years	16	0.1556	0.1429	156.73 (0.53236)	158.46 (0.59609)
Sukkur					
19 years	11	0.1551	0.1426	156.65 (0.46367)	158.09 (0.91513)
20 years	12	0.1551	0.1427	156.68 (0.84448)	158.11 (0.76362)
21 years	15	0.1552	0.1427	156.69 (0.87684)	158.12 (0.60557)
22 years	15	0.1552	0.1428	156.78 (0.51511)	158.15 (0.70495)
Ghotki					
19 years	06	0.1921	0.1797	159.34 (0.65965)	160.88 (1.38108)
20 years	09	0.1923	0.1799	159.45 (1.00263)	160.90 (1.52768)
21 years	17	0.1923	0.1799	159.63 (0.85926)	160.93 (0.97471)
22 years	20	0.1924	0.1801	159.78 (1.15661)	161.14 (1.01935)
Shikarpur					
19 years	06	0.1581	0.1577	158.07 (1.00097)	159.77 (0.53488)
20 years	14	0.1582	0.1581	158.14 (0.98074)	159.79 (0.80353)
21 years	15	0.1583	0.1582	158.29 (0.58041)	159.85 (0.81503)
22 years	10	0.1585	0.1582	158.46 (0.68147)	159.86 (0.82335)
Larkana					
19 years	13	0.1708	0.1701	158.22 (1.10646)	158.94 (0.64442)
20 years	03	0.1710	0.1702	158.25 (0.04321)	158.96 (0.35693)
21 years	08	0.1710	0.1704	158.27 (0.47053)	158.96 (0.75082)
22 years	10	0.1711	0.1704	158.31 (0.44960)	158.99 (0.70429)

TABLE. III. INVESTIGATED MEAN PITCH AND DURATION OF RECORDED SOUNDS OF 2 LETTER WORDS

District and	n	μ Duration in ms	μ Duration in ms	μ Pitch in Hz	μ Pitch in Hz
Age Group		(Male)	(Female)	and σ (Male)	and σ (Female)
Khairpur					
19 years	09	0.2161	0.1716	136.58 (0.48095)	254.05 (0.46912)
20 years	14	0.2162	0.1716	136.56 (0.53630)	254.01 (0.64991)
21 years	22	0.2164	0.1717	136.59 (0.57440)	254.12 (0.49295)
22 years	16	0.2165	0.1719	136.61 (0.47028)	254.17 (0.30033)
Sukkur					
19 years	11	0.2160	0.1715	136.51 (0.42895)	254.21 (0.31493)
20 years	12	0.2162	0.1716	136.55 (0.45345)	254.11 (0.39143)
21 years	15	0.2162	0.1717	136.56 (0.38056)	254.28 (0.53760)
22 years	15	0.2163	0.1717	136.61 (0.41055)	254.31 (0.45835)
Ghotki					
19 years	06	0.2398	0.1903	137.78 (0.69318)	255.27 (1.64697)
20 years	09	0.2399	0.1905	137.82 (0.45722)	255.28 (0.54939)
21 years	17	0.2402	0.1905	137.83 (0.56371)	255.35 (0.90806)
22 years	20	0.2403	0.1906	137.87 (0.89304)	255.39 (0.56721)
Shikarpur					
19 years	06	0.2211	0.1832	137.01 (0.34761)	254.49 (0.34137)
20 years	14	0.2211	0.1833	137.17 (0.37842)	254.33 (0.58161)
21 years	15	0.2215	0.1834	137.23 (0.41825)	254.42 (0.22334)
22 years	10	0.2216	0.1834	137.24 (0.35847)	254.52 (0.29957)
Larkana					
19 years	13	0.2189	0.1814	136.69 (0.40352)	254.17 (0.52989)
20 years	03	0.2189	0.1814	136.60 (0.17664)	254.36 (0.01415)
21 years	08	0.2191	0.1815	136.71 (0.36486)	254.40 (0.43064)
22 years	10	0.2192	0.1815	136.76 (0.49667)	254.47 (0.52972)

TABLE. IV. INVESTIGATED MEAN PITCH AND DURATION OF RECORDED SOUNDS OF 3 LETTER WORDS

District and	**n**	**µ Duration in ms**	**µ Duration in ms**	**µ Pitch in Hz**	**µ Pitch in Hz**
Age Group		**(Male)**	**(Female)**	**and σ (Male)**	**and σ (Female)**
Khairpur					
19 years	09	0.2975	0.2309	142.55 (0.52502)	268.50 (0.41492)
20 years	14	0.2977	0.2309	142.54 (0.41127)	268.49 (0.42730)
21 years	22	0.2977	0.2313	142.59 (0.24799)	268.52 (0.38579)
22 years	16	0.2978	0.2312	142.64 (0.36012)	268.51 (0.35541)
Sukkur					
19 years	11	0.2974	0.2309	143.21 (0.51085)	268.29 (0.23413)
20 years	12	0.2976	0.2311	143.22 (0.36499)	268.31 (0.37265)
21 years	15	0.2977	0.2316	143.26 (0.25099)	268.36 (0.25007)
22 years	15	0.2979	0.2321	143.29 (0.22663)	268.37 (0.26895)
Ghotki					
19 years	06	0.2991	0.2378	144.75 (0.23296)	269.41 (0.32935)
20 years	09	0.2993	0.2374	144.78 (0.21889)	269.40 (0.42042)
21 years	17	0.2993	0.2380	144.83 (0.31262)	269.42 (0.17314)
22 years	20	0.2994	0.2383	144.80 (0.24741)	269.45 (0.13532)
Shikarpur					
19 years	06	0.2983	0.2392	143.63 (0.31172)	268.85 (0.29603)
20 years	14	0.2981	0.2389	143.59 (0.25574)	268.86 (0.32054)
21 years	15	0.2984	0.2389	143.59 (0.28339)	268.86 (0.24416)
22 years	10	0.2988	0.2390	143.61 (0.25999)	268.88 (0.17135)
Larkana					
19 years	13	0.2977	0.2317	143.45 (0.14444)	268.59 (0.30207)
20 years	03	0.2975	0.2318	143.48 (0.00817)	268.62 (0.01415)
21 years	08	0.2980	0.2322	143.47 (0.10747)	268.63 (0.11554)
22 years	10	0.2983	0.2323	143.45 (0.22199)	268.60 (0.19979)

TABLE. V. INVESTIGATED MEAN PITCH AND DURATION OF RECORDED SOUNDS OF 4 LETTER WORDS

District and	n	µ Duration in ms	µ Duration in ms	µ Pitch in Hz	µ Pitch in Hz
Age Group		(Male)	(Female)	and σ (Male)	and σ (Female)
Khairpur					
19 years	09	0.3828	0.3904	147.46 (0.39525)	265.33 (0.31539)
20 years	14	0.3831	0.3902	147.53 (0.39305)	265.38 (0.32641)
21 years	22	0.3832	0.3907	147.54 (0.34606)	265.39 (0.39097)
22 years	16	0.3834	0.3906	147.57 (0.29944)	265.43 (0.29307)
Sukkur					
19 years	11	0.3832	0.3906	147.48 (0.26711)	265.36 (0.32485)
20 years	12	0.3832	0.3909	147.49 (0.35555)	265.36 (0.31182)
21 years	15	0.3837	0.3911	147.56 (0.30175)	265.38 (0.19963)
22 years	15	0.3835	0.3913	147.58 (0.39915)	265.39 (0.30115)
Ghotki					
19 years	06	0.4003	0.4126	147.95 (0.30904)	266.17 (0.16351)
20 years	09	0.3999	0.4128	147.93 (0.18997)	266.16 (0.15427)
21 years	17	0.4006	0.4128	147.97 (0.19232)	266.19 (0.14414)
22 years	20	0.4004	0.4129	147.98 (0.23106)	266.22 (0.19558)
Shikarpur					
19 years	06	0.3946	0.3952	147.59 (0.11591)	265.56 (0.17244)
20 years	14	0.3943	0.3951	147.61 (0.16562)	265.57 (0.21075)
21 years	15	0.3945	0.3954	147.63 (0.13525)	265.59 (0.17424)
22 years	10	0.3946	0.3951	147.64 (0.21977)	265.60 (0.12845)
Larkana					
19 years	13	0.3839	0.3912	147.54 (0.23933)	265.41 (0.19807)
20 years	03	0.3843	0.3909	147.56 (0.05354)	265.46 (0.19442)
21 years	08	0.3841	0.3915	147.54 (0.16477)	265.43 (0.18507)
22 years	10	0.3842	0.3918	147.53 (0.13131)	265.45 (0.12633)

TABLE. VI. INVESTIGATED MEAN PITCH AND DURATION OF RECORDED SOUNDS OF 5 LETTER WORDS

District and	n	µ Duration in ms	µ Duration in ms	µ Pitch in Hz	µ Pitch in Hz
Age Group		(Male)	(Female)	and σ (Male)	and σ (Female)
Khairpur					
19 years	09	0.3958	0.3637	127.80 (0.28794)	235.56 (0.29829)
20 years	14	0.3959	0.3639	127.78 (0.21504)	235.56 (0.06095)
21 years	22	0.3957	0.3640	127.76 (0.24179)	235.53 (0.17719)
22 years	16	0.3961	0.3641	127.78 (0.37423)	235.55 (0.10069)
Sukkur					
19 years	11	0.3952	0.3636	127.77 (0.10296)	235.58 (0.36873)
20 years	12	0.3955	0.3637	127.78 (0.21985)	235.61 (0.13479)
21 years	15	0.3956	0.3636	127.81 (0.20695)	235.62 (0.09041)
22 years	15	0.3956	0.3637	127.82 (0.30479)	235.60 (0.25177)
Ghotki					
19 years	06	0.3986	0.3689	128.17 (0.11972)	235.89 (0.07528)
20 years	09	0.3986	0.3687	128.14 (0.09877)	235.85 (0.08179)
21 years	17	0.3987	0.3688	128.16 (0.12916)	235.87 (0.14492)
22 years	20	0.3989	0.3689	128.15 (0.12045)	235.86 (0.16146)
Shikarpur					
19 years	06	0.3962	0.3695	127.87 (0.22657)	235.67 (0.13881)
20 years	14	0.3962	0.3695	127.88 (0.17217)	235.67 (0.13989)
21 years	15	0.3965	0.3694	127.89 (0.12628)	235.68 (0.08189)
22 years	10	0.3966	0.3693	127.88 (0.07362)	235.69 (0.06588)
Larkana					
19 years	13	0.3957	0.3641	127.81 (0.11046)	235.59 (0.17436)
20 years	03	0.3959	0.3645	127.80 (0.04321)	235.62 (0.07257)
21 years	08	0.3961	0.3643	127.82 (0.05196)	235.61 (0.14169)
22 years	10	0.3962	0.3644	127.79 (0.07655)	235.58 (0.13409)

TABLE. VII. INVESTIGATED MEAN PITCH AND DURATION OF RECORDED SOUNDS OF 6 LETTER WORDS

District and	n	µ Duration in ms	µ Duration in ms	µ Pitch in Hz	µ Pitch in Hz
Age Group		(Male)	(Female)	and σ (Male)	and σ (Female)
Khairpur					
19 years	09	0.5422	0.4791	150.94 (0.17349)	240.79 (0.15727)
20 years	14	0.5423	0.4794	150.94 (0.15469)	240.81 (0.15933)
21 years	22	0.5426	0.4792	150.97 (0.20631)	240.78 (0.20991)
22 years	16	0.5425	0.4792	150.95 (0.21795)	240.80 (0.21755)
Sukkur					
19 years	11	0.5426	0.4723	150.97 (0.23584)	240.82 (0.21329)
20 years	12	0.5426	0.4724	150.98 (0.22628)	240.83 (0.22329)
21 years	15	0.5429	0.4727	150.96 (0.20865)	240.83 (0.20465)
22 years	15	0.5431	0.4726	150.97 (0.22993)	240.85 (0.21357)
Ghotki					
19 years	06	0.5576	0.4853	151.62 (0.29456)	241.37 (0.21347)
20 years	09	0.5574	0.4851	151.63 (0.24536)	241.36 (0.20554)
21 years	17	0.5575	0.4854	151.66 (0.22054)	241.37 (0.22534)
22 years	20	0.5573	0.4854	151.64 (0.20722)	241.38 (0.23852)
Shikarpur					
19 years	06	0.5486	0.4792	151.13 (0.22138)	240.77 (0.21618)
20 years	14	0.5486	0.4788	151.11 (0.20167)	240.75 (0.21122)
21 years	15	0.5487	0.4793	151.14 (0.21753)	240.74 (0.21711)
22 years	10	0.5487	0.4793	151.11 (0.24956)	240.78 (0.24376)
Larkana					
19 years	13	0.5438	0.4728	150.96 (0.22847)	240.29 (0.19608)
20 years	03	0.5440	0.4728	150.95 (0.23424)	240.29 (0.16083)
21 years	08	0.5438	0.4729	150.95 (0.24546)	240.30 (0.16553)
22 years	10	0.5439	0.4727	150.97 (0.22628)	240.33 (0.23825)

The calculated results of words having 6 letters are given in Table VII. The least MDM of 0.5422ms and the highest MDM of 0.5576ms are measured. The minimum MDF is 0.4723ms and the maximum MDF is 0.4854ms. The lowest MPM is 150.94Hz and the highest MPM is 151.66Hz. The lowest MPF of 240.29Hz and largest calculated MPF of 241.38Hz are received from the recorded male and female recorded sounds. The calculated results of Mean SD are shown in “Fig. 9”. The mean SD of Khairpur shows that male and female pitch with large words are same and the peoples speak slightly. The high SD results are received with the speakers of Ghotki.

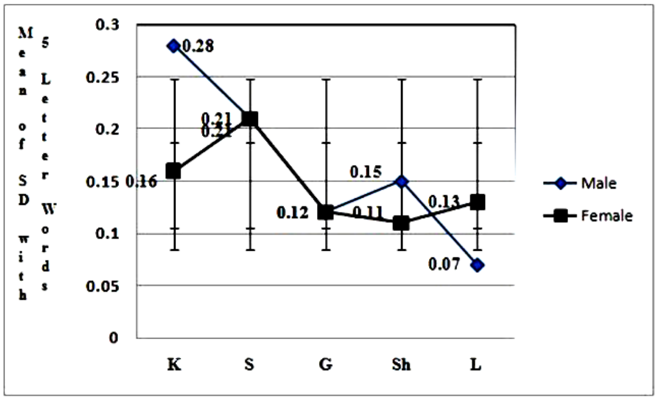


Fig. 8. Mean SD of Male and Female Pitch with 5 Letter Words.

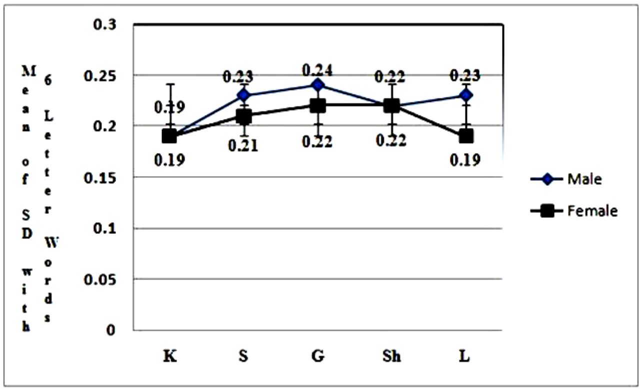


Fig. 9. Mean SD of Male and Female Pitch with 6 Letter Words.

IX. Discussion and Conclusion

According to the phonology, prosody is conceptual phenomena acquiring from the recorded speech. The deep phonetic understanding of cognitive concepts like rhythm, intonation and accentuation are the notations of pitch and duration as these both are core transmitters of the acoustic information and mandatory for designing, developing and validation of prosody module generation for Sindhi language. Hence, we have evaluated and investigated the two core parameters i.e. pitch and duration of the recorded sounds of Sindhi male and female participants and presented in this paper. The outcome of this research study will be helpful to develop efficient and effective Sindhi prosody generation module.

The sentences were given to speakers for investigation of pitch and duration from the recorded sounds. For this, 245 undergraduate students are selected from the five Districts; Khairpur, Sukkur, Ghotki, Shikarpur and Larkana of Upper Sindh having different ages. Eight sentences were given individually which are randomly selected from the prepared 65 sentence. The total words were spoken by male and female speakers comprises on 1960.

Furthermore, all the recordings are recorded at the radio station, Khairpur on different timings and due to short time only limited time was given by the authority. The recorded sounds are segmented into words and then stored into the computer. The PRAAT speech analyzer is used for segmentation of sounds of words and analysis of the recorded sounds.

The duration and pitch of the recorded sounds are separately calculated and presented according to the words based on the number of letters. The lowest MPM 127.76Hz is calculated with 5 letter words and the highest MPM of 159.78Hzis recorded with 1 letter word. Experimented results proved that the pitch in Sindhi sounds is entirely based on the syllabification. Almost, 1 letter word has one syllable while it is possible the number of syllables increases when word is based on more than 1 letters. It is also observed that the pitch of Sindhi people is high at the start of the word and at the end of the syllable particularly when syllable ends with ‘Jazm’. The cumulative Mean SD of all words spoken by Male and Female speakers is depicted in “Fig. 10”.

On the basis of received results and the calculate Mean SD of all words spoken by the male speakers it is found that speakers pertaining to district Ghotki have high pitch as compare to others. And the speakers of Larkana district have low pitch. However, little variation in pitch is observed with the male speakers of Khairpur, Sukkur and Shikarpur Districts. Approximately same results of the female speakers are received during the experiments. It is found that the pitch of the female speakers are high then the pitch of male speakers but the variation is only 0.072.

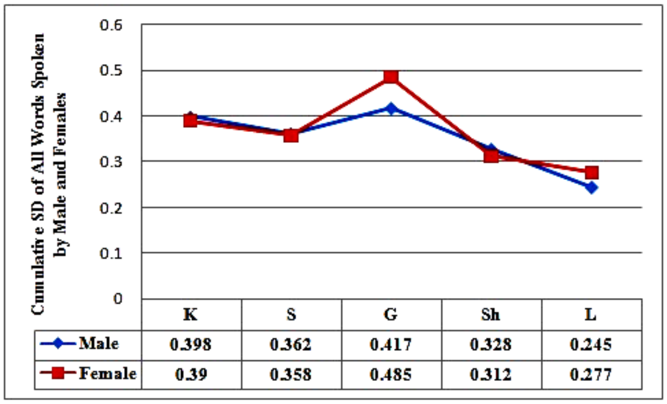


Fig. 10. Cumulative SD of Male and Female Pitch.

References

[1] Hassan, “Assimilation and Incidental Differences in Sindhi Language”, Eurasian Journal of Humanities, vol. 2, Issue 1, pp. 1-16, 2016.

[2] J. A. Mahar, G. Q. Memon, “Phonology for Sindhi Letter-to-Sound Conversion”, Journal of Information & Communication Technology, Vol. 3, No. 1, pp. 11-21, Spring 2009.

[3] S. Hoffmann, B. Pfister, “Employing Sentence Structure: Syntax Trees as Prosody Generators”, In 13th Annual Conference of the International Speech Communication Association, pp. 470-473, September 2012.

[4] P. B. Dasgupta, “Detection and Analysis of Human Emotions Through Voice and Speech Pattern Processing”, International Journal of Computer Trends and Technology, Vol. 52 No. 1, pp. 1-3, 2017.

[5] K. Waghmare, S. Kayte, B. Gawali, “Analysis of Pitch and Duration in Speech Synthesis”, Communications on Applied Electronics, Vol. 4, No. 4, pp. 10-18, February 2016.

[6] X. Sarasola, E. Navas, D. Tavarez, L. Serrano, I. Saratxaga, I. Hernaez, “Application of Pitch Derived Parameters to Speech and Monophonic Singing Classification”, Applied Science, Vol. 9, 3140, pp. 1-16, August 2019.

[7] H. Shaikh, J. A. Mahar, G. A. Malah, “Digital Investigation of Accent Variation in Sindhi Dialects”, Indian Journal of Science and Technology, Vol. 6, No.10, pp.5429-33, 2013.

[8] J. A. Mahar, G. Q. Memon, H. A. Shah, “Perception of Syllables Pitch Contour in Sindhi Language”, Proceeding of the IEEE Natural Language Processing and Knowledge Engineering, pp. 593-597, September 2009, China.

[9] A. M. Abbasi, S. Hussain, “The Role of Pitch between Stress and Intonation in Sindhi”, ELF Annual Research Journal, Vol. 17, pp. 41-54, 2015.

[10] A. M. Abbasi, H. Pathan, M. A. Channa, “Experimental Phonetics and Phonology in Indo-Aryan & European Languages”, Journal of Language and Cultural Education, Vol. 6, No. 3, pp. 21-52, 2018.

[11] A. Keerio, N. Channa, Y. A. Malkani, B. Qureshi, J. A. Chandio, (2014), “Acoustic Analysis of the Liquid Class of Consonant Sounds of Sindhi”, Sindh University Research Journal (Science Series.), Vol.46, No. 4, pp. 505-510, 2014.

[12] M. Farooq, “Acoustic Analysis of Corner Vowels in Six Indigenous Languages of Pakistan”, Journal of Research in Social Sciences, Vol. 6 No. 2, pp. 2305- 6533, 2018.

[13] M. Swain, A. Routray, P. Kabisatpathy, “Databases, Features and Classifiers for Speech Emotion Recognition: A Review”, International Journal of Speech Technology, Vol. 21, pp. 93-120, 2018.

[14] HT Lathadevi, S. P. Guggarigoudar, “Objective Acoustic Analysis and Comparison of Normal and Abnormal Voices”, Journal of Clinical and Diagnostic Research, Vol. 12, No. 12, pp. 1-4, December 2018.

[15] M. P. Gelfera, L. Sara, “Speaking Fundamental Frequency and Individual Variability in Caucasian and African American School-Age Children”, American Journal of Speech-Language Pathology, Vol. 23, pp.395-406, 2014.

Classification of C2C e-Commerce Product Images using Deep Learning Algorithm

Herdian[1], Gede Putra Kusuma[2], Suharjito[3]
Computer Science Department
BINUS Graduate Program-Master of Computer Science
Bina Nusantara University, Jakarta, Indonesia, 11480

***Abstract*—C2C (consumer-to-consumer) is a business model where two individuals transact or conduct business with each other using a platform. A consumer act as a seller put their product in a platform later will be displayed to another consumer act as a buyer. This condition encourages platform to maintain high quality product information especially image that is provided by the seller. Product images need to be relevant to the product itself. It can be controlled automatically using image classification. In this paper, we carried out a research to find out the best deep learning model in image classification for e-commerce products. A dataset of 12,500 product images is collected from various web sources to be used in training and testing process. Five models are selected and fine-tuned using a uniform hyperparameter set-up. Those hyperparameters are found by using a manual process by trying a lot of hyperparameters. The testing result from every model is presented and evaluated. The result shows that NASNetLarge yield the best performance among all evaluated models with 84% testing accuracy.**

***Keywords*—Image classification; e-commerce; product images; deep learning; hyperparameter tuning**

I. Introduction

The current adoption of e-commerce in Indonesia is high. Das et al. [1] mentioned in 2016, 78% or more than 80 million users had made online purchases. The online transactions that occur in Indonesia, one of which occurs in the C2C business model, where two individuals seller and buyer transact with each other [2].

In the C2C business model, there is a platform that mediates between sellers and buyers. The seller advertises their products on a platform, which will then be seen by the buyers. This causes weak control of the information contained in the products displayed on the platform to the buyers [2]. Product information, such as product image is an important factor for successful transaction. Several researches show that the image of the product is very important in buying interest [3], [4], [5]. Therefore, a mechanism is needed to maintain the quality of product images uploaded by the seller. An automatic approach using image classification method can be used to achieve it.

There are many algorithms that can be used for image classification. One algorithm that is currently popular to solve image classification problems is Convolutional Neural Network (CNN) [6], [7], which is one of the deep learning algorithms. Deep learning itself in recent years has received a lot of attention from researchers, communities and industry. Deep learning is able to provide excellent results for various tasks such as the traffic signs identification [8], mandarin letter writing identification [9], etc.

Deep learning implementation from scratch requires huge amount of dataset. This can be an obstacle because the huge data collection requires a lot of resource. Besides that, in various cases, the data needed is difficult to collect. Under these conditions, deep learning can still be applied by transfer learning method. Transfer learning refer to the situation where what has been learned in one setting is exploited to improve generalization in another setting [10]. In case of image classification using deep learning, this can be done by using a previously trained model that is often called the pre-trained model and fine-tuned that model using a target dataset.

The use of pre-trained models is very helpful because it can save time and costs of the training process. In this study several pre-trained models will be re-trained using fine-tuning method on a 12,500 product images dataset. Then their performances will be compared based on the testing accuracy. Although there are previous studies related to image classification using deep learning, it is relatively difficult to find study on the image classification process that addresses everything from data collection to the use of several models and comparing the results, specifically in the area of e-commerce.

The rest of this paper is organized as follows. Previous works on related topics are presented in Section II. Background theories of the pre-trained models, which are used in our research, are described in Section III. We describe our datasets in Section IV. Experiments and Results are presented in Sections V and VI, respectively. Discussion on the experimental results is presented in Section VII. Finally, we conclude our paper in Section VIII.

II. Related Work

There are several papers related to the importance of the product information's quality on e-commerce, such as [11], [12] and [13]. Those papers discuss the quality of information, which includes content accuracy, completeness, and relevance. The accuracy, which is one of the dimensions of information quality, represents the perception of consumers that the information presented for a product or other content on the platform is true [13]. One of the product information dimension that will help provide a buyer understanding of a product is image. In the process of creating product information, there may be a mistake that causes a decrease in the quality of product information. So to prevent this, a mechanism to maintain the quality of information or product images remain relevant and accurate is needed. One way is to automatically classify which images are good and which of them are not good.

Regarding the image classification itself, paper [14] provides an overview of the process of image classification. Generally, it covers image pre-processing, feature extraction, and classification. Image pre-processing is needed before an image is analysed. It can be in the form of image normalization. The feature extraction than do the image transformation to understand the image. In the end, the classification process is done to identify an image as a class from a group of classes.

Various deep learning algorithms have been used for image classification, such as MobileNet, NASNet, and DenseNet. MobileNet is a small-sized model optimized for use on mobile devices [15], [16]. Although the size of the model is small but its performance on Imagenet outperforms GoogleNet, which was the winner of the 2014 ILSVRC. There is also NASNet, which managed to match the performance of SENet, which is the winner of ILSVRC 2017 [17]. There is also DenseNet, which performed as good as ResNet, which is the winner of ILSVRC 2015 [18].

III. Pre-trained Models

Imagenet pre-trained models are used for image classification. The experiments are performed using Keras library. All models are available in the Keras library. The pre-trained models are MobileNetV1 [15], MobileNetV2 [16], NASNetMobile [17], NASNetLarge [17], and DenseNet121 [18]. Transfer learning process will be done using fine-tuning method for the 5 models. The last layers that are related to the classification layer will not be included but replaced by new layers.

A. *MobileNetV1*

MobileNet is the smallest model among the other 4 models. It contains around 4 million parameters. It has a total of 87 layers without top layers that are related to classification layers. The last layers related to classification are removed and replaced using new layers that is suitable for the category in the target dataset. For MobileNetV1 the first 75 layers will be frozen and will use 224x224 as input dimension. It's layer architecture generally consist of several block of depthwise separable convolution as shown in Fig. 1.

B. *MobileNetV2*

This is an improvement to MobileNetV1. For the ImageNet dataset, this newly MobileNet architecture improves the state of the art for wide range of performance points [16]. Using Keras library, total layer of MobileNetV2 is 155 layers without top layers. The first 135 layers will be frozen and will use 224x224 as input size. In MobileNetV2, there are two types of blocks. One is residual block with stride of 1. Another one is block with stride of 2. Those two blocks are stacked to form MobileNetV2 as shown in Fig. 2.

C. *NASNetMobile*

NASNetMobile is one of the variants of NASNet architecture for mobile platforms. Based on the Keras version 2.2.4, NASNetMobile has a total of 769 layers. It has more layers than MobileNet models. The number of layers that will be frozen is 724 and will use input dimension similar with MobileNet that is 224x224. NASNet architecture consist of several cell stacked together: normal cell and reduction cell. For NASNetMobile, every 4 normal cell stacked together followed by a reduction cell, as shown in Fig. 3.

D. *NASNetLarge*

This is the largest model from NASNet. This model achieved top-1 accuracy for ImageNet at 82.7%. This performance is similar to SENet as the winner of ILSVRC 2017. Based on Keras library, this model has 1039 layers. The first 950 layers will be frozen. This model will use 331x331 input size. NASNetLarge has similar architecture with NASNetMobile it is consist of several cells stacked together.

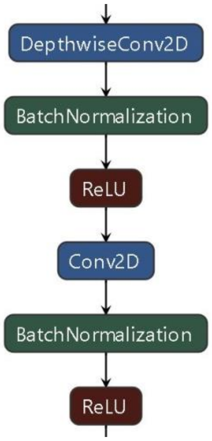


Fig. 1. Depthwise Separable Convolution Block.

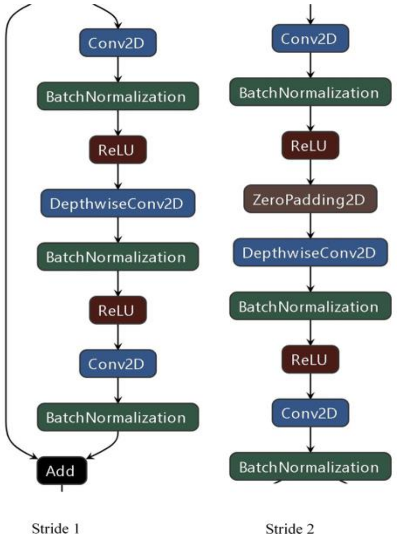


Fig. 2. MobileNetV2 Architecture Blocks.

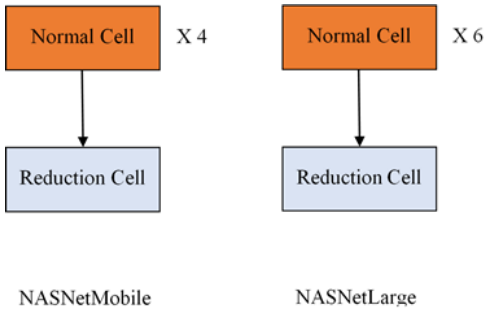


Fig. 3. NASNet Cells.

E. DenseNet121

The smallest variant of DenseNet available in Keras library is chosen. It has been shown to yield performance similar to ResNet101, but with less parameters [18]. Based on Keras library, it has a total of 427 layers. The first 411 layers will be frozen and will use input size of 224x224. DenseNet121 consist of 4 dense blocks. Every dense block consists of several convolution block. Fig. 4 shows the convolution block that is stacked together to construct a dense block.

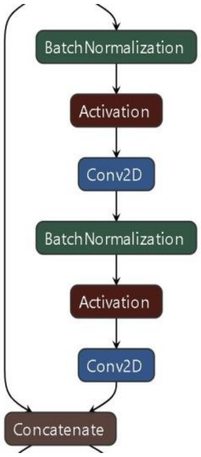


Fig. 4. Convolution Block in Densenet.

IV. Datasets

A total of 12,500 product images are collected from several C2C e-commerce websites. The collection process is done manually by visiting a web page that contains images related to a category. For example, to collect images for trouser category, a webpage containing all trouser images is visited and then by using a chrome extension all images are downloaded and saved to a local folder. This process is repeated until a category contains desired total number of images. Fig. 5 shows the process of collecting the product images for the dataset.

The dataset is organized into two sub-groups: 10.000 training data and 2.500 testing data. Each sub-group has 10 balance categories with 1.000 images for each category in training dataset and 250 images for each category in testing dataset. Validation dataset will be obtained by performing a random split on the training dataset with split ratio of 0.2. Fig. 6 shows the dataset distribution across all categories. Meanwhile, Fig. 7 shows sample images from the collected dataset.

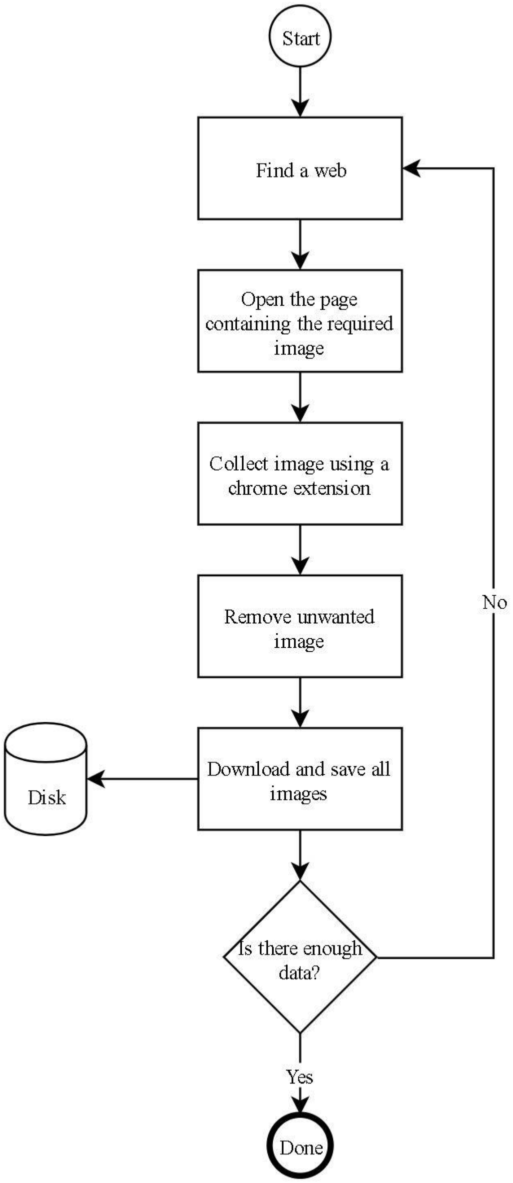

Start
Find a web
Open the page containing the required image
Collect image using a chrome extension
Remove unwanted image
Download and save all images
Disk
Is there enough data?
No
Yes
Done

Fig. 5. Images Dataset Collection Process.

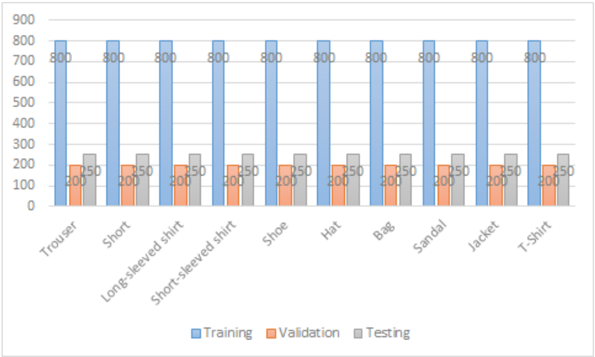


Fig. 6. The Distribution of 12,500 Product Images Dataset.

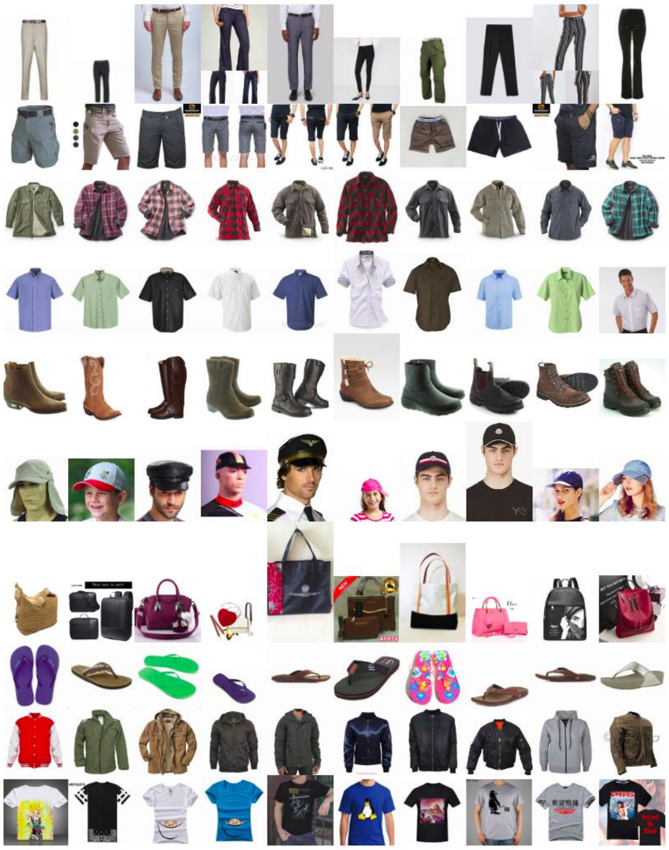


Fig. 7. Sample of 12,500 Product Images Dataset.

V. Experiments

Training for each model is done using a uniform hyperparameter set-up. Those hyperparameters is found by using a manual process by trying a lot of hyperparameter and training a model using each hyperparameter. The number of epoch is set to 100. It is chosen due to the fact that some models beginning to show overfitting at the end of 100 epoch. Also, NASNetLarge requires more time for training.

The experiment for every model is performed according to the following steps:

1) *Finding hyperparameter:* This is a set of activities to find suitable hyperparameter to train all models.

2) *Data augmentation:* This is performed to the training and validation data.

3) *Model training:* It uses Keras' fit_generator to feed sample data using batch to the model.

4) *Model testing:* The model then tested using 2.500 testing sub-group from the dataset to find out testing accuracy.

These steps are applied similarly to all 5 models. The differences are just in the input dimension and the total number of frozen layers.

A. Finding Hyperparameter

Model training is done using transfer learning. A fine-tuning process is performed for all models using the same hyperparameter. Those hyperparameter are learning rate, validation split, and the number of bottleneck layers from the original model being trained. It can be per block or per cell. Validation split is the number of validation data ratio from the training data. Meanwhile, number of bottleneck layer refers to how many layers can be re-trained from the original model architecture. These values are obtained by experimenting with several different hyperparameter values to one model and the results are compared. After the optimal values are obtained, the values are implemented for the other models. The model chosen for training to find hyperparameter values is MobileNet. This is due to the fact that MobileNet requires relatively faster training time compared to the other models.

B. Data Augmentation

The image dataset will be loaded batch by batch, then for every batch of images, data augmentation process is performed. It is used to introduce variation and noise to the model. The input image is rotated, shifted, shear, zoomed, and flipped horizontally. Augmentation process will not be performed for the testing data. Fig. 8 shows the data after augmentation process.

C. Model Training

The training process uses transfer learning method for 5 pre-trained models and uses the Tensorflow and Keras library. The 5 pre-trained models are available in the Keras library. The training process is done one-by-one for each model. The original weight of the model is used, and the classification layer of the original model is not included, so that it is possible to add new classification layer that matches the 12,500 product images dataset characteristic. Training process is performed for 100 epochs using the same hyperparameter values for all 5 models.

D. Model Testing

After the training process is complete, it continues with the testing process using 2,500 testing images from the dataset. The result of testing accuracy from this process will be evaluated and compared to find out which model has the best performance.

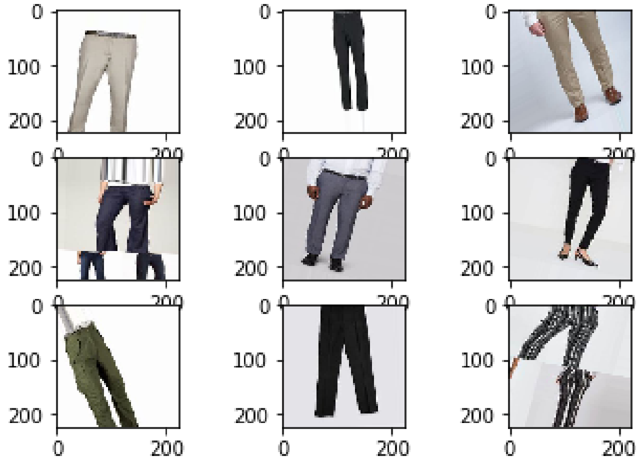


Fig. 8. Sample Data after Augmentation Process.

VI. Results

This section provides the results from every experiments, which includes validation and testing performances.

A. *Finding Hyperparameter*

The result of this process is optimal hyperparameter values for learning rate, validation split, and also number of bottleneck layer included in the training.

There are three values compared for every hyperparameter. Each value is then used to train MobileNet and the most optimal hyperparameter, based on validation accuracy, will be used to train another models. Learning rate are compared at 0.001, 0.0001, and 0.00001. The result shows that learning rate of 0.0001 gives optimal result based on validation accuracy. Validation split are evaluated at 0.20, 0.25, and 0.30. The most optimal result is validation split ratio of 0.20. Different numbers of bottleneck layer can be included in re-training process are also tested. The first experiment uses 0 layer, which means all layer from original model architecture is frozen. The second training is started from the last 1 block, and the third one is from the last 2 blocks. The result shows that the training from the last 2 blocks gives the most optimal result. It yields the highest validation accuracy after 100 epochs.

B. *MobileNetV1*

MobileNetV1 gives validation accuracy of 89.34% and testing accuracy of 82%. Fig. 9 shows the confusion matrix of the testing results for MobileNetV1.

Table I summarizes the classification report on testing dataset for MobileNetV1. As shown in the table, it gives an average precision of 79%, average recall of 77%, and average f1-score of 77%.

C. *MobileNetV2*

MobileNetV2 gives validation accuracy of 78.15% and testing accuracy of 75%. Fig. 10 shows the confusion matrix of the testing results for MobileNetV2.

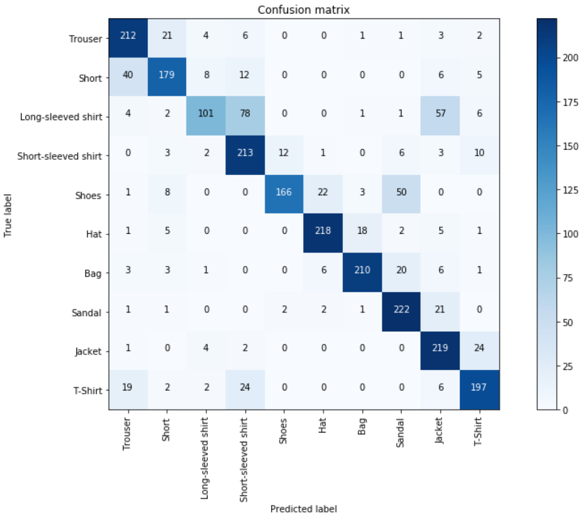


Fig. 9. Confusion Matrix for MobileNetV1.

TABLE. I. Classification Report of MobileNetV1

Category	Precision	Recall	F1-Score
Trouser	75%	85%	80%
Short	80%	72%	76%
Long-sleeved shirt	83%	40%	54%
Short-sleeved shirt	64%	85%	73%
Shoes	92%	66%	77%
Hat	88%	87%	87%
Bag	90%	84%	87%
Sandal	74%	89%	80%
Jacket	67%	88%	76%
T-Shirt	80%	79%	79%
Average:	**79%**	**77%**	**77%**

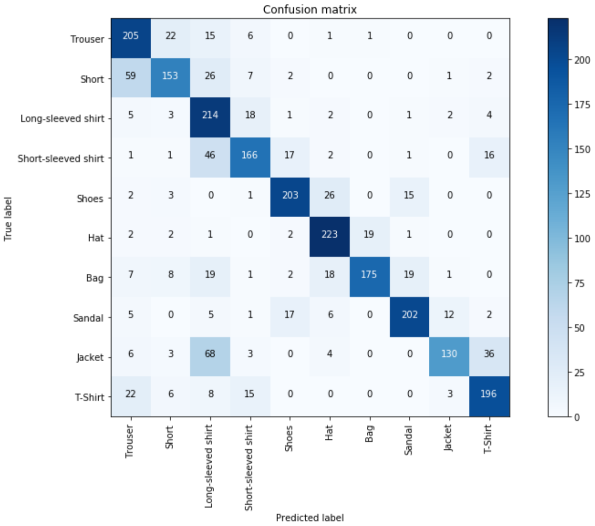


Fig. 10. Confusion Matrix for MobileNetV2.

Table II summarizes the classification report on testing dataset for MobileNetV2. As shown in the table, it gives an average precision of 77%, average recall of 75%, and average f1-score of 75%.

TABLE. II. Classification Report of MobileNetV2

Category	Precision	Recall	F1-Score
Trouser	65%	82%	73%
Short	76%	61%	68%
Long-sleeved shirt	53%	86%	66%
Short-sleeved shirt	76%	66%	71%
Shoes	83%	81%	82%
Hat	79%	89%	84%
Bag	90%	70%	79%
Sandal	85%	81%	83%
Jacket	87%	52%	65%
T-Shirt	77%	78%	77%
Average:	**77%**	**75%**	**75%**

D. NASNetMobile

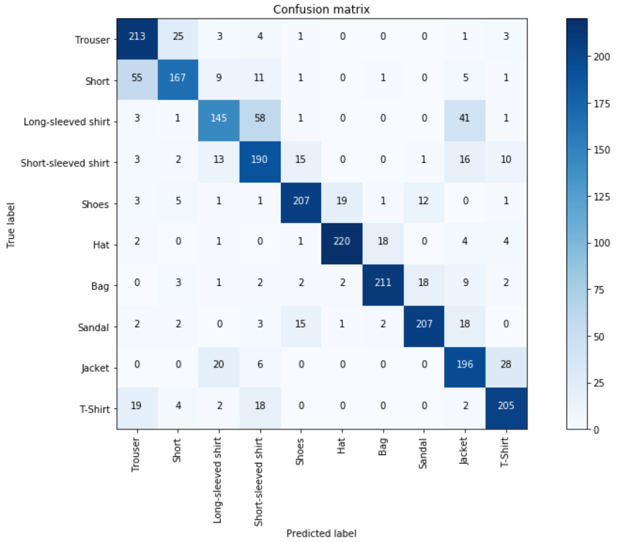
NASNetMobile gives validation accuracy of 84% and testing accuracy of 78%. Fig. 11 shows the confusion matrix of the testing results for NASNetMobile.

Table III summarizes the classification report on testing dataset for NASNetMobile. As shown in the table, it gives an average precision of 79%, average recall of 78%, and average f1-score of 78%.

E. NASNetLarge

NASNetLarge is the biggest model in term of architecture. It gives validation accuracy of 90.69% and testing accuracy of 84%. Fig. 12 shows the confusion matrix of the testing results for NASNetLarge.

Table IV summarizes the classification report on testing dataset for NASNetLarge. As shown in the table, it gives an average precision of 84%, average recall of 84%, and average f1-score of 84%.

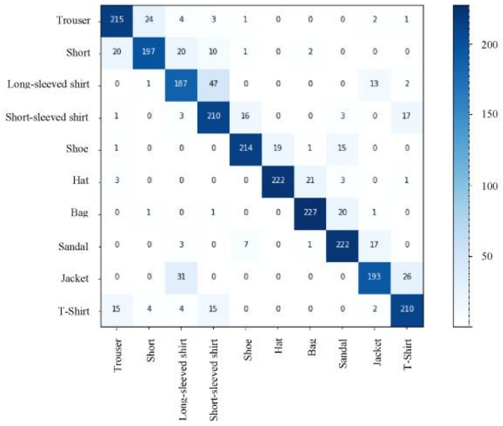


	Trouser	Short	Long-sleeved shirt	Short-sleeved shirt	Shoes	Hat	Bag	Sandal	Jacket	T-Shirt
Trouser	213	25	3	4	1	0	0	0	1	3
Short	55	167	9	11	1	0	1	0	5	1
Long-sleeved shirt	3	1	145	58	1	0	0	0	41	1
Short-sleeved shirt	3	2	13	190	15	0	0	1	16	10
Shoes	3	5	1	1	207	19	1	12	0	1
Hat	2	0	1	0	1	220	18	0	4	4
Bag	0	3	1	2	2	2	211	18	9	2
Sandal	2	2	0	3	15	1	2	207	18	0
Jacket	0	0	20	6	0	0	0	0	196	28
T-Shirt	19	4	2	18	0	0	0	0	2	205

Fig. 11. Confusion Matrix for NASNetMobile.

TABLE. III. CLASSIFICATION REPORT OF NASNETMOBILE

Category	**Precision**	**Recall**	**F1-Score**
Trouser	71%	85%	77%
Short	80%	67%	73%
Long-sleeved shirt	74%	58%	65%
Short-sleeved shirt	65%	76%	70%
Shoes	85%	83%	84%
Hat	91%	88%	89%
Bag	91%	84%	87%
Sandal	87%	83%	85%
Jacket	67%	78%	72%
T-Shirt	80%	82%	81%
Average:	**79%**	**78%**	**78%**



	Trouser	Short	Long-sleeved shirt	Short-sleeved shirt	Shoe	Hat	Bag	Sandal	Jacket	T-Shirt
Trouser	215	24	4	3	1	0	0	0	2	1
Short	20	197	20	10	1	0	2	0	0	0
Long-sleeved shirt	0	1	187	47	0	0	0	0	13	2
Short-sleeved shirt	1	0	3	210	16	0	0	3	0	17
Shoe	1	0	0	0	214	19	1	15	0	0
Hat	3	0	0	0	0	222	21	3	0	1
Bag	0	1	0	1	0	0	227	20	1	0
Sandal	0	0	3	0	7	0	1	222	17	0
Jacket	0	0	31	0	0	0	0	0	193	26
T-Shirt	15	4	4	15	0	0	0	0	2	210

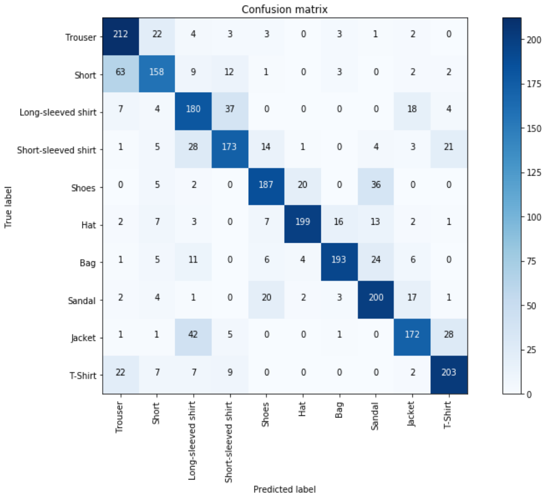
Fig. 12. Confusion Matrix for NASNetLarge.

TABLE. IV. CLASSIFICATION REPORT OF NASNETLARGE

Category	**Precision**	**Recall**	**F1-Score**
Trouser	84%	86%	85%
Short	87%	79%	83%
Long-sleeved shirt	74%	75%	75%
Short-sleeved shirt	73%	84%	78%
Shoes	90%	86%	88%
Hat	92%	89%	90%
Bag	90%	91%	90%
Sandal	84%	89%	87%
Jacket	85%	77%	81%
T-Shirt	82%	84%	83%
Average:	**84%**	**84%**	**84%**

F. DenseNet121

DenseNet121 is the smallest variant from DenseNet model. It is comparable to other mobile models. DenseNet121 gives validation accuracy of 84.09% and testing accuracy of 75%. Fig. 13 shows the confusion matrix of the testing results for DenseNet121.



	Trouser	Short	Long-sleeved shirt	Short-sleeved shirt	Shoes	Hat	Bag	Sandal	Jacket	T-Shirt
Trouser	212	22	4	3	3	0	3	1	2	0
Short	63	158	9	12	1	0	3	0	2	2
Long-sleeved shirt	7	4	180	37	0	0	0	0	18	4
Short-sleeved shirt	1	5	28	173	14	1	0	4	3	21
Shoes	0	5	2	0	187	20	0	36	0	0
Hat	2	7	3	0	7	199	16	13	2	1
Bag	1	5	11	0	6	4	193	24	6	0
Sandal	2	4	1	0	20	2	3	200	17	1
Jacket	1	1	42	5	0	0	1	0	172	28
T-Shirt	22	7	7	9	0	0	0	0	2	203

Fig. 13. Confusion Matrix for DenseNet121.

Table V summarizes the classification report on testing dataset for DenseNet121. As shown in the table, it gives an average precision of 76%, average recall of 75%, and average f1-score of 75%.

Based on the experimental results of the 5 pre-trained models above, we can see that NASNetLarge has the best performance with testing accuracy of 84%. NASNetLarge also gives highest precision, recall, and f1-score at 84%. The validation and testing performances of the 5 evaluated models are summarized in Table VI.

During training, at the last epoch, NASNetLarge also yields the lowest validation loss among all evaluated models at 0.2910 and the highest validation accuracy of 90.69%. The NASNetLarge results have correlation with its original performance on ImageNet. It is the best among the evaluated models in this paper. It is also yield jointly best performance with winner of ILSVRC 2017.

Fig. 14 shows the validation accuracy graphs and Fig. 15 shows the validation loss graph for all 5 models. From these figures, we can see that NASNetLarge performance is also better than the other models during training. It consistently achieves the highest validation accuracy and also the lowest validation loss at every epoch. However, the main drawback of NASNetLarge is its training time. It requires much more time to train compared to the other models.

TABLE. V. CLASSIFICATION REPORT OF DENSENET121

Category	**Precision**	**Recall**	**F1-Score**
Trouser	68%	85%	76%
Short	72%	63%	68%
Long-sleeved shirt	63%	72%	67%
Short-sleeved shirt	72%	69%	71%
Shoes	79%	75%	77%
Hat	88%	80%	84%
Bag	88%	77%	82%
Sandal	72%	80%	76%
Jacket	77%	69%	73%
T-Shirt	78%	81%	80%
Average:	**76%**	**75%**	**75%**

TABLE. VI. MODELS PERFORMANCE RESULTS

No.	**Models**	**Validation Accuracy**	**Validation Loss**	**Testing Accuracy**
1	NASNetLarge	90.69%	0.2910	84%
2	MobileNetV1	89.34%	0.3121	82%
3	NASNetMobile	84.00%	0.4653	78%
4	DenseNet121	84.09%	0.5373	75%
5	MobileNetV2	78.15%	0.6939	75%

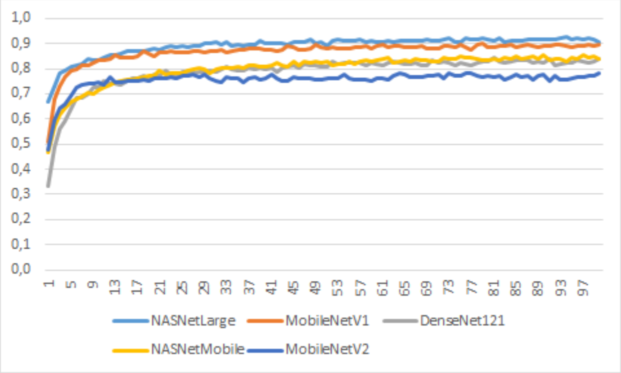


Fig. 14. Validation Accuracy for 100 Epochs.

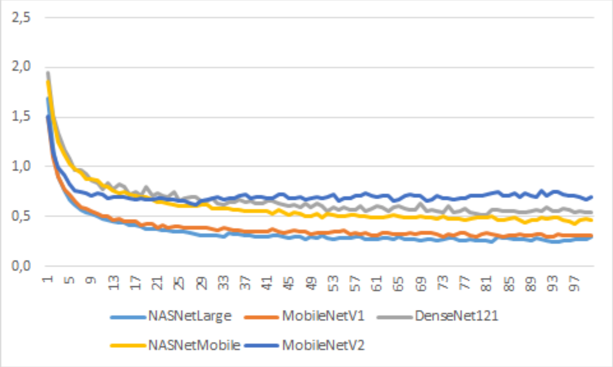


Fig. 15. Validation Loss for 100 Epochs.

VII. DISCUSSION

A clothing dataset has been collected from various sources to be used for experiments on 5 models, which consist of training and testing processes. The training process is carried out to obtain an optimal model. It will then be used in the testing process to find out how the model's performance against the clothing dataset.

The classification performance of each model for the long-sleeved shirt class is the lowest compared to other classes. Many images of long-sleeved shirts are incorrectly classified as short-sleeved shirt or jacket. This can be caused by the three classes having similar characteristics. Several images of long-sleeved shirts and short-sleeved shirts are folded. Furthermore, long-sleeved shirts and jackets have the same characteristic, which is long sleeved.

NASNetLarge provided the best performance on ImageNet according to [17]. This also matches the experimental results in this study. NASNetLarge provides the best performance compared to the other 4 models. Another finding in this study is that MobileNetV1 provides a fairly good performance, which is ranked second under NASNetLarge, but with a much smaller model size [15]. This model can be considered during the implementation process due to its light weight.

VIII. Conclusions

In this work, e-commerce product images classification has been demonstrated using deep learning algorithm. We have collected and labelled a total of 12,500 product images dataset. The images were crawled form several C2C e-commerce websites. Five deep learning models have been evaluated on the dataset, which include MobileNetV1, MobileNetV2, NASNetMobile, NASNetLarge, and DenseNet121.

Based on the experimental results, NASNetLarge achieves the best performance for image classification with testing accuracy of 84%. Also, it shows the best performance during training with validation accuracy of 90.69%. It outperforms the other four models that are trained using similar hyperparameter. However, this performance comes with a cost of larger architecture and longer training time compared to the other models.

Further research related to this study can be done by making variations to the dataset, using imbalanced dataset, or increasing the number of images. Another research can also be done by conducting experiments using different sets of hyperparameters.

Acknowledgment

This work was supported by the Directorate General of Strengthening for Research and Development, Ministry of Research, Technology, and Higher Education, Republic of Indonesia, through research grant number 23/AKM/MONOPNT/2019.

References

[1] K. Das, M. Gryseels, P. Sudhir, and K. T. Tan, “Unlocking Indonesia’s digital opportunity,” no. October, pp. 1–28, 2016.

[2] C. Dan, “Consumer-To-Consumer (C2C) Electronic Commerce: The Recent Picture,” Int. J. Networks Commun, 2014.

[3] X. Li, M. Wang, and Y. Chen, “the Impact of Product Photo on Online Consumer Purchase Intention: an Image-Processing Enabled Empirical Study,” PACIS 2014 Proc., 2014.

[4] C. Lumb, “The Customer Decision Process and User Interaction in E-commerce,” 2014.

[5] E. Huang and C.-C. Liu, “A Study on Trust Building and Its Derived Value in C2C E-Commerce,” J. Glob. Bus. Manag, 2010.

[6] A. KrizhKrizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances In Neural Information Processing Systems, 1–9.evsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Adv. Neural Inf. Process. Syst., 2012.

[7] J. Patterson and A. Gibson, Deep Learning A Practitioner’s Approach, 1st ed. boston: oreilly, 2016.

[8] D. Cireşan, U. Meier, and J. Schmidhuber, “A committee of neural networks for traffic sign classification,” in Proceedings of the International Joint Conference on Neural Networks, 2011.

[9] D. Ciresan, U. Meier, and J. Schmidhuber, “Multi-column Deep Neural Networks for Image Classification,” Int. Conf. Pattern Recognit, 2012.

[10] A. C. Ian Goodfellow, Yoshua Bengio, Deep Learning. 2016.

[11] T. Singh Chhikara, “Information Quality -Crucial Aspect of E-Commerce,” IOSR J. VLSI Signal Process. Ver. II, 2015.

[12] W. H. DeLone and E. R. Mclean, “The DeLone and McLean Model of Information Systems Success: A Ten-Year Update,” J. Manag. Inf. Syst. / Spring, vol. 19, no. 4, pp. 9–30, 2003.

[13] B. H. Wixom and P. A. Todd, “A Theoretical Integration of User Satisfaction and Technology Acceptance,” Info. Sys. Res., 2005.

[14] P. Kamavisdar, S. Saluja, and S. Agrawal, “A survey on image classification approaches and techniques,” Int. J. Adv. …, 2013.

[15] A. G. Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” in arxiv, 2017.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018.

[17] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, “Learning Transferable Architectures for Scalable Image Recognition,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018.

[18] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017.

Design and Learning Effectiveness Evaluation of Gamification in e-Learning Systems

Mohammad T. Alshammari
College of Computer Science and Engineering
University of Hail, Hail, Saudi Arabia

***Abstract*—This paper proposes a gamification design model that can be used to design and develop gamified e-learning systems. Furthermore, a controlled and carefully designed experimental evaluation in terms of learning effectiveness of gamification is offered. The experiment was conducted with 44 participants randomly assigned to an experimental 'gamified' condition and a controlled 'non-gamified' condition. In both conditions the same learning material, to teach computer security, were used. The main difference between the two conditions was the integration of gamification in an e-learning system designed based on the proposed model. The results indicate that learning using the gamified version of the e-learning system produces better short-term and medium-term learning gain than learning using the non-gamified e-learning version. Future avenues of research are also provided.**

***Keywords*—Gamification; e-learning systems; interaction design; experimental evaluation**

I. Introduction

Learning is a complex issue that can be influenced by different factors including learner characteristics, learning content, learning environment and teaching style. Traditional learning approaches through classrooms can be rigid and unattractive to some learners given the time and place constraints. In contrast, e-learning systems provide the ability for learners to learn anytime, anywhere and offer different interactivity levels that may not well be supported by traditional learning approaches. Moreover, learning content can be incorporated in e-learning systems with different types such as examples, simulations, problem-solving tasks and explanations. These types can also be offered in different multimedia formats including written and spoken material in addition to videos and games.

Motivating learners to use such systems in order to enhance their learning and gain some knowledge is a challenging task [1]. Gamification is usually put forward as a proposed solution to motivate, engage and support learners through their interaction with learning material [2]. The term 'gamification' was firstly introduced in the scientific community in 2010 [3]. Since then, it is still widely adopted by researchers in different research domains and areas including learning, business and health [4]. Gamification can be defined as the use of game elements, thinking and mechanisms in a non-game context [3].

E-learning systems can be designed to integrate different game elements such as points, badges, progress bars and levels to support learners in having a gamified learning experience [5]. Gamification in learning should be differentiated form the concept of game-based learning which principally aims at delivering learning content and objectives in an entertaining approach implemented as a game so that learners can play the game in order to learn.

Based on the literature and recent reviews on gamification [4], [14], [15], several issues have been found. First, most gamification studies were not supported with empirical evidence following well-designed, controlled and thorough experimental evaluations that have control conditions. Second, mixed results can also be found generally with limited sample sizes. Third, studies were mostly based on short-term applications of some types of game elements. Fourth, some studies do not rely on such models as a foundation to the design and development of gamified e-learning systems. This issue can be challenging when other researchers need to replicate their work. Furthermore, gamification is still in its infancy and quickly developing requiring more focused and thorough studies [14].

These issues need to be carefully addressed when applying gamification in learning. This paper bridges this gap by offering a gamification design model that can be used to develop different instances of gamified e-learning systems. The model is consisted of three major components including the courseware module, the learner profile and the gamification component. The paper also offers an evaluation of gamification in an e-learning system, developed based on the proposed model. A carefully designed and controlled experimental evaluation was conducted with 44 participants in a real learning environment. The participants used the developed gamified e-learning system to learn some concepts related to a course on computer security, the application domain. This study primarily aims to explore the learning effectiveness of the integration of gamification in e-learning systems. The main research question of this study is that does the integration of gamification in e-learning systems enhance learning?

The rest of the paper is structured as follows. Section II provides related work to gamification in e-learning systems. Section III details the proposed gamification design model. Section IV outlines the evaluation method. Section V offers the results. Section VI concludes the paper highlighting the main findings. Section VII points out to future directions of research.

II. Related Work

Many attempts have been proposed to investigate gamification in online-learning research. For instance, a blended study (classroom and online) conducted by [6] to evaluate the learning effect of gamification yielding positive results. In [7], design phases were proposed to integrate gamification in e-learning systems supported with some results related mainly to learning outcomes. Gamified quizzes besides lectures were also offered in [8]; there were positive results in terms of perceived usefulness of gamification based on a survey. Yet, quizzes can represent one aspect of learning so that they have not focused on the whole learning process. A gamified e-learning system to teach Java programming language was also developed and evaluated in terms of learning gain and motivation [9]. However, their results cannot be generalized since their study was applied only to first year undergraduate students.

Moreover, research in [10] compared different forms of gamification including competitive, collaborative and adaptive gamified learning activities concluding that gamification can be beneficial in learning when diverse game elements are incorporated in e-learning systems. However, their study used four experimental conditions where each condition was lasted for a very short learning time affecting the controllability and feasibility of their experiment. Research in [11] investigated gamification based on feedback as a single game element yielding negative results; they argued that the integration and arrangements of various game elements can be more valuable in learning. Similarly, the work in [12] investigated the effect of badges as the key game element on motivation and performance of learners without yielding positive findings. The effort made by [13] focused on gamifying learning activities to learn some concepts of C-programming language producing encouraging results in terms of learner achievements and engagement. However, the work of [13] was based on a very small size so that their results cannot be generalized and further studies with more participants are needed. Accordingly, more careful studies are needed to investigate the effect of gamification in e-learning systems which justify the work presented in this paper.

III. Gamification Design Model

It is important to highlight how gamification can be designed in e-learning systems before attempting to investigate their effect on learning [3], [7]. Therefore, a gamification design model is proposed and depicted in Fig. 1. The model is consisted of a number of components. It mainly includes the gamification component, the courseware module and the learner profile. Information stored in these components can be retrieved by the gamification business logic in order to generate the gamified learning experience, and then deliver it to the learner. The output also feeds into the learner profile for updates.

In the proposed model, there are also three stakeholders: learner, instructor and admin. The learner is the person who is conducting some tasks in order to gain some knowledge and understanding of something. The instructor is the person who facilitates the creation of such courses (objectives, units, lessons and assessments) to support learners. The admin can manage the implemented model by handling the technological perspectives, and by adding gamification elements. The admin can also be enabled to play the role of an instructor in the model.

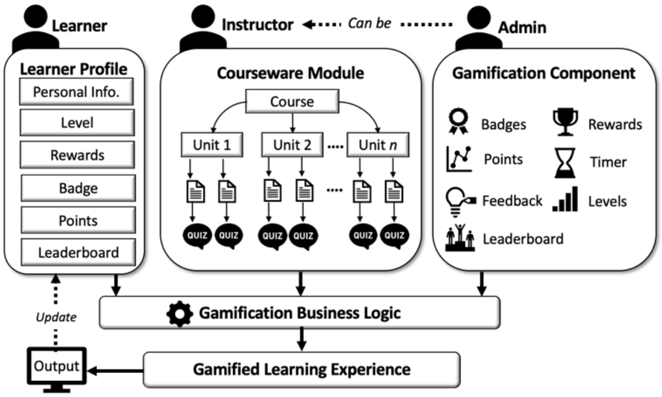


Fig. 1. The Gamification Design Model.

A. Gamification Component

The gamification component aims to represent, manage and maintain the gamification elements by system admins. There are seven main game elements in the gamification component including badges, rewards, points, timer, feedback, levels and leaderboard because they are widely used in related work [1]. Each element is described and presented in Table I.

The model, when implemented, can be flexible in a way that enables admins to create and maintain different information related to each element easily. For example, the admin can create a number of levels, name each level and then store them in the system such as beginner level, intermediate level and advanced level. Another example, badges can be created to unlimited number assigning a picture and a title to each one. Different rules and requirements can be applied to all the elements in the gamification component.

It is important to note that these game elements can be created and maintained in isolation from other components of the model. Though, they can easily be assigned to each learning activity. This is to allow instructors to mainly focus on the creation of courses and learning activities at the first place. Then, they can map the game elements to learning activities by selecting suitable ones as preferred. Moreover, not all elements are necessarily be activated in the system but rather enabling instructors to select elements that are expected to enhance learning and motivation and learners.

TABLE. I. Description of Gamife Elements

Game Element	Description
Badges	A distinctive emblem awarded when accomplishing specified learning tasks.
Rewards	A stimulus given based upon successful completion of a particular learning task.
Points	Points are incremented based on the learner progress toward learning.
Timer	A set of timers created to be mapped to learning lessons and tests.
Feedback	The provision of corrective information for learning misconception based on the results of tests.
Levels	A position of the learner on a pre-defined scale based upon the learning progress.
Leaderboard	A score board presenting the names and current points of the leading learners.

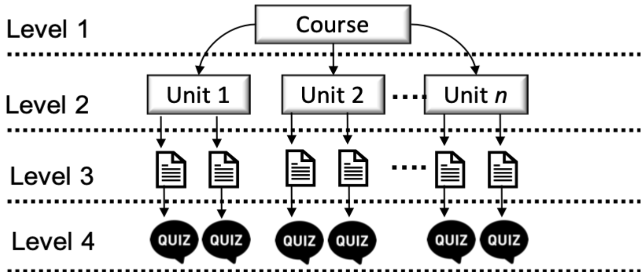


Fig. 2. The Courseware Module Structure with Four Levels. Level 1 Contains the Course Node. Level 2 has a Number of Learning units. Level 3 Involves Some Learning Lessons Associated with Related Learning units. Level 4 is consisted of a Number of Quizzes, and Each Quiz can be Linked to a Particular Learning Lesson.

B. *Courseware Module*

The courseware module is an important component in the gamification design model. Its main purpose is to represent, store and maintain learning material related to a specific course. It uses a hierarchical representation in a tree-like structure with four levels as shown in Fig. 2.

The root or level 1 represents the course node. It contains information or metadata related to the course such as the title, description, keywords and objectives to describe the course. Level 2, which entirely connected to the course node, is consisted of a number of learning units. Each unit node has some metadata to define its properties the same as the course node. A number of learning lessons can be linked to specific learning units. These lessons are located in level 3 and have additional metadata such as pre-requisites, type and multimedia used. Quizzes can also be created and then explicitly be mapped to specific learning lessons in level 4.

It is important to highlight the point that the hierarchical structure representation of the courseware module can enable instructors to author learning content in a more organized way without thinking of how to integrate gamification through the learning content. This particular representation is helpful as learning activities can automatically or manually be mapped to game elements. In addition, it is possible to generate distinctive courses related to different learning areas supporting the generalization of the proposed model and its components.

C. *Learner Profile*

According to [16], a learner profile is “what enables a system to care about a student”. The learner profile represents learner characteristics and maintains the learner level, badges, rewards, leaderboard and points. The information stored in the learner profile is continually updated as the learner progresses through learning using the gamified e-learning system. For example, when a learner successfully studies a lesson and completes a quiz related to that lesson, some points can be added to the learner profile depending on the quiz score and time spent. Accordingly, when a set of quizzes are completed by a learner, other gamification elements can be updated and then stored in the learner profile.

IV. Method

This section details the methodology used to evaluate the effectiveness of gamification on learning. A gamified e-learning system was developed taken into account the proposed gamification design model to answer the main research question. Fig. 3 shows a screenshot example of the gamified system. This example shows the profile of the learner highlighting the rank of the learner among his/her peers, the number of awards obtained, the number of points earned, the medal achieved which is a silver medal and the level of the student (i.e., a beginner level). The main purpose of developing the system is to validate the proposed model, and to evaluate the learning effectiveness of gamification through a controlled experiment. All the seven game elements in the model were incorporated in the system to offer a complete gamified learning experience to learners.

A pilot testing was conducted in two sessions prior to the main controlled experiment with 10 participants (5 students, 3 instructors and 2 interaction design experts) to learn some topics related to computer security (i.e., the application domain of the system) and to walkthrough the implemented system. This is to mimic the actual experiment, to find out issues that may occur and to ensure that the system is working properly. All participants in the pilot testing contributed and commented on some ways to improve the system, the measurement tools and learning material. By the pilot testing, it was also possible to indicate the time needed to complete the experiment. Some amendments were made before conducting the main experiment according to the findings and observations of the pilot test.

A. *Hypotheses*

Two main hypotheses are proposed for this study as follows:

- Hypothesis (H1). Learning using the gamified version of the e-learning system produces better (short-term) learning gain than learning using the non-gamified version.
- Hypothesis (H2). Learning using the gamified version of the e-learning system produces better (medium-term) learning gain than learning using the non-gamified version.

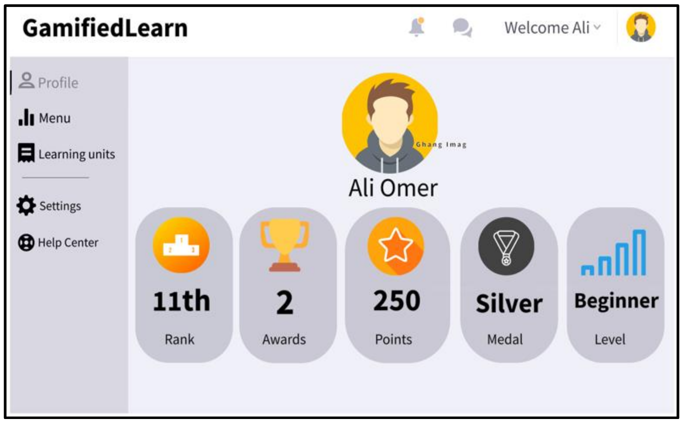


Fig. 3. A Screenshot Example of the Gamified e-Learning System. this Example Shows the Profile of The Learner Highlighting the Rank of the Learner Among his/her Peers, the Number of Awards Obtained, the Number of Points Earned, the Medal Achieved which is a Silver Medal and the Level of the Student (Beginner).

B. Measurement Tools

Learning gain was evaluated using three main tests: pre-test, post-test and follow-up test. Two types of learning gain are measured based on the tests. First, short-term learning gain is measured as the difference between the post-test and the pre-test in order to evaluate the immediate learning effectiveness of gamification. Second, medium-term learning gain is calculated as the difference between the follow-up test and the pre-test in order to evaluate the sustained knowledge. Each test involves 25 items related to a course on computer security. Each item has five multiple-choices with the fifth being '*I don't know!*' in order to reduce the chance of random speculating by participants.

It should be noted that all tests are similar except for the ordering, arrangements and phrasing of both the items and the responses. A Cronbach alpha was also measured for the pre-test, post-test and follow-up test having good reliability as 0.96, 0.82 and 0.83, respectively.

C. Experimental Procedure

Participants were first welcomed and introduced to the process of the experiment after signing consent forms of participation. Then, all participants were randomly assigned to one of two conditions: (A) a gamified version of the e-learning system (experimental) or (B) a non-gamified version (control). Participants in both A and B completed the pre-test, and then studied some learning material related to computer security using the system according to their assigned condition. The learning material in both conditions are similar and the difference is whether students were exposed to gamification or not based on their assigned condition.

The learning task was completed in a computer laboratory for three weeks where there were two experimental sessions per week. Each session was lasted about 120-180 minutes. At the end of all the experimental sessions, participants were immediately directed to complete the post-test. Two to three weeks after completing the main experiments, all participants were invited to only complete the follow-up test.

D. Data Analysis

The data of the experiments were collected by the system automatically and stored in a database. Then, they were converted into a format that enables for the data analysis process. Minitab 19 software was used for data analysis. The type of the tests performed is outlined when presenting the results in the following section.

V. Results

A number of 44 Male undergraduate students form the College of Computer Science and Engineering, University of Hail, Saudi Arabia completed the experiments. The gender variable was controlled to eliminate any variance and confounding effect on the results. The conditions being experimental 'gamified' and control 'non-gamified' were balanced having 22 participants each.

There were five variables related to learning gain (pre-test, post-test, follow-up test, short-term learning gain and medium-term learning gain) in the experiment. Fig. 4 summarizes the mean score results for those variables.

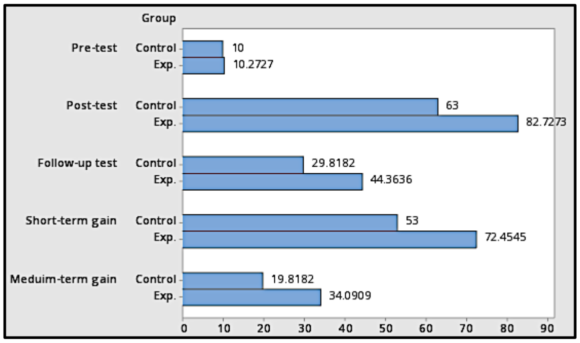


Fig. 4. Summary for the Mean Scores of the Experimental Variables.

Regarding the pre-test results, an independent sample t-test was conducted to compare the group who interacted with the gamified version of the e-learning system and the non-gamified version. There was no statistically significant difference in the scores for the gamified group (M=10.27, SD=3.92) and the non-gamified group (M=10, SD=4.58); t(41)=-0.21, p=0.83. These results suggest that all participants had the same knowledge level prior to the experiment which enabled for more reasonable comparison between the two experimental conditions. Hence, any positive learning effect found was primarily caused by using the gamified e-learning system.

Reporting on the post-test results, there was a statistically significant difference in the scores for the gamified group (M=82.73, SD=9.83) and the non-gamified group (M=63, SD=14.3); t(37)=-5.32, p=0.000, according to an independent sample t-test. These results suggest that participants who used the gamified e-learning system achieved better learning gain than those in the non-gamified version.

An independent sample t-test was also conducted concerning the follow-up test. The results indicated a statistically significant difference in the scores for the gamified group (M=44.4, SD=10.4) and the non-gamified group (M=29.8, SD=15.2); t(37)=-3.71, p=0.001. The learning gain scores according to the follow-up test in comparison to the post-test scores have been decreased by about 50% in both conditions because participants took the follow-up test after the experiment two to three weeks later affecting the sustained knowledge. Still, the results suggested positive findings related to the experimental 'gamified' condition.

By obtaining the results related to the pre-test, post-test and follow-up test, it is possible to test the hypotheses H1 and H2. Regarding the short-term learning gain, calculated as the difference between the post-test scores and the pre-test scores, an independent sample t-test was conducted. There was a statistically significant difference in the short-term learning gain scores for the gamified group (M=72.5, SD=10.5) and the non-gamified group (M=53, SD=14.7); t(37)=-5.04, p=0.000. According to the results, the hypothesis H1 can be confirmed. It can be suggested that learning using the gamified version of the e-learning system produces better (short-term) learning gain than learning using the non-gamified version.

Reporting on the medium-term learning gain based on the conducted independent sample t-test, there was also a statistically significant difference in the medium-term learning gain scores for the gamified group (M=34.1, SD=11.8) and the non-gamified group (M=19.8, SD=15.9); t(37)=-3.39, p=0.002. Based on the results, the hypothesis H2 can be confirmed. It is suggested that learning using the gamified version of the e-learning system produces better (medium-term) learning gain than learning using the non-gamified version.

VI. Conclusion

This paper contributed to the current literature by the provision of the gamification design model. The model can be used as a foundation to design different instances of gamified e-learning systems. It is consisted of major components needed to develop such systems including the learner profile, the courseware module and the gamification component. An implementation of the model was also completed resulting into a gamified e-learning system in order to validate the proposed model and to evaluate the learning effectiveness of gamification.

Moreover, a thorough experimental evaluation in terms of learning gain was conducted considering the limitations of published research. Learning gain was also reported not only according to the immediate (short-term) learning effect measured by a post-test taken by participants after completing the experiment, but also according to a delayed (medium-term) learning effect measured based on a follow-up test taken by participants two to three weeks later after completing the experiment. The findings indicated that learning using the gamified version of the e-learning system produces better (short-term and medium-term) learning gain than learning using the non-gamified e-learning version.

VII. Future Work

Pointing out to future work, it is planned to conduct a longer-term evaluation in a semester long duration adding more learning resources and focusing on another application domain in order to generalize the current findings with a larger sample size. It is also true that this paper reports on learning gain as an important variable; yet, other variables can be considered in future experiments such as motivation and satisfaction since the study presented in this paper builds up the foundation to conduct more experiments by the proposed model and by the careful experimental evaluation approach.

Another potential future direction is to compare the effect of unique combination of different game elements on different cognition and psychological factors with careful considerations to experimental design. Another possible research direction is to investigate different personalities and learning style of learners and to explore how they relate to certain game elements. Also, the culture and the learning context can play an important role in motivating or demotivating learners to use gamified e-learning systems. Therefore, more research is required to compare the effect of using gamified e-learning systems in different learning contexts and cultures. Additionally, researchers can conduct more research to explore the extent to which learners are engaged in the learning process when using gamified systems in long-term evaluation studies. Moreover, the proposed gamification model can further be improved to cater for collaborative and social learning besides the integration of game elements in e-learning systems to support active learning. An essential point to emphasize, generally in online-learning research, is that gamified e-learning systems should be designed following sound instructional theories and models.

References

[1] M. Sailer, J. U. Hense, S. K. Mayr, and H. Mandl, “How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction,” Computers in Human Behavior, vol. 69, pp. 371–380, 2017.

[2] I. V Osipov, E. Nikulchev, A. A. Volinsky, and A. Y. Prasikova, “Study of gamification effectiveness in online e-learning systems,” International Journal of advanced computer science and applications, vol. 6, no. 2, pp. 71–77, 2015.

[3] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, “From game design elements to gamefulness: defining gamification,” in Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments, 2011, pp. 9–15.

[4] G. Baptista and T. Oliveira, “Gamification and serious games: A literature meta-analysis and integrative model,” Computers in Human Behavior, vol. 92, pp. 306–315, 2019.

[5] I. Varannai, P. L. Sasvári, and A. Urbanovics, “The use of gamification in higher education: an empirical study,” International Journal of Advanced Computer Science and Applications, vol. 8, no. 10, pp. 1–6, 2017.

[6] M. Tan and K. F. Hew, “Incorporating meaningful gamification in a blended learning research methods class: Examining student learning, engagement, and affective outcomes,” Australasian Journal of Educational Technology, vol. 32, no. 5, 2016.

[7] D. Strmečki, A. Bernik, and D. Radošević, “Gamification in e-Learning: introducing gamified design elements into e-learning systems,” Journal of Computer Science, vol. 11, no. 12, pp. 1108–1117, 2015.

[8] J. Filippou, C. Cheong, and F. Cheong, “A Model to Investigate Preference for Use of Gamification in a Learning Activity,” Australasian Journal of Information Systems, vol. 22, 2018.

[9] F. L. Khaleel, N. S. Ashaari, and T. S. M. T. Wook, “An empirical study on gamification for learning programming language website,” Jurnal Teknologi, vol. 81, no. 2, 2019.

[10] T. Jagušt, I. Botički, and H.-J. So, “Examining competitive, collaborative and adaptive gamification in young learners’ math learning,” Computers & Education, vol. 125, pp. 444–457, 2018.

[11] K. Welbers, E. A. Konijn, C. Burgers, A. B. de Vaate, A. Eden, and B. C. Brugman, “Gamification as a tool for engaging student learning: A field experiment with a gamified app,” E-Learning and Digital Media, vol. 16, no. 2, pp. 92–109, 2019.

[12] E. Kyewski and N. C. Krämer, “To gamify or not to gamify? An experimental field study of the influence of badges on motivation, activity, and performance in an online learning course,” Computers & Education, vol. 118, pp. 25–37, 2018.

[13] M.-B. Ibanez, A. Di-Serio, and C. Delgado-Kloos, “Gamification for engaging computer science students in learning activities: A case study,” IEEE Transactions on learning technologies, vol. 7, no. 3, pp. 291–301, 2014.

[14] J. Koivisto and J. Hamari, “The rise of motivational information systems: A review of gamification research,” International Journal of Information Management, vol. 45, pp. 191–210, 2019.

[15] S. Subhash and E. A. Cudney, “Gamified learning in higher education: A systematic review of the literature,” Computers in Human Behavior, vol. 87, pp. 192–206, 2018.

[16] J. A. Self, “The defining characteristics of intelligent tutoring systems research: ITSs care, precisely,” International Journal of Artificial Intelligence in Education, vol. 10, pp. 350–364, 1999.

Establishing News Credibility using Sentiment Analysis on Twitter

Zareen Sharf[1]
Department of Computer Science
SZABIST
Karachi, Pakistan

Zakia Jalil[2], Wajiha Amir[3]
Department of Computer Science
International Islamic University
Islamabad, Islamabad, Pakistan

Nudrat Siddiqui[4]
Research Scholar
Karachi, Pakistan

***Abstract*—The widespread use of Internet has resulted in a massive number of websites, blogs and forums. People can easily discuss with each other about different topics and products, and can leave reviews to help out others. This automatically leads to a necessity of having a system that may automatically extract opinions from those comments or reviews to perform a strong analysis. So, it may help out businesses to know the opinions of people about their products/services so they can make further improvements. Sentiment Analysis or Opinion Mining is the system that intelligently performs classification of sentiments by extracting those opinions or sentiments from the given text (or comments or reviews). This paper presents a thorough research work carried out on tweets' sentiment analysis. An area-specific analysis is done to determine the polarity of extracted tweets for make an automatic classification that what recent news people have liked or disliked. The research is further extended to perform retweet analysis to describe the re-distribution of reactions on a specific twitter post (or tweet).**

***Keywords*—Sentiment analysis; tweets; opinion mining**

I. Introduction

The study performed an area-specific sentiment analysis on tweets to extract people's opinions or comments on the recent news. The retweet analysis is also performed in order to describe the re-distribution of reactions on a particular twitter post.

The massive increase of social media (including networking sites, blogs, forums, communities, etc.) on the web has taken a new turn in the form of public opinion. The organizations, businesses and companies now consider public opinion (or feedback) an important aspect in decision making. But filtering out the necessary information from such social sites presents an issue that needs to be resolved. The credibility of social sites is very important to be analyzed as well as the diversity of languages on such social platforms is another problem to deal with. This raises an utmost need of a smart and intelligent analysis to filter out the desired public opinion on a certain issue.

Sentiment Analysis is the solution to the aforementioned problem which is an application of natural language processing (NLP). This analysis system is capable of extracting the opinions or views of public regarding a certain topic. It consists of a system that collects opinions (or feedbacks) on a specific topic from various blogs, social networking sites and reviewing panels. There are two major challenges in sentiment analysis. First one is the opinion words and the second one is the manner in which these opinion words are expressed. The words used in an opinion can be positive (for a specific topic or issue) and negative at the same time for some other issue. Whereas the manner depicts the style or tone in which the words are being used. Either the words are used in a positive sense or the same words are used to taunt something. On the other side, language is another hurdle for many of the smart and accurate sentiment analysis systems. A number of systems have been developed to process English, but an intelligent multilingual system is an utmost requirement to cope with diversity of languages on such social sites.

Due to an emerging trend of communicating and sharing personal opinions, people participate in different events on such sites particularly the ones owned by press media. This participation comes in the form of comments and reviews on a particular news or report. These reviews and comments embody a lot of subjective information. So, some intelligent ways are required to be devised to extract meaningful information from such opinions or comments. In this regard, the NLP systems propose a term Subjective Analysis to handle the mentioned issue. This is an enclosed term that covers sentiments, emotions, opinions and evaluation. The two major approaches that are incorporated for a meticulous and accurate analysis of viewer's responses are Sentiment Analysis and Opinion Mining.

The sentiment or opinion analysis for tweets or comments is far more difficult and challenging than the systems that only collect user's feedback in the form of likes and dislikes. For the tweets (or comments), it might be possible that people have disliked something because of the unnecessary details provided on a certain topic and they are not interested to get into such details. This is the reason sentiment analysis is not merely to check each and every word, rather the system has to be guided to identify and extract what is beneficial for the analysis. It is a fact that tweets contain slang language, various internet writing styles, jokes, icons or commonly used web phrases. All these things make syntax analysis complicated and may lead to wrong classification of tweets. So, this is the major requirement to devise a powerful sentiment analysis system that not only focuses on the extraction of information but also analyses the subjectivity of the tweet. Some of the major challenges for sentiment analysis systems are:

Named Entity Extraction–This is one of the major challenges. Being existed in every tweet, the named entity extraction refers to the extraction of main idea behind the usage of any name, like the ‘Dreams of Mango People’ is a term that is used to represent the thoughts and desires of common people.

Information Extraction–Another biggest challenge is the extraction of meaningful information. ‘Was the tweet informative or fake?’ This is the question that a sentiment analysis system is devised to answer accurately.

Sentiment Determination–It is about determining the sense in which the tweet has been posted. It is either in positive or negative sense. Since sentiments are usually used in a subtle manner that makes it difficult to be analyzed from a single word or sentence. For example:

The private educational institutes are increasing fee by 20%. But these institutes are providing a high-quality of education.

The first part “The private educational institutes are increasing fee by 20%.” can be considered as a fact, while the other part is based on a personal opinion. This makes it further clear that identifying a single keyword for subjectivity is not that easy as it seems.

Parsing-In sentiment analysis, this phase is referred to as Semantic Parsing. This actually aims at highlighting the semantic constituents (subject/object, verb/adjective) are identified. It is basically a formal analysis of a sentence performed by the computer.

Anaphora Resolution: This is the phase in which nouns and pronouns are identified. Basically, an anaphora relates an expression to another one preceding it in the discourse. It presents a biggest challenge in Sentiment Analysis. Here the anaphoric expression can be explained by relating them with the context of the whole sentence. Means it can be viewed as a summarization of the context by extracting different sentences. It helps out in Sentiment Analysis by identifying which discourse identities are used repeatedly.

Twitter is one of the widely used social networks that has a massive number of users including politicians, celebrities and companies. The given research focuses on the sentiment analysis performed on Twitter data. The paper is organized as follows:

Section 2 provides a brief background about social media as well as Twitter and tweets. Section 3 gives a thorough literature review. Section 4 deals with the proposed technique that is news credibility using Sentiment Analysis. Section 5 gives implementation details while Section 6 provides performance measures, experimental results. And Section 7 gives conclusion with the recommendations for future work.

II. Background

With the advent of modern technologies and a gigantic increase of social network communication has made everything online. This is the reason With the advent of modern technologies and a gigantic increase of social network communication has made everything online. This is the reason online communication is getting cheaper. Almost 70% of the platforms are completely free and this leads to a massive participation of a layman in the form of reviews, comments and discussions regarding news updates, products, services, etc. People from all over the world can give their opinions on such platforms, means the region’s restriction is no more. This attracts researchers and analysts who want to analyze such opinions and comments. This is where NLP plays its significant role in which such comments, opinions or tweets are analyzed to extract the useful information with the help of computer programs. This entire process is referred to as ‘Opinion Mining’.

The following paper focuses on the Sentiment Analysis carried out on tweets. For this purpose, an area-specific analysis is performed; the user’s tweets are monitored on different Pakistani political as well as non-political issues. In order to extract the user’s feedback on a specific news, a mechanism (called Sentiment Analysis) is devised that basically analyzes the polarity of those tweets (checking positivity or negativity) to classify either people have liked it or not.

A. *Retweet Analysis Description*

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”.

It refers to the second major phase covered in this paper and plays a significant role in classifying the tweets and their credibility. From different studies, it is found that important news events or updates are retweeted (or shared) more. Similarly, the tweets with negative impact like news related to natural disasters are retweeted more than others. This is why the retweet analysis description is performed to describe the re-distribution of expressions or reactions on a particular tweet or post.

III. Literature Review

The major concept behind Sentiment Analysis is to determine the attitude of a user (or a writer or speaker) regarding a specific issue. Or it finds out the overall polarity of a document. Many of the work done in this realm is dependent upon machine learning approach. Due to the huge spread of user generated content through social media and forums, a massive work has been done on the sentiment analysis for a social network. Also referred as Opinion Mining, this process is considered to be highly accurate way to test credibility of any news or topic on social networks. (Denecke, Kerstin, 2008) [1] devised a method that automatically determines the polarity of sentences which are not in the same language. This is called multilingual sentiment analysis. First the text language is determined (if it was other than English) and translated into English in order to carry out sentiment analysis in an easy way. The research also determined the type of document, whether it is subjective or objective. Due to the diversity and complexity involved in human languages, the research came across many difficulties. Moreover, sentiment of text became more crucial, so the whole text was not analyzed.

Most of the work done on Sentiment Analysis had utilized the approaches of Support Vector Machine (SVM) and other classifiers with binary uni-gram weights. (Paltoglou, Georgios, and Mike Thelwall, 2010) [2] came up with the idea that classification accuracy can be enhanced with information retrieval scheme when more weighting schemes were used. The work modeled a document as an unordered collection of words called a bag of words. The approach came out to be computationally efficient. More sophisticated term weighting functions were adopted from SMART retrieval system to devise a probabilistic model.

(Caladoa, Edgar RochaaAlexandre et al., 2011) [3] had thoroughly discussed about the user contents posted on their respective Twitter profiles as well as what type of people mostly post tweets. In order to carry out this research, they first extracted a list of features. The list comprised of a number of friends, followers, tweets and re-tweets. Moreover, the user's tweeting behavior was tested incorporating feature extraction method. Finally, the public behavior on real time news events was also examined. For this purpose, an algorithm was devised that was used to extract re-tweet chains and timestamp of messages. The algorithm performed user profiling for which the limitation was that this profiling was not enough for analyzing the behaviors of users posting tweets or replying to the tweets regarding news events.

Another work performed on Sentiment Analysis by (Takaoka, Kouichi, and Akiyo Nadamoto, 2011) [4] came up with a new system titled 'Words of Wisdom'. They proposed a system that was based on multi-dimensional sentiments vector, in which two-system approach was utilized. First, a multi-dimensional vector (based on 10 categories of sentiments) was proposed, and then values for these proposed sentiments were calculated. A frequency vector was used to calculate the frequency of sentiments. The results came out to be quite accurate and the authors proposed in future work that distances could also be calculated and adoption for news data could also be incorporated.

Due to a rapid and huge spread of social networking and blogging platforms, researchers are going deeper into this area. Online Opinion Mining is another form of sentiment analysis that has been greatly worked on but it is considered a difficult form of sentiment analysis. (Haddi, Emma et al., 2013) [5] worked on the role of text pre-processing for sentiment analysis and demonstrated how the sentiment analysis can be further significantly improved by using appropriate feature selections and representations. They made use of Support Vector Machine (SVM) and performed a comparison of their accuracies with the accuracies acquired in topic categorization. The research mainly focused on the product reviews (on social blogs) and determined the importance of a product on the basis of those reviews. The supervised learning for feature sections and representations was used as the major technique and the accuracies achieved were in the range of 75% to 83%.

Going beyond the researchers have shown a great interest in Twitter-specific sentiment analysis. This kind of sentiment analysis is a bit different from the conventional one. Due to the limited character length (up to 140 characters), Twitter messages are full of slangs, short forms, abbreviations. This makes it far difficult from other forms of sentiment analysis. Most of the Twitter sentiment analysis is done using machine learning approach. The two major reasons for using Machine Learning techniques are:

1) A huge amount of Twitter data is available for training datasets.

2) The test data is also available that is user-labeled for sentiment with emoticons, so there is no need of manual annotation of data for training.

The emergence of Web 2.0 has brought improvements in the way people used to perceive Internet. Micro-blogging is one of the most popular Web 2.0 applications that have facilitated users to collaborate, share, discuss and leave their feedbacks on different topics, news and products. Twitter being one of the most popular micro-blogging platforms, has been the hot area of research for many years. (Kontopoulos, Efstratios, et al., 2013) [6] had worked on the same discipline and discussed about the ontology-based techniques for an efficient sentiment analysis of Twitter posts. Their research is divided into two phases: a) creation of domain ontology and b) sentiment analysis on a set of tweets based on the concepts and properties of ontology. This work had utilized FCA (Formal Concept Analysis) which is a mathematical data analysis theory and typically used in knowledge representation and information management. They came across a difficulty in advertising tweets in which an unpleasant ratio was involved.

(Montejo-Ráez, Arturo, et al., 2014) [7] performed Sentiment Polarity Classification on Twitter posts using a novel approach. They extracted a vector of weighted nodes from the graph of WordNet. Then, these weights were used in SentiWordNet in order to compute the final polarity. The method proposed a non-supervised approach that was domain independent. Since Twitter publishes a vast range of information including political, economic, business and more contexts, the scoring of posts is done as per the degree of positive and negative opinions expressed therein. For this purpose, SentiWordNet scores were combined with a random walk analysis of the concepts found in the text over the WordNet graph. Random Walk is an algorithm that was particularly used for mathematical formulization to perform random steps. A graph was also constructed for configuration of the results that demonstrated different parts of WordNet subgraph for the solid terms. The proposed method was intended to calculate the global polarity of Twitter posts by expanding a fewer concepts that were in tweets. The limitation of this work is that they didn't consider whole tweet text form analysis. They also came up with the fact that taking too many concepts will introduce noise in understanding the latent semantic of the text.

(Abbasi, Mohammad-Ali, and Huan Liu., 2013) [8] focused on the social media for information of the upcoming news events in the world. Since people are more interested in getting first hand news, the paper worked on the same issue and proposed a method to measure user credibility in social media. For this purpose, they proposed CredRank algorithm that was solely devised to measure user credibility in social media. It analyzes the users' online behavior to measure the said credibility. The proposed methodology worked in the given steps:

- Detect and cluster coordinated (i.e. dependent) users together.
- Weight each cluster on the basis of the cluster's size.

Due to the anonymous and unmonitored mature of the Internet, the user generated content on Twitter might be incredible. This incredibility leads researchers to work on different ways to perform credibility analysis. (Gupta, Aditi, and Ponnurangam Kumaraguru., 2012) [9] worked on this realm and performed credibility analysis on the information contain in a tweet corresponding to fourteen high impact news events of 2011 occurring globally. They conducted research with the help of Regression Analysis through which they identified the important content and source-based features; helpful for predicating the credibility of information contained in a tweet. The ranking of tweets (according to their credibility scores) was performed using the supervised machine learning and feedback approach. The performance of the ranking algorithm was significantly improved when re-ranking strategy was applied. With all the data analyzed, it came out to be known that pn average 30% of total tweets posted about an event comprised of situational information. While the remaining 14% contained spam tweets. Whereas only 17% of the total tweets with situational awareness information were credible. Pseudo Relevance Feedback (PRF) was used for re-ranking purpose. This technique is also known as Blind Relevance Feedback and one of the most prominent re-ranking techniques used in information retrieval tasks to improve the performance of re-ranking. PRF works by extracting K ranked documents and then re-rank them on the basis of denoted score. The algorithm extracted the most frequent unigrams from the top K tweets and re-rank them by utilizing the text similarity between those most frequent unigrams and K tweets. PRF was basically applied to the best set of results that was acquired by previous analysis (that is the ranking results using both message and source. Using the metric BM25, the text similarity between a tweet T and query set Q was determined for each event occurred. Around 50% of tweets on an event are composed of the tweets which were related to the event but didn't provide any useful information. So, it was concluded that a ranking algorithm is based on both the user properties and content. And it turns out to be very effective in determining the credibility of information in these tweets.

(Amiri, Fatemeh et al., 2015) [10] had performed sentiment analysis for Persian text through lexicon-based approach. Since a very little amount of work has been done so far on Persian language, therefore in order to gain insights from different online sites and social media, sentiment analysis was performed. But the researchers came up with a novel approach of incorporating a manually created lexicon that was enriched with sentiment scores, coupled with hand-coded grammar rules. The work also addressed some of the Persian language issues including difference between formal and informal writing styles, context sensitivity, complexity due to frequent morphological operations, lexicon intricacy etc. To perform the proposed idea, they first manually collected Persian adjectives, words and expressions from two online resources. After the collection, sentiment annotation was performed in which the collected words were annotated with corresponding sentiment scores. Then, a lexicon based sentiment analysis pipeline was created. This was comprised of steps including: Tokenizer – that splitted text into very simple tokens. Sentence Splitter – that fragmented text into sentences. POS Tagger – It produced a part of speech tag as an annotation for each word or symbol. Gazetteer – This was the basis of the proposed methodology, as each gazetteer entry when appeared in the text, it got marked and was assigned a sentiment score accordingly. After passing the data through this pipeline, the JAPE rules (hand-coded grammar rules) were devised. These rules were formed in two phases – phase I formed word-level rules whereas phase II worked on sentence-level rules. Finally, the Groovy scripting processing resource was utilized through which the number of positive and negative annotations (in a given text piece) were counted and an overall polarity was determined as well. The proposed method yielded around 60 – 70% accuracy rates for the initial version of lexicon-based sentiment analysis API. Although, the method did not come up that much efficient as most of the ML based approaches. But it showed value and could be combined with some ML based approach to produce a hybrid system.

(Sharma, Nitesh, et al, 2018) [11] designed a web-based application for performing sentiment analysis of live tweets. Due to a massive use of social media, people use this platform for expressing their opinions on almost every topic. That is why the researchers came up with an idea of building a web-based application that not only performs sentiment analysis on live tweets but also visualizes the measured sentiments associated with keyword (hashtag, words or phrases) of Twitter messages. So, this enables users to measure the sentiment of these messages in terms of geography. The application is designed on the framework 'Flask' using Python programming language. This framework is built in a way that user enters a keyword and application fetches the live tweets (related to the entered keyword), extracts text from each tweet and calculates the user location and sentiment for each tweet and finally plots the results on a map. Flask framework contains an initial configuration file that is used by Views module. This module is responsible for rendering web pages and communicating with API. It intercepts the incoming requests and transfers the control to the back-end layer for processing. As the result is ready, the views module generates web pages. In order to extract tweets, Twitter streaming API is utilized that fetches live tweets corresponding to the entered keyword. Moreover, the attributes (like location, tweet text, followers count, friends count, tweet time) are extracted using the meta data. For extracting the user location, a parser is created that maps user countries to the countries extracted from data. In this regard, there is one additional feature particularly for the US residents that the app generates their state wise plotting on map as well. For sentiment calculation, Python's text-blob library is incorporated. The sentiment score assigned to words range from -1 to +1 and a polarity score of 0 is termed as neutral sentiment. An in-depth analysis of data is also provided in a way that system also calculates the mean and weighted polarity values. Furthermore, the system also calculates the number of tweets from each country and from each within state of the USA. Map plotting is done with plotly and Tableau. This application is unique in a sense that it can perform sentiment analysis on live tweets as well as on previously gathered tweets stored in a database. When the system was evaluated using a search string #Watercrisis, the results were found that a total of 1,164 tweets were extracted related to this hashtag (or keyword) from around the world. And their respective sentiment scores were successfully calculated and visualized on map as well.

(Lauren, Paula, et al., 2018) [12] have conducted a massive research by generating word embeddings from an extreme learning machine for the sentiment analysis and sequence labeling tasks. Word embeddings are basically the low-dimensional distributed word representations for a set of language modeling and feature learning techniques. The words or phrases from the designed vocabulary are mapped on to vectors of real numbers (in low dimensional space). This research focuses on ELM based word embeddings for the sentiment analysis and sequence labeling tasks. There are already different models for generating word embeddings; Word2Vec and Global Vectors (Glo Ve) are the popular of them. In this research, they have also done a comparative study in which ELM based word embeddings are compared with the aforementioned models. Both models use word-pair co-occurrences (also called as Word-Context Matrix) for learning the low-dimensional representation of words. Word2Vec computes this matrix one row at a time while Glo Ve computes the matrix at once and then applies matrix factorization. The first model is known as Predict-based as it performs line-by-line computation, whereas the other one is called as Count-based method because the word pair counts are performed all at once in the first. Their results are comparable, but Word2Vec takes more time as it needs to train the neural network also, while Glo Ve consumes more memory. On the other hand, Extreme Learning Machine (ELM) is actually a type of feed forward neural network and its efficiency is attributed to the random non-updated hidden layer weights as well as the efficient learning of output layer weights. This study has utilized an Autoencoder architecture based on ELM for its feature learning functionality. Autoencoder itself is a type of neural network that performs feature learning by compressing the input feature vector (in the hidden layer) and then decompressing it back in the output layer. This compression feature is very useful because in generating word embeddings using an ELM, the compressed representation is desired because word embeddings correspond to the low-dimensional representation of Word-Context matrix. Furthermore, a recurrent neural network (RNN) is also utilized in this study for the sequence labeling task. An RNN is a feed-forward neural network that contains recurrent connections back to the previous layer. This feature is useful for processing sequential data. They have incorporated Elman RNN for this study. Then for sentiment analysis task (to assess the models of word embeddings), Logistic Regression is applied as a classification algorithm in which the outcome Y=1 means positive result, whereas Y=0 means the outcome is negative. For conducting research, three separate datasets are utilized in which two datasets belong to Sentiment Analysis task and one for Sequence Labeling task. Word embeddings were generated using three models: ELM, Word2Vec and Glo Ve. Furthermore, Word2Vec includes four models – two from Skip-Gram and two from Continuous Bag of Words (CBOW). The parameters used across all of the models were: minimum word count (means the number of times a word is present in the corpus), word vector dimension size and window context size. Moreover, Word2Vec and Glo Ve required different hyperparamters including Learning rate, Weighting, Iteration and more. Whereas, ELM required only one hyperparamter that is Regularization. For generating ELM word embeddings, the foremost step is to build a vocabulary and its major parameter is minimum word count that finds out how many times a word should appear in the corpus to be included in the vocabulary. Then, frequency counts are rendered for each of the word pairs from the training corpus. The window context size is utilized that determines how far to the left and right each word has to be used for the co-occurrence counts. A weighting scheme is incorporated to get a numerical representation of context, so the closer words are given higher weight and the distant ones are assigned with lower weight values. The word context matrix is a square matrix with the dimensions matching to the size of vocabulary. The square root transformation of the word-context matrix with l_2- normalization is done before applying ELM. The word vector dimension size corresponds to the number of neurons in the hidden layer of ELM. For this purpose, the MATLAB ELM Autoencoder source files were used. The study has performed both intrinsic and extrinsic evaluation The intrinsic evaluation means word embeddings are evaluated on the basis of semantic relatedness whereas in extrinsic evaluation the word embeddings are assessed in the downstream NL Task like text categorization. The results on sentiment analysis task are evaluated using Precision, F1 and Recall measures. For the sequence labeling task, Precision and Recall are measured and F1- score is applied utilizing the Precision and Reall equations. In sentiment analysis task, the results of F1-score demonstrate that ELM word embeddings are competitive with Skip-Gram and Glo Ve word embeddings. And for the sequence labeling, Precision, Recall and F1-score are averaged across 20 executions using the test set incorporating the RNN for six models. The experimental results show that CBOWHS model did well overall in comparison of other three Word2Vec models. While the ELM word embeddings show a slightly better average in terms of F1-score as compared to other five models.

(Shirsat, Vishal S et al., 2019) [13] discussed about sentence level sentiment identification by performing research on news articles. The data (of news articles) was extracted from BBC news and sentence-level negation identification was basically applied. Sentiment Analysis is basically categorized as: Document level sentiment analysis in which the polarity of entire document is determined, Sentence-level sentiment analysis in which each sentence is analyzed and polarity is determined. And Aspect-level sentiment analysis that applies analysis on objects and their respective features. This study uses sentence-level sentiment analysis on news articles using Machine Learning Algorithms Support Vector Machine and Naïve Bayes. The proposed methodology consists of five major steps in which the foremost step was to preprocess the dataset in which the irrelevant text (in news articles) like HTML tags, advertisements and scripts are removed. So, the data is prepared for text classification in further steps. The practical implementation involves applying Machine Learning algorithms to perform the classification task. Naïve Bayes algorithm is one of them that determines the probability of an occurrence given the probability of another occurrence that has already occurred. Another algorithm that is used in study is Support Vector Machine which is non-probabilistic algorithm and works well for both sequential and non-sequential data. After preprocessing of data, the next step is to apply Stemming in which the entire document is transformed into lower case to gain uniformity. Stemming is basically the truncation of a word to its root form. After this, the Term Document Matrix is determined, this matrix defines the frequency of terms that appear in the preprocessed dataset. The rows of matrix are collection and columns correspond to the related terms. Then, sentiment score generation is performed with the aid of positive and negative dictionary. For this purpose, each word in the preprocessed dataset is compared to the word in dictionary to determine whether it is positive or negative. Finally, the Naïve Bayes and Support Vector Machine algorithms are applied to perform classification and estimate the accuracy. The experiment was conducted on the five categories of news articles including Entertainment, Business, Politics, Sports and Technology. They utilized Bing Liu dictionary for determining positive and negative words. This dictionary contains 2006 positive and 4783 negative words. The experimental results demonstrated that Naïve Bayes achieved an accuracy of 96.46% for Entertainment category and the lowest accuracy was 92.63% that was for Business category. While Support Vector Machine achieved highest accuracy 94.16% for Politics category and the lowest as 69.01% in Sports category.

(Iqbal, Farkhund, et al., 2019) [14] came up with a novel approach in which they designed and developed a hybrid framework for performing sentiment analysis. This framework combined ML based algorithm along with lexical database to automatically analyze the online content (including reviews, social media). Then a Genetic algorithm based feature reduction solution is provided through which further accuracy and scalability is achieved. Moreover, they have proposed a cross-disciplinary area of geopolitics to which they have applied the proposed framework as a case study application. It is a complete unique approach in which they tested and experimented the accuracy of the proposed approach by applying it to the topics like terrorism, global conflicts, etc.

IV. Methodology

This section gives a detailed discussion about the proposed methods for performing the desired task. The main focus of the work is to perform sentiment analysis on the public opinions and reviews on Twitter on the daily news updates. Therefore, the most important part is to determine how people react about a certain news or topic. In this section, the proposed methodology gives an overview about how data is extracted for analysis as well as the method applied to perform extraction. The next phase deals with data pre-processing and how the application shows accuracy in results. In this regard, a tool is proposed that is used to extract data from Twitter and analyzes it according to the applied algorithm.

A. Data Collection

Data extraction and collection are the major parts of this research. Extraction is actually a time-consuming process. In order to perform this task, first an account was created on Twitter and an application was also created for authentication to extract the required tweets. Twitter API was incorporated to achieve this work. API is an application programming interface, in which a set of protocols and tools for development of different software applications is defined. This API is open source from where the code (related to Twitter data) is extracted and can further be used in different respects of research work.

B. Proposed Architecture

The proposed architecture of the system is divided into two steps: Sentiment Analysis and Retweet Analysis. Let's analyze both steps here:

Sentiment Analysis Architecture–The proposed architecture of sentiment analysis depicts the entire scenario of sentiment analysis that how it works on the provided data. The major performance measures applied for the manual annotation (during sentiment analysis) are:

Confusion Matrix–Also referred to as contingency table, matching matrix or errors matrix; this measure is applied on both supervised and unsupervised learning systems. This measure is used in order to have a visual overview for the overall performance of the model used in work. Actually, the matrix compares the actual class values with the predicted ones. This way, it analyzes whether the presented model is confused between the two classes or not. The comparison of predicted values with the ground true values gives a clear picture about the proposed model. The matrix is drawn in the form of table shown in Fig. 1.

It is clear from the matrix that there are two classes positive and negative. But they are further divided in boxes and termed as TP, FP, TN and FN

TP corresponds to the true positive terms. These are the values that are positive in actual and the proposed model has also predicted them as positive.

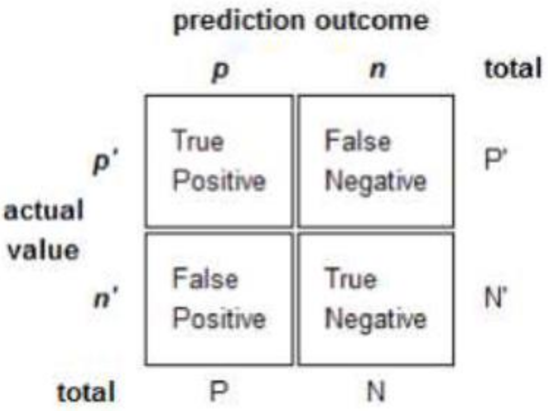
FN means false negative. Means these terms are actually negative but the proposed model did not predict them as negative.

TN represents the values that are negative in actual and the proposed model has also predicted them to be negative.

FP represents the values that are positive in actual but the proposed model didn't predict them to be positive.

Precision–Precision is basically defined as the ratio of all those instances which are correctly predicted by the classifier as positive. Although, it is a performance measure for binary classification (and we have used multi classification), but it can be used for multi class classification. The formula is defined below:

$$\frac{TP}{TP + FP}$$



		prediction outcome		
		p	**n**	**total**
actual value	**p'**	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
	total	P	N	

Fig. 1. Confusion Matrix.

Sensitivity-It corresponds to the ability of test for correctly identifying the condition. Also known as Recall, it checks for the strength of classifiers' probability to select instances of a specific class from the whole dataset. It, in actual, corresponds to the true positive rate having the shown formula:

$$Sensitivity = \frac{TP}{TP + FN}$$

Specificity–It is the measure of the proportion of negative instances (or values) correctly identified from the data set. Means it is opposite of sensitivity. The formula is:

$$Specificity = \frac{TN}{TN + FP}$$

Kappa Coefficient–This performance measure is very useful in comparing the actual accuracy of a system with random accuracy. As defined by R. Landis and G. Kochh, total accuracy of a system is an observational probability of agreement, whereas random accuracy is a hypothetical expected probability of agreement under an appropriate set of baseline constraints. The Kappa accuracy formula is given here:

$$Kappa = \frac{Total\ accuracy - Random\ accuracy}{1 - Total\ accuracy}$$

Re-tweet Analysis Architecture–A description is given below:

Re-tweet Analysis–The retweet analysis is performed in parallel with sentiment analysis. First, the data is downloaded from Twitter, then source of the original posts is analyzed. These steps are followed by the steps in which re -tweets are checked and counted as well. While analyzing the re-tweets, it is important to be considered that source should not be the same. Means the person who has basically tweeted must not re-tweet it as it comes under an effort to increase the ranking of his post or tweet. For re-tweet count, if it is greater than 35 the news is credible.

Support Vector Machine-Abbreviated as SVM, this is one of the best classifiers which is used for binary classification. Its major purpose is to use the entire data on a high dimension space and try to acquire the maximum margin hyperplane between both data sets.

The Hyperplane-There might be any number of hyperplanes (for SVM) for the specified data points. Moreover, they are further classified into thick and thin hyperplanes. The main objective of SVM classification is to find out a hyperplane that is linearly separable and has the largest margin.

Following constraints are required in order to calculate the hyperplane:

- x i- the vector that contains the attribute values of all instances of which is the vectors that contains the attribute values of all the instances of i.
- *W* - the vector containing weights of all attributes.
- *b* - is a real number created for representing the y-intercept.

We set the decision boundary on the points that the following equation comes out to be true:

$x \,.\, w + b = 0$

Suppose, we have two points that lie on the decision boundary. So,

x a . $w + b = x$ b . $w + b = 0$

Thus, we can also say that:

w (x b- x a) = 0

where, we know that both these points are parallel to the decision boundary.

The main formula for the hyperplane as generated by *w* and *b* is given below:

$$f(xi) = xi.w + b$$

So, for some point, if x i . $w + b > 0$; it will lie above the hyperplane. And if x i . $w + b < 0$, it will lie below the hyperplane. We can represent the classes as 1,0 and they can be written in the form

Y={ 0, if x i . $w + b > 0$}

Y={1, if x i . $w + b < 0$}

These points can be named as Support Vectors.

Distance-The following rule is used to give the distance between margins and the decision boundary:

$$D = \frac{2}{||w||}$$

We have to estimate the parameters and *b*. w throughout the course of SVM learning.

As mentioned above the key criterion behind SVM classification is to have a correct classification of all points. Fig. 2 presents a visual representation of this.

$x_i.\ w + b \geq 1$ if y=1

$x_i.\ w + b \geq -1$ if y=0

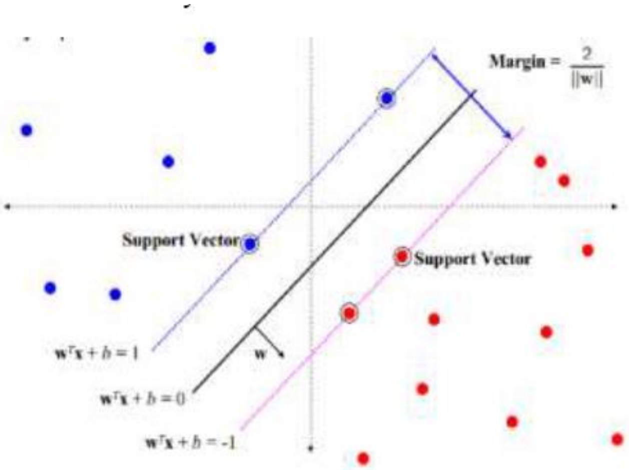

Margin = $\frac{2}{||\mathbf{w}||}$
Support Vector
Support Vector
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$
w

Fig. 2. The SVM Hyperplane.

Furthermore, it is also mentioned above that for SVM's hyperplane, the margin must be the largest. This can be achieved by minimization of the following formula:

$$f(w) = \frac{1}{2}||w||^2$$

To carry out minimization, the following constraint must be fulfilled:

$Y_i(w.x_i + b \geq 1)$ for $1 \leq i \leq N$

Lagrange Multipliers is incorporated to optimize or minimize the constraint. The formula for the multiplier is given here:

$L(x, \lambda) = f(x) + \Sigma_{i=1 \text{ to } m} \lambda_i g_i(x)$

The following two steps are required to solve Lagrange Multiplier:

$\frac{\partial L}{\partial xi} = 0$ for $1 \leq i \leq n$

$\frac{\partial L}{\partial xi} = 0$ for $1 \leq i \leq m$

V. Model Implementation

This section describes the implementation in detail with the description of tool used and its working. ASP.Net framework is used in this study to perform the task. This is a web application framework developed and marketed by Microsoft. It incorporates programming languages including C#, VB.Net that enable a programmer to build dynamic and interactive web applications. There are many interactive controls like text boxes, buttons, labels etc. that easily configure and manipulate the code. ASP.Net has extended web forms (as event-driven model) for web applications. As the browser submits a web form to the web server, a markup page is generated in response. All client side activities are sent to the server for full state of processing. It is the sole responsibility of server to process the output in accordance with client's action and triggers the reactions.

Information is stored in two states, Page and Session states. Page State is basically the client's state in which the contents of various input fields are present in the web forms. While the Session State represents the server's state that keeps track of the information globally collected over a session. ASP.NET carries pages to and from the server at runtime and also codes the state of server side components in hidden fields. This is done in order to make server aware of overall running applications and operate it in a two-tiered connected way.

ASP.NET uses and accesses the following data sources:

- Databases (e.g., Access, SQL Server, Oracle, MySQL)
- XML documents
- Business Objects
- Flat files

ASP.NET also hides all the details related to the processes of data access, so a higher level of classes and objects is provided that facilitates the data access. Moreover, ADO,NET is another technology incorporated by ASP.NET that works as a bridge between backend data source and ASP.NET control objects.

Twitter: All the data used in this study is real-time and extracted from Twitter. The whole analysis is performed on the posts, reviews and reactions on the news posts uploaded on Twitter. We have performed the research by restricting the location to Pakistan, so only the news posts belonging to Pakistan were considered. APIS are the downloaded data, they are not used to check location. The data on which experiment is conducted is less as many of the duplicate posts were removed. But in order to check the effect on accuracy by increasing the training data, the most recent Pakistani news are considered.

Test Data: In order to analyze the sentiments of Twitter posts (related to news events), only those tweets are extracted that contain words about news headlines (against a certain threshold value) were collected. These tweets were gathered while considering the location (that is restricted to Pakistan related news updates). All the data was from July 2014 to August 2014.

The second important task is to check the number of replies (known as Re-tweets) on these gathered news based posts. But the source of re-tweet must not be the same as of original tweet (as mentioned earlier). If the source comes out to be same, it will get difficult for finding the credibility of that post because sometimes the same person re-tweets his/her post in order to have an increased re-tweet count.

Total Initial Data: The test data (as described above) is downloaded from Twitter. In order to download the most recent posts, we first checked the trending topics on Twitter's home page. All this data collection was done with the help of APIs. The table (given below) shows complete details including total number of tweets, number of trending topics and their count with positive, negative and neutral tweets.

Trending Topic: It corresponds to the topics or trends that are currently in search. These trending topics are uploaded daily on Twitter to give the users an overview about the trending news or updates.

Keyword Matching: In order to extract tweets, a specific keyword is given to the application. Then, application starts searching tweets (on the website) related to that entered keyword. As the relevant tweets are found, the application connects with Twitter API to download those tweets. Then, API performs authentication of user with oath keys (that are special keys provided to Twitter users who want to download data from the website). Further analysis is performed after the downloading of data.

Sentiment Analysis: Basically, there are three types of Sentiment Analysis: i) Word based Sentiment Analysis, ii) Sentence based Sentiment Analysis, and iii) Document based Sentiment Analysis. Our study focuses on sentiment analysis of sentence as well as words. We have researched about the reactions, emotions and feelings of people on different news (related to Pakistan). After matching and extracting words related to the feelings, sentiment analysis categorizes the extracted words into three categories: Positive, Negative and Neutral.

Positive Sentiment: The reactions (or comments) on the news (or Twitter post) showing happiness and positivity are categorized as positive sentiments. We have, at the backend, incorporated separate dictionaries with both positive and negative words. Moreover, two languages are considered, English and Urdu. Since, we have performed research on Pakistani news based Twitter posts, so most of the tweets are in Urdu language. The most common positive words are happy, nice, good, well-done, thanks, welcome, I like it and more.

Negative Sentiment: The same procedure (as used in positive sentiments) is applied for negative sentiments. If an extracted post has more negative words than positive and neutral, the overall post is considered negative. The most common negative words are worried, sad, sorrow, angry, crying, bad etc.

Neutral Sentiment: The sentiments that don't show any particular feelings like happy, sad, angry, etc. are the neutral sentiments. It gets difficult to identify the neutral tweets because they contain both positive and negative words. If post contains maximum number of positive words, the polarity is measured positive and if maximum number of negative words, the polarity is negative. Otherwise the post is considered to have neutral polarity.

Polarity Check: To measure polarity of the tweet text, the words are compared with both dictionaries (English and Urdu) in parallel. As the word is found in a dictionary, then polarity is classified, in the second step, as positive, negative or neutral. Then number of words are counted and on the basis of this count, polarity of overall post (i.e. positive, negative or neutral) is assigned.

Training Data: In order to achieve accuracy in results, it is required to train application on specific performance measures. In this regard, human annotation is performed before applying performance measures and it is manually done by checking twitter posts and the sentiments against each trending topic and then applying performance measures.

Manual Annotation on Training Data: Twitter posts are manually labeled and then results achieved are compared with the results provided by the application. The performance measures are used to measure accuracy and all comparisons are made on the basis of these measures. A few examples of positive, negative and neutral tweets which are annotated are provided in Tables I and II.

Classification Results: SVM classifier was utilized to get the desired results. The obtained accuracy is measured to be 95%.

TABLE. I. EXAMPLES OF POSITIVE POSTS

Twitter Posts	Category
#PTI will make #KARACHI the city of lights again in Naya #Pakistan InshAllah #PTI4Karachi #bekhaufJunooniKarachiwalay	Positive
I congratulate @ImranKhanPTI ; other leaders of PTI on successful public meeting in Karachi this evening.	Positive
Happy Defence Day...Long Live Pakistan @MHaris_ @__Shaikho @jdk_update @OyeeParkashaayy @KyunBatow @jiyaleem	Positive
Proud Soil, Proud Nation with Determined Force…… National Air Defence Day…….Pakistan Zindabad	positive
Allah protect our families in Multan and the surrounding area from the flood threat@MultanInFlood	positive
PML-N &amp; CM Punjab are famous for progressive work in Lahore. One big rain disclosed all efficiency in Lahore.	positive

TABLE. II. EXAMPLES OF NEGATIVE POSTS

Twitter Posts	Category
Ahmad_Noorani: Geo ISD Real fascist face of PTI. MQM much better than PTI. Karachi ppl wil reject PTI terrorists in next elections..	negative
Failed show by PTI in Karachi yesterday, hardly 20K people came. #BurnolForPTI	negative
After So Much Rain Still 14 Hour LoadShedding in Lahore camp; in result of that People have attacked Lesco office Today burned its furniture !!	negative
Heavy rain starts again in Lahore with a flood warning issued. - Bubblews http://t.co/VL8at9ytTT via @GoBubblews	negative
RT @ABKool: PMLN brought gullu butts in police to attack azadi march and CM KP called people from KP to come protect Khan	negative
Atleast five armed militants attacked Astana Fazal in Sargodha on September 6, Pakistan Defence Day. - http://t.co/11Beh1ejLQ	negative

VI. Experimental Results and Discussion

This section discusses the experimental results with complete details including tables and graphs. The step by step process is given below:

Keywords selection: This is the step in which user goes to Twitter's homepage and searches for the trending topics. As per the location (of the user) specified, Twitter generates the trending topics. The information is presented in Table III.

Extraction of tweets: The keywords are added through analysis box. Press the create button to add analysis word in the analysis drop down menu. As the Run Analysis button is clicked, the program extracts tweets as per the entered keyword. The tweets are generated in the form of bundles like 300 tweets in one bundle.

Sentiment analysis: Sentiment analysis on extracted tweets is then performed that shows positive, negative or neutral tweets as per the given polarities. For this purpose, this application analyzes the polarity of words in each tweet. And the number of positive and negative words are counted. With higher number of positive words, the polarity is defined to be positive and similarly with negative words. The figure and table shown below specify this step clearly.

Retweet analysis: Retweet Analysis is performed in parallel to show the credibility of results. This analysis works on the concept that if a specific keyword has 35 or more retweets in a particular time interval, that topic is considered to be an interesting one that's why more people are interesting in that update. So, it is labeled as a credible news update. This is presented in Table IV and can be visualized in Fig. 3.

TABLE. III. Sentiment Analysis of Trending Topics

Keywords	Tweets Count	Negative	Positive	Neutral	Credible/Non-credible
Flood-in-Multan	891	519	78	294	Credible
One Nation Day	298	40	113	145	Credible
Punjab	4162	3011	407	744	Credible
Red Zone	1190	323	270	597	Credible
Defence Day	595	200	229	166	Credible
Flood in Pakistan	892	831	25	36	Credible
Operation Zarb E Azb	594	267	82	245	Credible
Civil Disobedience	1784	756	343	685	Credible
Umar Akmal	981	181	219	491	Credible
Rain in Lahore	298	257	9	32	Credible
iPhone 6	826	199	178	449	Credible
Tsunami in Karachi	298	282	7	9	Credible
Azadi March	893	424	175	294	Credible
PMLN	1188	523	199	466	Credible
Free-and-fair-Election	398	77	268	53	Credible
PTI in Karachi	1785	384	661	740	Credible

TABLE. IV. Retweet Analysis Credibility Table

SR.#	Id analysis	text	retweet_start_Date	retweet_end_Date	retweet_result	Retweet score
1	1184785917	Flood-in-Multan	9/14/2014 0:00	9/14/2014	credible	81
2	1230244833	One Nation Day	9/13/2014 0:00	9/13/2014	credible	38
3	1283045768	Punjab	9/13/2014 0:00	9/13/2014	credible	67
4	1387115085	Red Zone	9/13/2014 0:00	9/13/2014	credible	35
5	1391593695	Defence Day	9/12/2014 0:00	9/12/2014	credible	52
6	2030857401	Flood in Pakistan	9/12/2014 0:00	9/12/2014	credible	60
7	2065829408	Operation Zarb E Azb	9/13/2014 0:00	9/13/2014	credible	37
8	2138605727	Civil Disobedience	9/13/2014 0:00	9/13/2014	credible	65
9	254201070	Umar Akmal	9/14/2014 0:00	9/14/2014	credible	53
10	289536224	Pak Army in Islamabad	9/12/2014 0:00	9/12/2014	NULL	NULL
11	366838366	Rain in Lahore	9/13/2014 0:00	9/13/2014	credible	65
12	474851049	iPhone 6	9/13/2014 0:00	9/13/2014	credible	59
13	553443382	Tsunami in karachi	9/12/2014 0:00	9/12/2014	credible	44
14	677447376	Azadi March	9/13/2014 0:00	9/13/2014	credible	62
15	973193056	PMLN	9/13/2014 0:00	9/13/2014	credible	58
16	1738732573	PTI-in-karachi	9/23/2014 0:00	9/23/2014	credible	71

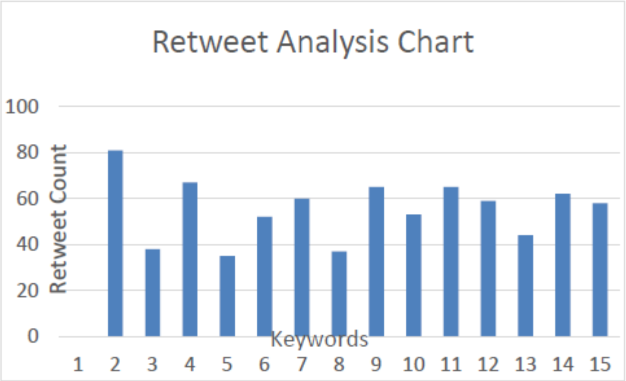


Fig. 3. Retweet Chart.

Output Result-In order to check performance, the performance measures used were precision and recall. A detailed account of the experiment conducted and results achieved is presented in the sections to follow.

Experimental Setup-Here the experimental results are shown in the form of tables according to the applied performance measures. Table V shows the dataset that we have used. This dataset comprises of the extracted tweets from Twitter and analysis was made on the recent and trending topics.

Now some of the extracted topics and their relevant results are discussed.

Topic= "Red Zone'-This topic was considered negative news before running analysis. The results acquired after running application over 1400 extracted tweets also prove that the news update is negative. The accuracy of the result was checked using performance measures and their respective values are given Table VI. Statistics related to confusion matrix are presented in Fig. 4.

Overall Accuracy: 81.71%
Kappa coefficient: 0.628
Sensitivity: 80.60%
Specificity: 82.24%
Values of chart area:
True Positive = 108, True Negative = 125, False Positive = 27, False Negative = 25

Topic= "Defense Day"-Obviously, this news update is positive and our application also generated the same result by achieving positive polarity for this news update. Among 975 extracted tweets, there were 375 positive ones, 252 negative and rest of the tweets were neutral. After the results were acquired, the performance measures were applied to check the accuracy and are presented in Table VII. Statistics related to confusion matrix are presented in Fig. 5.

Overall Accuracy: 81.65%
Kappa coefficient: 0.623
Sensitivity: 86.67%
Specificity: 76.88%
Values of chart area:
True Positive = 195, True Negative = 130, False Positive = 37, False Negative = 36

TABLE. V. EXPERIMENTAL SETUP

Dataset details	
Attributes no.	Five classes
Characteristics of attribute	Integers
Missing values	Nil
Predicted variables	Positives and negative tweets

TABLE. VI. PRECISION AND RECALL VALUES (RED ZONE)

	Class 1	Class 2	Classification overall	Procedure Accuracy(precision)
Class 1	108	27	135	80%
Class 2	26	125	151	82.78%
Truth overall	134	152	286	
User accuracy (recall)	80.59 %	82.23 %		

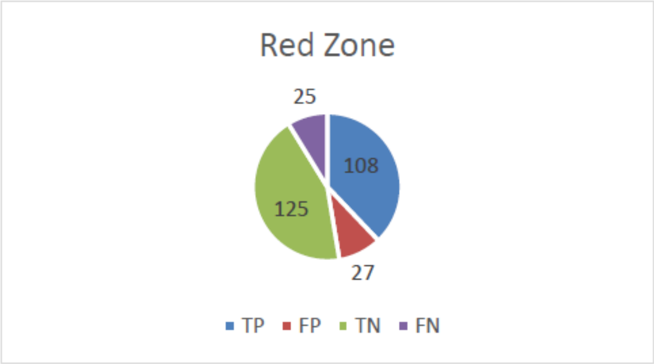


Fig. 4. Confusion Matrix Values (Red Zone).

TABLE. VII. PRECISION AND RECALL VALUES (DEFENSE DAY)

	Class 1	Class 2	Classification overall	Procedure Accuracy(precision)
Class 1	195	37	232	80.05%
Class 2	36	130	166	78.31%
Truth overall	231	167	398	
User accuracy (recall)	84.41%	77.84%		

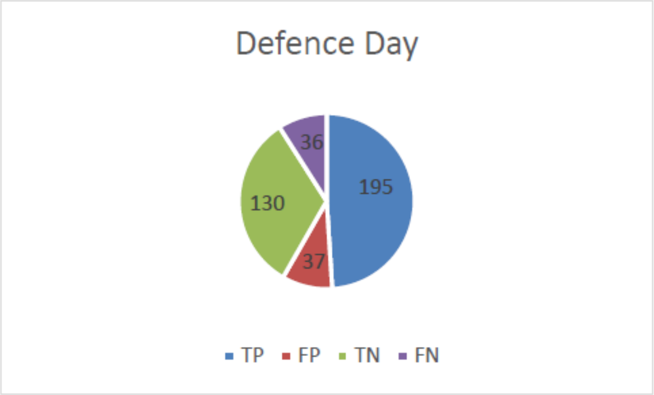


Fig. 5. Confusion Matrix Values(Defense Day).

Topic= “Rain in Lahore”-The polarity found for this news update was negative. In the month of September, heavy rain occurred that was quite destructive for Lahore and the cities around Punjab. The application gave 298 tweets in which 257 were negative tweets and merely 9 were positive ones. The accuracy results (using performance measures) are provided in Table VIII. Statistics related to confusion matrix are presented in Fig. 6.

Overall Accuracy: 81.57%
Kappa coefficient: 0.513
Sensitivity: 54.55%
Specificity: 92.59%
True Positive = 30, True Negative = 125, False Positive = 10, False Negative = 25

The overall accuracy achieved for each of the extracted topics depicts the authenticity of the proposed system and the potential it has to get enhanced for more advanced results in future.

TABLE. VIII. PRECISION AND RECALL VALUES (RAIN IN LAHORE)

	Class 1	Class 2	Classification overall	Procedure Accuracy(precision)
Class 1	30	10	40	75%
Class 2	25	125	150	83.33%
Truth overall	55	135	190	
User accuracy (recall)	54.54%	92.59%		

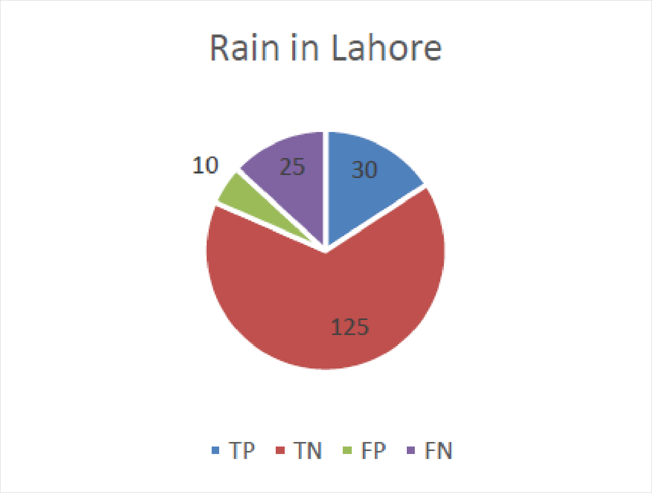


Fig. 6. Confusion Matrix Values (Rain in Lahore).

VII. CONCLUSION

The research is aimed at solving a practical problem of Sentiment Analysis of Twitter posts related to news updates. We have presented the background, related literature review as well as possible approaches, techniques, features and assumptions for Sentiment Analysis about news events. The data collection process is also thoroughly specified that we have collected the required data with the help of Twitter APIs. After collecting the tweets related to news updates, they were analyzed to determine their type as per their respective polarities. The chances of noisy data were reduced by downloading the tweets through keyword matching process.

One point is to be cleared that the tweets extracted are all related to the trending news events means the topics that are in great discussion and appear on Twitter’s homepage on a daily basis. The reviews or reactions on such tweets are used to determine the polarity as well as credibility of the news updates. It was found during the study that sentiment analysis of tweets can be performed independently without concern of their context. But feature extraction is the crucial element that matters a lot. In this regard, uni-gram and bi-gram proved to be the better features than other ones for performing reliable sentiment analysis.

In the next phase, the re-tweet analysis was another challenging task to be performed. By setting a threshold value >= 35, the approach worked quite better. This analysis was performed in parallel and credibility of the news events was proved on the basis of re-tweet analysis. If a particular keyword has 35 or more retweets, the topic was considered to be interesting. And more people were found to be taking interest in that topic. This further proved that news is credible.

There were some challenges faced during this study. The collection of tweets related to news updates was quite a challenging and difficult task. After overcoming this challenge, we came across with another challenge that was in the form of manual annotation. In manual annotation, it is difficult to decide whether a given tweet is positive, negative or neutral. Since the tweets are related to trending news events and there is always a possibility that tweet is dependent on the context that makes it further difficult. In this regard, a subjective approach is required because annotation is highly subjective. The limit of 140 characters in a tweet also brings some difficulties, as people find difficulty in expressing their views or opinions in such a restricted limit. So, they use abbreviations, short forms and slangs that make sentiment analysis further challenging.

More future work and studies can be done in this realm. By proposing more complex algorithms and utilizing appropriate measures, a more scalable approach can be proposed. The advanced techniques can be used to enhance the work for resolving issues like detecting fraudulent or criminal activities on social media. Moreover, we focused on English and Urdu languages while ignoring the slang terms. The future work may incorporate the technique that may be capable of analyzing the slang terms as well.

REFERENCES

[1] Denecke, Kerstin, “Using sentiwordnet for multilingual sentiment analysis.” IEEE 24th International Conference on Data Engineering Workshop, 2008.

[2] Paltoglou, Georgios, and Mike Thelwall, “A study of information retrieval weighting schemes for sentiment analysis” Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010.

[3] Caladoa, Edgar RochaaAlexandre P. Franciscoa Pável, and H. Sofia-Pintoa.” User profiling on Twitter”, 2011.

[4] Takaoka, Kouichi, and Akiyo NadamotoK. Elissa, “Title of paper if known,” unpublished.” Words-of-wisdom search based on multi-dimensional sentiment vector”, Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services. ACM, 2011.

[5] Haddi, Emma, Xiaohui Liu, and Yong Shi.” The role of text pre-processing in sentiment analysis.”, Procedia Computer Science, 2013.

[6] Kontopoulos, Efstratios and Berberidis, Christos and Dergiades, Theologos and Bassiliades, Nick.” Ontology-based sentiment analysis of twitter posts.”, Expert systems with applications, 2013.

[7] Montejo-Raez, Arturo and Martinez-C’amara, Eugenio and Martin-Valdivia, M Teresa and Urena-Lopez, L Alfonso.” Ranked wordnet graph for sentiment polarity classification in twitter”, Computer Speech & Languag, 2014.

[8] Abbasi, Mohammad-Ali, and Huan Liu.” Measuring user credibility in social media.”, International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction., 2013.

[9] Gupta, Aditi, and Ponnurangam Kumaraguru.” Credibility ranking of tweets during high impact events.”, Proceedings of the 1st workshop on privacy and security in online social media. ACM. 2012.

[10] Amiri, Fatemeh, Simon Scerri, and Mohammadhassan Khodashahi.” Lexicon-based sentiment analysis for Persian Text.”. Proceedings of the International Conference Recent Advances in Natural Language Processing. 2015.

[11] Sharma, Nitesh and Pabreja, Rachit and Yaqub, Ussama and Atluri, Vijayalakshmi and Chun, Soon and Vaidya, Jaideep.” Web-based application for sentiment analysis of live tweets.”. Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age..2018.

[12] Lauren, Paula and Qu, Guangzhi and Yang, Jucheng and Watta, Paul and Huang, Guang-Bin and Lendasse, Amaury.” Generating word embeddings from an extreme learning machine for sentiment analysis and sequence labeling tasks.”. Cognitive Computation.2018.

[13] Shirsat, Vishal S., Rajkumar S. Jagdale, and Sachin N. Deshmukh. "Sentence Level Sentiment Identification and Calculation from News Articles Using Machine Learning Techniques." Computing, Communication and Signal Processing. Springer, Singapore, 2019. 371-376.

[14] Iqbal, Farkhund and Hashmi, Jahanzeb Maqbool and Fung, Benjamin CM and Batool, Rabia and Khattak, Asad Masood and Aleem, Saiqa and Hung, Patrick CK.” A Hybrid Framework for Sentiment Analysis using Genetic Algorithm based Feature Reduction.”.IEEE Access.2019.

A Nested Genetic Algorithm for Mobile Ad-Hoc Network Optimization with Fuzzy Fitness

NourElDin S. Eissa[1], Ahmed Zakaria Talha[2], Ahmed F. Amin[3], Amr Badr[4]
Department of Computer Engineering
Arab Academy for Science, Technology and Maritime Transport (AASTMT), Cairo, Egypt[1, 2, 3]
Department of Computer Science, Cairo University, Cairo, Egypt[4]

***Abstract*—One of the major culprits that faces Mobile Ad-hoc networks (MANET) is broadcasting, which constitutes a very important part of the infrastructure of such networks. This paper presents a nested genetic algorithm (GA) technique with fuzzy logic-based fitness that optimizes the broadcasting capability of such networks. While normally the optimization of broadcasting is considered as a multi-objective problem with various output parameters that require tuning, the proposed system taps another approach that focuses on a single output parameter, which is the network reachability time. This is the time required for the data to reach a certain percentage of connected clients in the network. The time is optimized by tuning different decision parameters of the Delayed Flooding with Cumulative Neighborhood (DFCN) broadcasting protocol. The proposed system is developed and simulated with the help of the Madhoc network simulator and is applied on different realistic real-life scenarios. The results reveal that the reachability time responds well to the suggested system and shows that each scenario responds differently to the tuning of decision parameters.**

Keywords—Broadcasting; DFCN; fuzzy logic; genetic algorithms; Madhoc simulator; MANET

I. INTRODUCTION

Mobile Ad-hoc Networks (MANETs) are dynamic types of network consisting of an uncontrolled setup of end-point communication devices known as terminals, which are able of arbitrarily connecting with each other without the need of a base station or a fixed infrastructure [1]. The types of devices that are usually found in MANETs are laptops and smartphones equipped with limited range wireless technologies such as Bluetooth and WiFi (802.11). This, in turn, limits the communication capability of such devices, but allows them to move while communicating.

This makes the MANET very unpredictable as it needs to continuously self-reconfigure itself to accommodate these dynamic changes [2]. This is considered a major drawback for the efficiency and effectiveness of the MANETs and, by failing to readjust, link breakage will start to take place and some of the routes can become undiscoverable [3]. For the devices to be able to reach a certain destination, they start sending route discovery requests to their neighboring nodes [4] which, in turn, do the same thing. This results in the network being overwhelmed with an extreme amount of broadcast traffic known as a broadcasting storm [5].

Since it is clear that broadcasting plays a very critical role in network discovery and assists the nodes in MANETs in discovering their neighborhood [6], optimizing it constitutes a major step as it will save both energy and time, especially since most of the devices in the network have limited energy as they are battery powered.

Due to the previously mentioned limitations, a key threat known as node ‘selfish behavior’ arises in the network, in which the nodes purposely tend to drop the messages that do not target it, in an effort to save its energy [7] [8]. In other words, the nodes are not encouraged to contribute to the forwarding process. This kind of self-regarding behavior negatively impacts the network because, as already stated, there is no solid infrastructure in MANET and all the nodes rely on the cooperation of other nodes in the network to deliver and forward their messages. Delayed Flooding with Cumulative Neighbors (DFCN) is a broadcasting protocol that can handle this behavior and, at the same time, can reduce the number of packets that need forwarding with minimal punitive actions on the final coverage [9]. This is achieved by dropping the forwarded message when enough of the neighborhood devices have already got it. Also, once a node decides to forward a certain packet, it waits for a specified amount of time before executing this action, which is then canceled if another node in the network actually forwards the message [10].

The work proposed in this paper tackles a specific type of MANET, known as Metropolitan Mobile Ad-Hoc Networks, which is characterized by a disparate density that is continuously changing, whereas highly dense areas can swing from being active to inactive over short periods of time. Because creating a real testbed for this type of network is very costly and challenging, and might also lack the reproducibility factor, it was decided that the best approach to handle it is by means of a simulation framework. The Madhoc [11] simulator has been selected to achieve this. An evolutionary algorithm-based technique that combines nested GA with fuzzy-based fitness is proposed and implemented. The technique integrates the Madhoc simulator in its core and considers DFCN optimization over multiple real life mobility scenarios.

The rest of this paper is organized as follows. Section II introduces the Madhoc simulator and gives an insight about its capabilities and the different modes of operations. In Section III, a review of the related work concerning the optimization of broadcasting techniques in MANETs is presented. Section IV highlights the main problems that this research aims to solve. Section V demonstrates the algorithms and techniques used to solve the problem. Section VI shows the obtained results and discusses them. Finally, Section VII concludes this work and proposes the potential future work.

II. Madhoc Simulator

Madhoc is a metropolitan MANET simulator completely written in Java and available to use publicly [12] on the author's website [13]. The simulator provides the ability to simulate MANET using different parameters and real-life constraints such as working area size, mobility speed, wall thickness, etc. It also supports many different wireless technologies (e.g. WiFi, Bluetooth, GSM, etc.). Most importantly, it implements the full DFCN broadcasting protocol with all the required decision parameters to optimize it. Madhoc can be executed as a standalone application or as an Application Programming Interface (API).

To be able to collect the required statistics and results, a Madhoc monitor class is used. A monitor is not a part of the physical network and does not have an instance in real networks, and is regarded to as an abstraction entity that only exists at simulation level. It mainly aims at maintaining a global perspective on all nodes and for carrying out the required operations such as node deployment and initialization. It mainly serves as an observer of the Ad-hoc decentralized process. Another major attribute of the Madhoc simulator is that it does not use an event-driven simulation architecture, but instead, the simulator's kernel iterates upon a discrete time domain, where the distance between two intervals is known as the resolution.

This parameter is defined by the user and should be fixed throughout all the related applications to guarantee comparable and consistent results. The higher this value is, the less accurate the simulation will become. This value should be carefully used according to the required application. In the case of DFCN, this value must be at least twice lower than the maximum RAD, otherwise the benefits of using RAD will be completely lost.

Another important factor to consider while choosing the resolution is the mobility scheme of the nodes, the resolution must be small enough to make sure that the nodes move in reasonable steps, otherwise, some connections that could have taken place in real life would not be simulated.

III. Related Work

In the literature, most research has been dedicated to solving the broadcasting issues by using a multitude of different methods. Evolutionary multi-objective approaches have been proven to be effective in solving broadcasting problems [14], however, they suffer from time and performance issues [15]. Other methods focus on combinatory numerical models but most of them fail to adequately reduce the routing overhead with highly scalable networks, which is a main feature of MANET. Those who focused on the DFCN protocol did not formulate a trending mobility model for optimizing the decision parameters. Some of the researchers directly focused on detecting the selfish nodes in the network and avoiding them to increase the efficiency of the broadcasting protocols, the most notable work in this regard is by S. Subramaniyan et al. [16], where a Record-and-Trust-Based Detection (RTBD) technique was simulated that can efficiently detect selfish nodes in MANET. The main focus of this work was to accelerate the detection of misbehaving selfish nodes. The proposed method managed to diminish the overhead, latency and overhead ratio which improved the broadcasting performance of the MANET. However, the authors did not demonstrate how the acquired security could be transferred to the neighboring nodes in the network so that they could avoid being compromised by the selfish nodes detected by RTBD, meaning that the technique is not scalable on larger networks and the performance will be degraded. Another key focus in the literature is intelligent rebroadcasting techniques that reduce the overhead by estimating the usefulness of rebroadcasts and the probability of causing a collision. S. S. Basurra et al. [17] discussed a Zone based Routing with Parallel Collision Guided Broadcasting Protocol (ZCG) to reduce redundant broadcasting and to accelerate the path discovery process. The authors compared ZCG with two other techniques, Dynamic Source Routing (DSR) and Adhoc Ondemand Distance Vector Routing (AODV). It was concluded that ZCG can speed up the routing process in MANET due to its on-demand parallel collision guided broadcasting. However, the proposed method lacked distribution fairness among the nodes and did not protect zone members from selfish behavior attributed to the Zone Leader. Another interesting finding in the literature is the clustering of MANETs as a mean to reduce the complexity of the routing table. M. Ahmad et al. [18] provided a comprehensive survey about the different clustering algorithms that address this issue. It concluded that the effectiveness of the clustering algorithms depends on a set of specific parameters, which the nodes are remaining power, the relative mobility, the overhead data, the trust value, and the node reputation.

IV. Problem Statement

In order to optimize the DFCN protocol, multiple decision parameters need to be considered. These parameters dictate how DFCN operates and they characterize the search space. Since the optimization heavily relies on each specific scenario, an individual optimization trend is expected for each scenario.

The reachability time t_r is the output benchmark that is used to measure the optimization result. It is the amount of time required for the network to reach a certain number of pre-defined nodes. The goal of this research is to optimize the DFCN parameters to decrease the reachability time of the nodes inside the MANET. The problem is formulated as follows:

m: $instance\ of\ Madhoc\ simulator, t_r$: $reachability\ time.$

$$t_r = m(LowerRAD, UpperRAD, ProD, MinGain, SafeDensity) \quad (1)$$

$$f(LowerRAD, UpperRAD, ProD, MinGain, SafeDensity) = min\ (t_r)$$

The function f corresponds to the proposed system where the target is to minimize the reachability time t_r for each instance of the simulator m. Table I below shows the DFCN parameters along with their respective threshold and domain values.

TABLE. I. DFCN PARAMETER DESCRIPTION

Parameter Name	Domain	Description	Unit	Threshold Value
LowerRAD	Real (ℝ)	Minimum time required to rebroadcast.	Second	[0, UpperRAD]
UpperRAD	Real (ℝ)	Maximum time required to rebroadcast	Second	[LowerRAD, 10]
ProD	Integer (ℤ)	Maximum Density for which it is still required to use proactive behavior (reacting to new neighbors)	Device	[0, 100]
MinGain	Real (ℝ)	Minimum gain for rebroadcasting.	-	[0, 1]
SafeDensity	Integer (ℤ)	Maximum density, below which the protocol will always broadcast.	Device	[0, 100]

As already stated, this will be done on three different mobility model scenarios, namely, Highway, Mall and Human mobility. The description for these scenarios is shown next.

A. *Highway Scenario*

The main feature of the highway mobility model is that the nodes move at significantly higher speeds compared to the other mobility models and the nodes are lower in numbers. The spot density is also set to one spot per square kilometer, which is very sparse, and the number of spots per simulation area is limited to three. In this scenario, most of the generated traffic comes from nodes moving in opposite directions to simulate cars moving on different and opposing lanes of a highway. Table II below shows the properties of this scenario.

B. *Mall Scenario*

The mall mobility scenario is composed of separate regions connected by relatively narrow areas. It represents a group of shops interconnected using corridors. In this scenario, the surface area is smaller than the highway one and the velocity is much slower. Also, the nodes move randomly for most of the time with no clear targets, representing humans wandering around and shopping in arbitrary shops. Table III illustrates the different parameters for this scenario.

C. *Human Mobility Scenario*

This scenario is more distinctive than the mall one and is considered one of the most daunting models. In this context, the focus is on the human mobility scheme, where the movements are not random, but instead, there is a list of target destinations that each node mostly moves towards. These targets can be far away, as well as a few meters around. Also, the targets can dynamically change with time depending on human behavior. For instance, a waiter in a restaurant can be regularly moving back and forth between the kitchen and customers' tables.

TABLE. II. HIGHWAY MOBILITY SCENARIO PARAMETERS

Parameter	Value	Units
Surface Area	1 x 1	km^2
Nodes Density	80	nodes / km^2
Velocity	[20 40]	m.s^{-1}

TABLE. III. MALL MOBILITY SCENARIO PARAMETERS

Parameter	Value	Units
Surface Area	0.3 x 0.3	km^2
Nodes Density	6,500	nodes / km^2
Velocity	[0.3 1]	m.s^{-1}

TABLE. IV. HUMAN MOBILITY SCENARIO PARAMETERS

Parameter	Value	Units
Surface Area	0.05 x 0.05	km^2
Nodes Density	80,500	nodes / km^2
Velocity	[0.3 1.5]	m/s^{-1}

The human mobility scheme is defined as a round simulation area, where fixed places that act as target spots are scattered and where the distance between two places cannot be less than 10 meters. Table IV shows the parameters for the human mobility model.

V. PROPOSED SYSTEM

The proposed technique consists of nested GA with fuzzy-based fitness. The aim is to optimize the DFCN decision parameters according to the reachability time and to find certain trends for each one of the different scenarios. The benchmark used is the reachability time for 10% of the nodes, which is the time required so that 10% of the nodes in the network successfully deliver their messages. The outer GA contains the DFCN parameters and the to-be-calculated output from the simulator. The inner GA evolves a set of rules for the fuzzy system, where each chromosome represents a complete fuzzy set and the inference output represents the inner fitness. The final inner fitness value that is calculated after the convergence has completed sets the fitness value of the outer GA. The proposed system is developed using C# language on Microsoft Visual Studio 2017 under 64-bit Windows 10 with 8GB of RAM and an Intel Core i5-6500 CPU. Because the proposed system is built using C# and the Madhoc simulator operates fully in Java, a mechanism that interfaces them was required. To be able to accomplish this, each time the simulator is required to calculate the reachability time, it is executed by the developed application as a command line program running inside a virtual sandbox process, where all the standard inputs and outputs are redirected to the application. Fig. 1 shows an overview of the system.

Fig. 2 shows the pseudo-code of the proposed system.

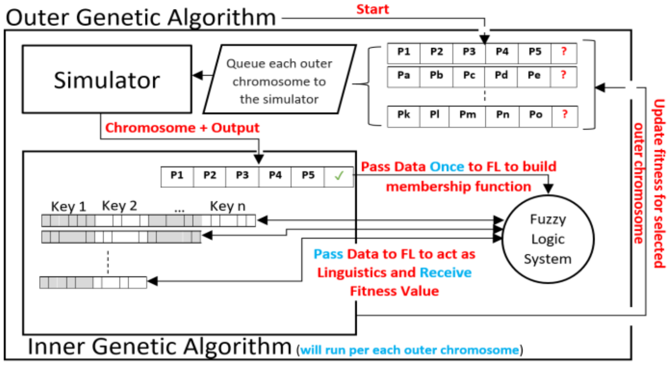

Outer Genetic Algorithm
Start
Simulator
Queue each outer chromosome to the simulator
P1 P2 P3 P4 P5 ?
Pa Pb Pc Pd Pe ?
Pk Pl Pm Pn Po ?
Chromosome + Output
Update fitness for selected outer chromosome
P1 P2 P3 P4 P5 ✓
Pass Data Once to FL to build membership function
Key 1 Key 2 ... Key n
Fuzzy Logic System
Pass Data to FL to act as Linguistics and Receive Fitness Value
Inner Genetic Algorithm (will run per each outer chromosome)

Fig. 1. Proposed System Illustration.

```
1  FUNCTION RunInnerGA(innerGenerationsCount, oChromosome)
2      i ← 0;
3      $P_i$ ← InitializeInnerPopulation(innerPopulationSize, keyMin, keyMax );
4      WHILE ( i < innerGenerationsCount -1 )
5          FL ← BuildFuzzySystem_VariableSets( oChromosome );
6          $P_{i+1}$ ← NULL;
7          j ← 0;
8          WHILE ( j < innerPopulationSize)
9              FL ← InitializeFuzzySystemLinguistics( $P_i$[ j ] );
10             Fitness( $P_i$[ j ] ) ← FuzzySystem_InferenceResult( );
11             j ← j + 1;
12         END WHILE
13         j ← 0;
14         WHILE ( j < innerPopulationSize / 2 - 1)
15             parents = RouletteSelect( $P_i$ );
16             offspring[0,1] = Crossover(parents, innerGACrossoverProbability);
17             $P_{i+1}$ ← $P_{i+1}$ + offspring[0,1 ];
18         END WHILE
19         j ← 0;
20         WHILE ( j < innerPopulationSize)
21             $P_i$[ j ] = Mutate( $P_i$[j ], innerGAMutationProbability);
22             j ← j + 1;
23         END WHILE
24         $P_{i+1}$ ← $P_{i+1}$ + GetFittest( $P_i$ );
25         i ← i + 1;
26     END WHILE
27     RETURN GetHighestFitnessValue( Pi );
28 END FUNCTION

30 FUNCTION RunOuterGA(outerGenerationsCount): MAIN
31     i ← 0;
32     $P_i$ ← InitializePopulation(outerGenerationsCount, thresholdsValues[ ]);
33     WHILE ( i < outerGenerationsCount -1 )
34         $P_{i+1}$ ← NULL;
35         j ← 0;
36         WHILE( j < outerPopulationSize)
37             Output( $P_i$[ j ]) ← GetMadhocOutput( $P_i$[ j ]);
39             Fitness( $P_i$[ j ] ) ← RunInnerGA ( $P_i$[ j ] );
40             j ← j+1;
41         END WHILE
42         j ← 0;
43         WHILE( j < OP_ outerPopulationSize  / 2 - 1)
44             parents = RouletteSelect($P_i$) ;
45             offspring[0,1] = Crossover(parents, outerGACrossoverProbability);
46             $P_{i+1}$ ← $P_{i+1}$ + offspring[0,1];
47         END WHILE
48         j ← 0;
49         WHILE ( j < outerPopulationSize)
50             $P_i$[ j ] = Mutate( $P_i$[ j ], outerGAMutationProbability);
51             j ← j + 1;
52         END WHILE
53         $P_{i+1}$ ← $P_{i+1}$ + GetFittest( $P_i$ );
54         i ← i + 1;
55         ExtractParametersAndOutput( $P_i$ );
56     END WHILE
57 END FUNCTION
```

Fig. 2. Pseudo-Code for the Proposed System.

The RunOuterGA function is the entry point of the program. The InitializeInnerPopulation function creates the initial population with random fuzzy logic keys that correspond to the linguistic strings. The oChromosome variable is the outer chromosome passed from the outer GA to the inner one, per generation. The keyMin and keyMax variables represent the range for the allowed number of keys per chromosome. At line 5, the fuzzy logic system is initialized and the fuzzy sets are created using the oChromsome genes, then at line 9, the linguistics are generated using the inner chromosome Pi[j], and finally at line 10, the fitness is calculated by getting the inference result for the developed fuzzy logic system.

The ExtractParametersAndOutput function is called per each outer GA generation to extract the current values of the decision parameters and the output from the fittest chromosome.

A. The Fuzzy Logic System

The fuzzy system is used to calculate the fitness for the inner GA. Each chromosome from the inner GA will act as complete fuzzy set. Each DFCN parameter will act as a linguistic variable with LOW, MED and HIGH as values. All of the variables have a triangular membership function that is equally divided over the maximum threshold of the respective parameters it represents. The rules for the fuzzy set are generated and optimized using the inner GA, which will be highlighted later.

In order to accomplish this, the inner chromosome is decoded from a numerical form to equivalent linguistic strings, according to Table V. To get the output values, the inference system uses a centroid defuzzifier with an interval of 1000. The interval represents the number of segments that the linguistic universe will be split into to perform the numerical approximation of the area center.

B. Outer Genetic Algorithm

The chromosome structure for the outer GA contains a hybrid of floating-point and integer values that correspond to the DFCN parameters, and also contain the output parameter which corresponds to the reachability time that will be calculated using the Madhoc simulator.

The chromosome size for the outer GA has a fixed length of six genes. Fig. 3 illustrates the chromosome structure. The crossover is a standard single-point operator that takes into consideration the gene placement to make sure the swapped parameters are still compatible and are within the specified thresholds. The mutation is performed through a non-uniform operator, which can be used to limit the lower and upper boundaries for the genes - which is crucial to avoid out-of-boundaries parameters - and also because it prevents the population from stagnating during the early evolution stages. The outer population size is fixed at 100 chromosomes and runs for a maximum of 300 generations. The crossover and mutation probabilities are fixed at 30% and 10%, respectively.

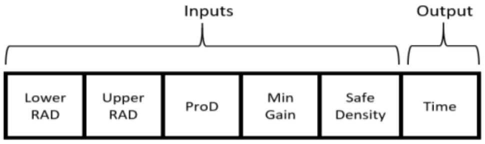


Fig. 3. Outer Chromosome Structure (Size=6).

TABLE. V. NUMERICAL-TO-LINGUISTIC STRING CONVERSION TABLE

Value	Equivalent Linguistic
1	LOW
2	MED
3	HIGH
-1	NOT LOW
-2	NOT MED
-3	NOT HIGH
0	NOT APPLICABLE

The selection is done through a traditional Roulette-Wheel operator. It is worth noting that the last gene (reachability time) is excluded from the evolution process and is stored inside the chromosome and passed later to the fuzzy system. All of the other aforementioned decision parameters are randomly generated within the threshold.

C. Inner Genetic Algorithm

The inner GA uses the same operators as the outer one. However, the chromosome structure is different. It consists of a variable number of genes ranging from 3 to 15. Each gene represents a key that encodes a linguistic string into numerical values as shown previously. This had to be done in order to be able to evolve the rules using the GA. Each key has a fixed length of 6 which corresponds to the number of input parameters and the output parameter.

The population size for the inner GA is set to 50 and the maximum number of generations is 100. Fig. 4 illustrates a sample inner GA with a population size of 7 and random chromosome sizes, denoted with S_n, where n is the chromosome number inside the population. It also shows an example of how the key is decoded into a linguistic string. The inner GA makes a complete run of 50 generations for each outer chromosome. The target is to diversify the linguistics of the fuzzy logic to reach the best possible output.

The defuzzified output value represents the fitness of the outer chromosome. After doing this for all the outer GA chromosomes, the best one is chosen and the outer GA transits into the next generation.

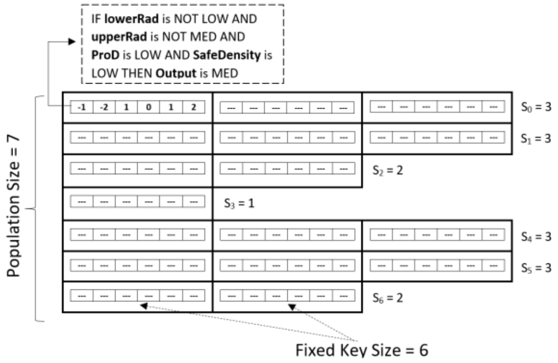

IF lowerRad is NOT LOW AND upperRad is NOT MED AND ProD is LOW AND SafeDensity is LOW THEN Output is MED
-1 -2 1 0 1 2
S₀ = 3
S₁ = 3
S₂ = 2
S₃ = 1
S₄ = 3
S₅ = 3
S₆ = 2
Population Size = 7
Fixed Key Size = 6

Fig. 4. Inner Genetic Algorithm Illustration.

VI. RESULTS AND DISCUSSION

The experiments are run five times and the results are averaged. The results show the convergence of decision parameters and the output (solid black line). The logarithmic trendline (red dotted line) is also calculated to provide a mathematical model for the decision parameters. Fig. 5 shows the results for the highway mobility environment. Table VI shows the output trendline for each decision parameter and the equivalent logarithmic regression expressions.

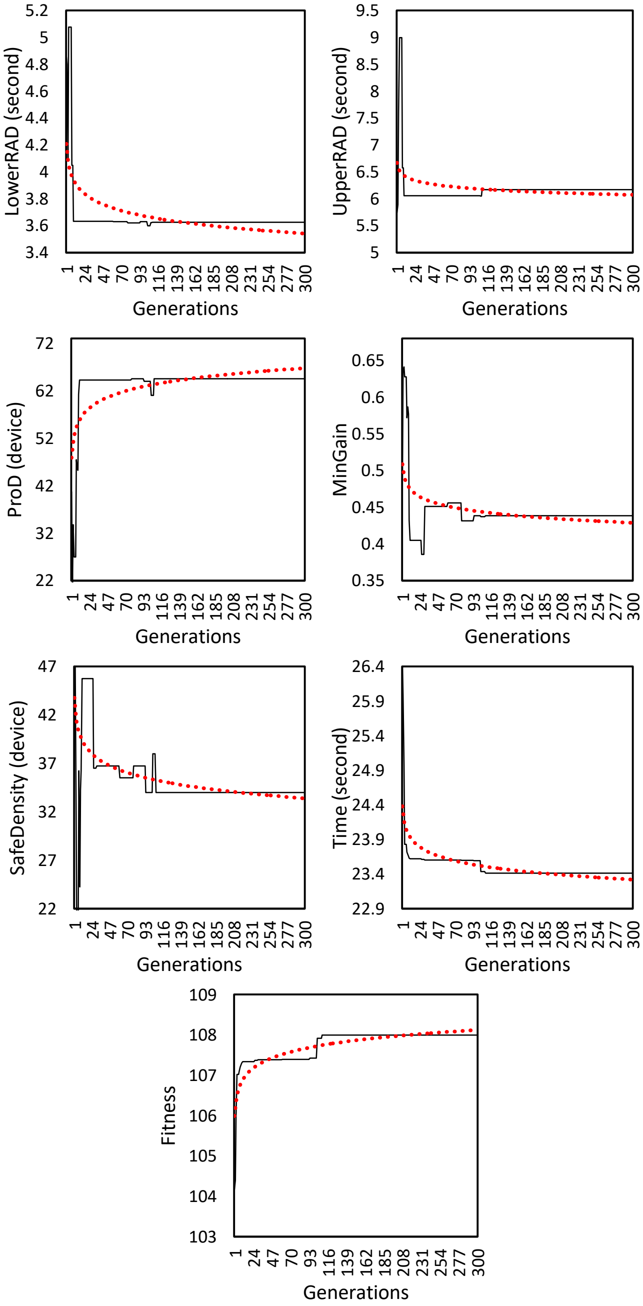


Fig. 5. Convergence for the High Way Mobility Model.

TABLE. VI. TRENDLINE PARAMETERS FOR HIGHWAY SCENARIO

Parameter	Trendline	Expression
LowerRAD	↓	$-0.117 * ln(G) + 4.2076$
UpperRAD	↓	$-0.105 * ln(G) + 6.6723$
ProD	↑	$3.3079 * ln(G) + 47.885$
MinGain	↓	$-0.014 * ln(G) + 0.5087$
SafeDensity	↓	$-1.818 * ln(G) + 43.76$

Fig. 6 shows the results for the Mall mobility scenario and Table VII shows the trendline for the decision parameters. The results for the human mobility model are shown in Fig. 7 and the respective trendline parameters are shown in Table VIII.

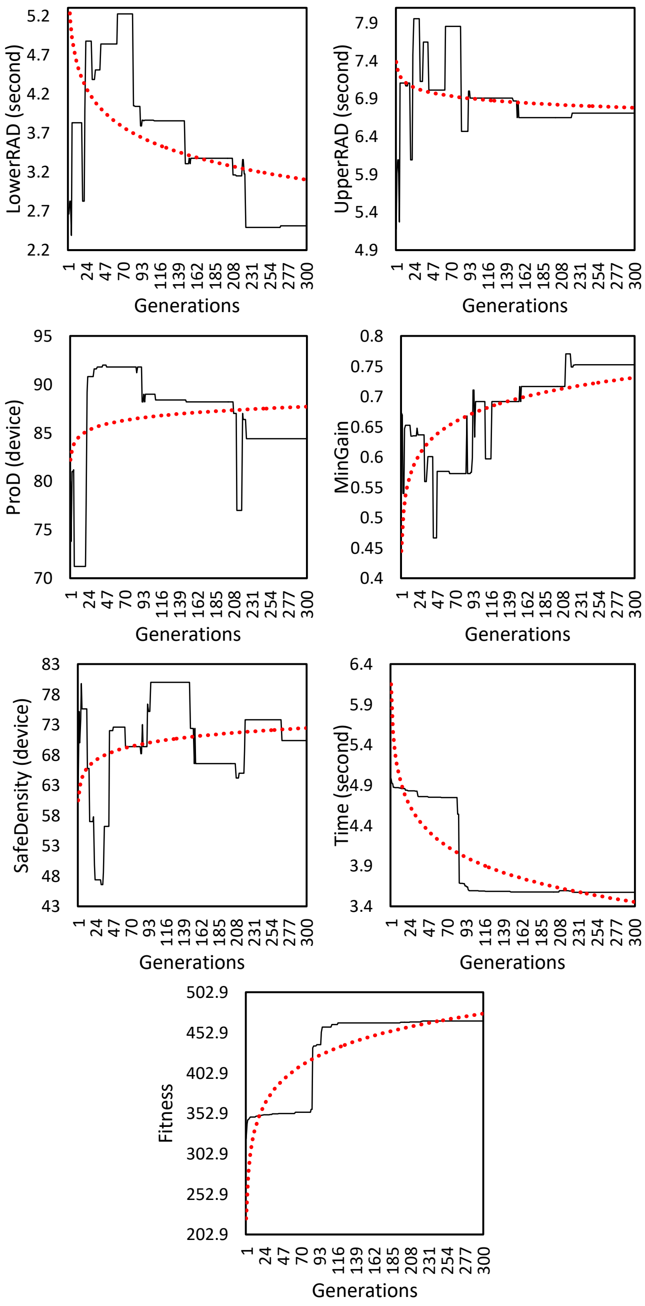


Fig. 6. Convergence for the Mall Mobility Model.

TABLE VII. TRENDLINE PARAMETERS FOR MALL SCENARIO

Parameter	Trendline	Expression
LowerRAD	↓	$-0.463 * ln(x) + 5.7409$
UpperRAD	↓	$-0.106 * ln(x) + 7.3855$
ProD	↑	$0.9608 * ln(x) + 82.223$
MinGain	↑	$0.0502 * ln(x) + 0.445$
SafeDensity	↑	$2.0984 * ln(x) + 60.488$

In the highway mobility scenario, the time to reach the destination decreased from 26.44 to 23.41 seconds, which amounts to 11.45%. Given that the number of nodes in this network is 80, the average time for a node to deliver a message decreased from 3.3 to 2.92 seconds.

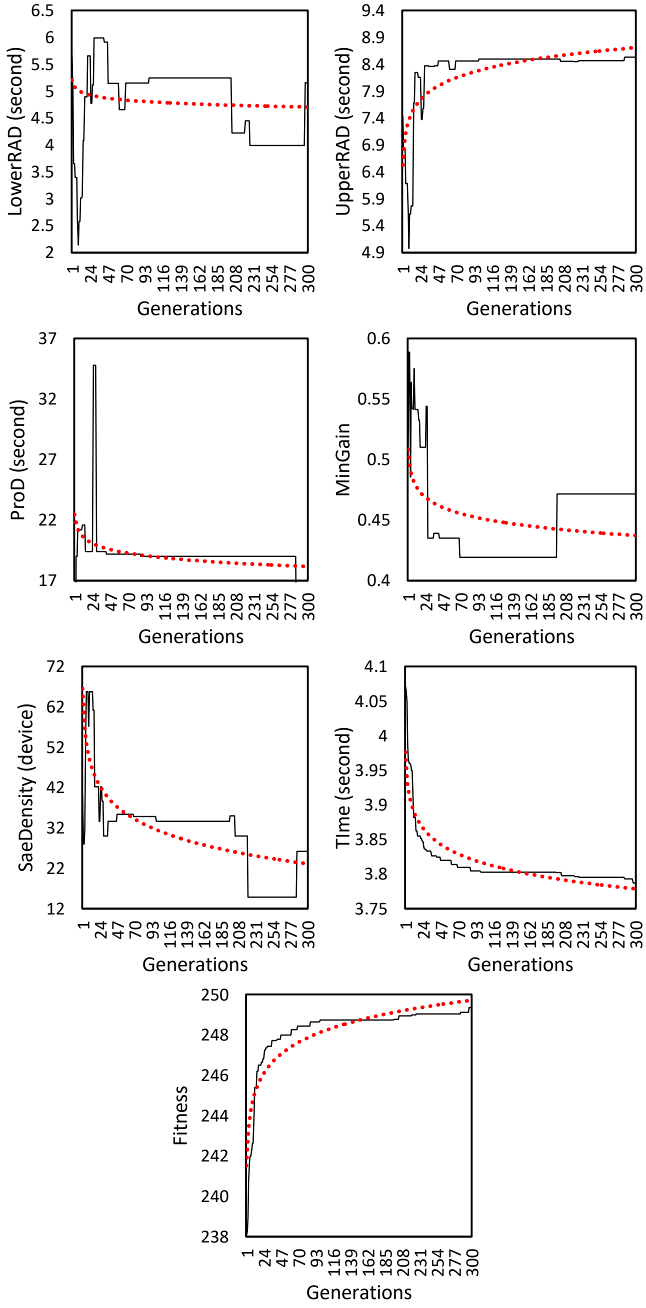


Fig. 7. Convergence for the Human Mobility Model.

TABLE VIII. TRENDLINE PARAMETERS FOR HUMAN MOBILITY SCENARIO

Parameter	Trendline	Expression
LowerRAD	↓	$-0.088 * ln(x) + 5.2098$
UpperRAD	↑	$0.3832 * ln(x) + 6.526$
ProD	↓	$-0.757 * ln(x) + 22.479$
MinGain	↓	$-0.012 * ln(x) + 0.5084$
SafeDensity	↓	$-7.621 * ln(x) + 66.544$

For the mall mobility scenario, the time to reach the nodes decreased from 4.98 seconds to 3.57 seconds which amounts to 28.3%, which brings down the average required time to deliver a message from 7.6ms to 5.49ms.

As for the human mobility scenario, the time to deliver the messages to their respective destinations decreased from 4.07 to 3.78 seconds, which amounts to 7.12%. The average time to deliver a message decreased from 0.5ms to 0.46ms.

By inspecting all the previous results, it appears that the mall mobility model benefited the most from the optimization of the DFCN decision parameters and the human mobility model benefited the least. While these two models have very close features, the major difference between them, as stated previously, is the randomness of the movements. The human mobility model is governed by human intentions of moving between a dynamic list of targets while the mall one is governed by random motion of shoppers moving between random shops. Also, by inspecting the highway scenario, it seems that the lack of enough nodes has significantly raised the average delivery time six times (6x) the delivery time in other scenarios.

To demonstrate the consistency of the results, a 5% confidence interval for the final reachability time is calculated and is shown in Table IX.

TABLE. IX. 5% CONFIDENCE INTERVAL FOR THE FINAL REACHABILITY TIME

Mobility Model	**5% Confidence Interval (seconds)**
Highway	23.4 ± 0.75
Mall	3.57 ± 0.34
Human	3.78 ± 0.02

VII. CONCLUSION AND FUTURE WORK

The proposed system managed to decrease the message delivery time for the three real-life scenarios (the highway, the mall and the human mobility models) by optimizing the decision parameters for the DFCN protocol. The mall mobility model benefited the most from the optimization of the DFCN parameters, which is mainly attributed to the randomness of the mobility, since the human mobility model also shares very close parameters but only differs in the movement intention. In the human mobility model, the mobility is governed by the intentions of the humans to reach a certain dynamic list of destinations and, therefore, the randomness significantly decreases. Also, the highway mobility model yielded the highest average message delivery time, which is attributed to the lack of nodes and the very high mobility speed, and since the DFCN protocol relies on 1-hop neighbors to deliver the messages to their destinations, this scenario severely affects it.

In the future, a mathematical model based on the found trendlines can be established and tested. This will help to achieve the results faster, instead of relying solely on metaheuristic techniques, which require a significant amount of time to converge to the optimal solution.

Also, Genetic Programming (GP) can be experimented with, to evolve programs and expressions related to each scenario. This way, the resulting programs can be used as a rigid optimization model, without the need to repeat the evolution process each time.

REFERENCES

[1] V. Rishiwal, S. K. Agarwal and M. Yadav, "Performance of AODV protocol for H-MANETs," in International Conference on Advances in Computing, Communication, & Automation (ICACCA) (Spring), Dehradun, India, 2016.

[2] L. J. G. Villalba, J. G. Matesanz, A. L. S. Orozco and J. D. M. Díaz, "Auto-Configuration Protocols in Mobile Ad Hoc Networks," Sensors (Basel), vol. 11, no. 4, p. 3652–3666, 2011.

[3] C. Dhakad and A. S. Bisen, "Efficient route selection by using link failure factor in MANET," in International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, India, 2016.

[4] P.-J. Chuang, P.-H. Yen and T.-Y. Chu, "Efficient Route Discovery and Repair in Mobile Ad-hoc Networks," in IEEE 26th International Conference on Advanced Information Networking and Applications, Fukuoka, Japan, 2012.

[5] V. Sharma and A. Vij, "Broadcasting methods in mobile ad-hoc networks," in 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 2017.

[6] M. Bakhouya, "Broadcasting approaches for Mobile Ad hoc Networks," in International Conference on High Performance Computing & Simulation (HPCS), Helsinki, Finland, 2013.

[7] H. Yadav and H. K. Pati, "A Survey on Selfish Node Detection in MANET," in International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida (UP), India, 2018.

[8] N. Ramya and S. Rathi, "Detection of selfish Nodes in MANET - a survey," in International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2016.

[9] L. Hogie, P. Bouvry, M. Seredynski and F. Guinand, "A Bandwidth-Efficient Broadcasting Protocol for Mobile Multi-hop Ad hoc Networks," in International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies (ICNICONSMCL), Morne, Mauritius, 2006.

[10] B. Dorronsoro, P. Ruiz, G. Danoy, Y. Pigné and P. Bouvry, "BROADCASTING PROTOCOL," in Evolutionary Algorithms for Mobile Ad Hoc Networks, John Wiley & Sons, Inc, 2014, pp. 135-138.

[11] L. Hogie and P. Bouvry, "An Overview of MANETs Simulation," Electronic Notes in Theoretical Computer Science, vol. 150, no. 1, pp. 81-101, 2006.

[12] L. Hogie, F. Guinand and P. Bouvry, The Madhoc metropolitan ad hoc network simulator, France: Luxembourg University and Le Havre University, 2006.

[13] L. Hogie, "Madhoc Metropolitan ad hoc network simulator," 2006. [Online]. Available: http://www.i3s.unice.fr/~hogie/madhoc/. [Accessed 25 January 2019].

[14] R. M. Chintalapalli and V. R. Ananthula, "M-LionWhale: multi-objective optimisation model for secure routing in mobilead-hocnetwork," IET Communications, vol. 12, no. 12, pp. 1406 - 1415, 2018.

[15] E. Alba, B. Dorronsoro, F. Luna and P. Bouvry, "A cellular multi-objective genetic algorithm for optimal broadcasting strategy in metropolitan MANETs," in IEEE International Parallel and Distributed Processing Symposium, Denver, CO, USA, 2005.

[16] S. Subramaniyan, W. Johnson and K. Subramaniyan, "A distributed framework for detecting selfish nodes in MANET using Record- and Trust-Based Detection (RTBD) technique," EURASIP Journal on Wireless Communications and Networkingvolume, p. Article 205, 2014.

[17] S. S. Basurra, M. D. Vos, J. Padget, Y. Ji, T. Lewis and S. Armou, "Energy Efficient Zone based Routing Protocol for MANETs," Ad Hoc Networks, vol. 25, pp. 16-37, 2015.

[18] M. Ahmad, A. Hameed, A. A. Ikram and I. Wahid, "State-of-the-Art Clustering Schemes in Mobile Ad Hoc Networks: Objectives, Challenges, and Future Directions," IEEE ACCESS, vol. 7, pp. 17067 - 17081, 2019.

Identification of Issues and Challenges in Romanized Sindhi Text

Irum Naz Sodhar[1]
Post Graduate Student, Department of Information Technology, Quaid-e-Awam University of Engineering Science and Technology, Nawabshah, Sindh, Pakistan

Muhammad Ibrahim Channa[3]
Professor, Department of Information Technology
Quaid-e-Awam University of Engineering, Science and Technology, Nawabshah, Sindh, Pakistan

Akhtar Hussain Jalbani[2]
Associate Professor, Department of Information Technology, Quaid-e-Awam University of Engineering Science and Technology, Nawabshah, Sindh, Pakistan

Dil Nawaz Hakro[4]
Associate Professor, Institute of Information and Communication Technology, (IICT)
University of Sindh, Jamshoro, Sindh, Pakistan.

***Abstract*—Now-a-days Sindhi language is widely used in internet for the various purposes such as: newspapers, Sindhi literature, books, educational/official websites and social networks communications, teaching and learning processes. Having developed technology of computer system, users face difficulties and problems in writing Sindhi script. In this study, various issues and challenges come in the Romanized Sindhi text by using Roman transliteration (Sindhi text (ST) forms of Romanized Sindhi text) are identified. These acknowledged issues are known as noise, written script of Romanized and its style, space issues in Romanized script, some characters not suitable in Romanized Sindhi, as a paragraph, rows, character issues, punctuation, row break and font style. However, this study provides the summary of issues and challenges of Romanized Sindhi text. This research work provides detailed information of issues and challenges faced by people during chatting in Romanized Sindhi text.**

Keywords*—Romanized Sindhi Text (RST); Sindhi language; issues and challenges; transliterator; social networks communication

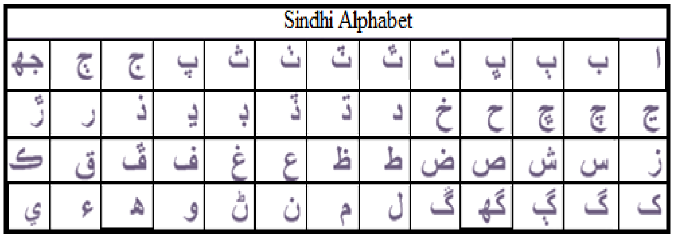
I. Introduction

Sindhi Language is a historical language of the world, the majority of Sindhi language speakers are inhabited in Sindh province of Pakistan. Around 12% peoples of Pakistan have mother tongue is Sindhi and an official language of the Sindh [1]. Sindhi language is also spoken in different part of the world with different ratio. Sindhi language has its own script and written format. In Sindhi Language 52 alphabetical letters (Fig. 1) were used for written as well as in speaking purposes [2-4]. Since, Sindh language contains more alphabetical letters than other languages, which causes difficulties for the new learners. Sindhi script writing is a right handed script, same as Arabic and Urdu Script. Urdu is morphological prosperous, having different type of characters in Urdu script. Sindhi script follows the rules as like Arabic Script and Perso-Arabic script [5].

In these days Sindhi language is considered as extensively used in internet for the various purposes such as: daily newspapers, Sindhi literature, books, educational/official websites, social network communication (What's App, Text messages, and social network), Teaching and learning processes. In this regard, the use of the keyboard (Sindhi) is being increased day by day and on the other hand people are still facing the problem of unavailability of Sindhi keyboards. However, the communication system of local users is carried on by android based mobile phones services; these mobile phones are unable to provide facilities to write Sindhi language containing 52 letters. Therefore, to overcome these problems, Romanized Sindhi text is one of the best options [6].

Romanized Sindhi text is when used in different plate forms may face many issues and problems in writing of Romanized Sindhi text or when use of different translators for Sindhi of Romanized text. Also the use of translators and other sources for normal users are very difficult and they need an easy way for the solution of the problem.

New issues and challenges of Sindhi language has been found, when communicating with each other in Romanized Sindhi text because Sindhi language has 52 letters of alphabets having different shapes, different symbols and different orientation of dots. So, it is very difficult to communicate using Sindhi language on different social media. Therefore it is very important to have such platforms where people of different Sindhi community can communicate easily and properly using Romanized format.



Sindhi Alphabet												
جھ	ڄ	ج	پ	ث	ٺ	ٽ	ٿ	ت	ڀ	ٻ	ب	ا
ڙ	ر	ذ	ڍ	ڊ	ڏ	ڌ	د	خ	ح	چ	ڇ	ڃ
ڪ	ق	ڦ	ف	غ	ع	ظ	ط	ض	ص	ش	س	ز
ي	ء	ھ	و	ڻ	ن	م	ل	ڱ	گھ	ڳ	گ	ک

Fig. 1. Sindhi Language Alphabet.

II. Related Work

In Sindhi data set construction, issues contain corpus acquisition; pre-processing and tokenization is discussed in this paper. The results of those issues based on observation which contains unigram, bigram and trigram frequencies; author explores the orthography and Sindhi script data construction [7]. The word corpus was used by German Scholar at first time. The plural of corpus was corpora, which was used for a huge number of text data consists of either millions or billions of data. Processing was challenging because scarcity of resources for computational linguistics and research, different text data have been developed in different languages of different countries [6].

A model for transliteration was provided by Leghari and Arain, this model provided two scripts of Sindhi language one was Perso-Arabic Script and other was Devanagari Script. Analyzing of both scripts, authors suggested that data on Roman Script also used for Sindhi Language and they proposed an algorithm for transliteration between two scripts [4].

In another research paper authors addressed the issues of Sindhi word Segmentation and provide different techniques to implement on different algorithms [8]. Current research on multi linguistic writing has been carried out in transliteration. Authors in [9] explored English related forms, writing with Romanized Greek characters. Authors in [5] fond out the challenges in Urdu text and tokenized the Urdu text and also detected the sentence boundary. This was very difficult task for comparison of tokenization and detected the sentence boundary.

OCR is an Optical Character Recognition used in written text or (as well as) in printed documents. Authors in [4] found out the issues and challenges in Sindhi OCR which contains many character dots, different placement and direction of dots. The authors also provided the summary for issues and challenges related to the development of Sindhi OCR.

A research work was done on the sentiment analysis of Arabic Language facing an issue that was unstructured and non-grammatical text. Results were analyzed by using various parameters: accuracy, precision, recall and F-score [10]. Authors in [11, 12] also worked on the sentiment analyses of Arabic language by using support vector machine technique. This technique was more accurate for classification for Sentiment summarization and analysis of the Sindhi text by using machine learning techniques DTM and TF-IDF. DTM and TF-IDF analysis was used by n-gram model. The supervised machine learning model was mostly used for Sindhi text Sentiment analysis [13].

Authors in [14] worked on the sentiment analysis of Urdu text by using sentiment classification model. This system extracts Senti Units and the target expressions through the shallow parsing based chunking. It is observed that dependency parsing algorithm created associations between these extracted expressions and measure the results either positive or negative.

Dootio and Wagan did research on the Development of Sindhi alphabet and reported that Sindhi language is widely used in all over the world. They added that mostly its literature is used in printed forms such as in books, in newspaper, in online learning websites and in different web pages on internet to construct Sindhi data set. The authors also used NLP techniques and developed the Sindhi text corpora for the use of Sindhi script [1].

Form literature, it is observed that there is still a huge space available for the research in Sindhi language to improve its written format. It is concluded from the literature that Sindhi script is widely used, but many issues and challenges come in writing forms different sources. But Romanized Sindhi text is also one of the ways to reduce the issues and challenges in Sindhi Text.

III. Materials and Methods

A. Data Set of Sindhi Script

In this study Sindhi script was used in the Romanized Sindhi Text. This Sindhi script was transliterated by online tools which are easily available. After the translation of Sindhi script into Romanized Sindhi text was checked for correct transliteration text and for errors in transliteration as shown in Fig. 2. In this relation, Sindhi data were selected from different sources and are easily available in online sources such as: newspapers, Poetry websites, Sindhi Facebook pages, Text messages, etc. The sources used in this study were then verified from the various sources and updated on a daily basis as described in Table I.

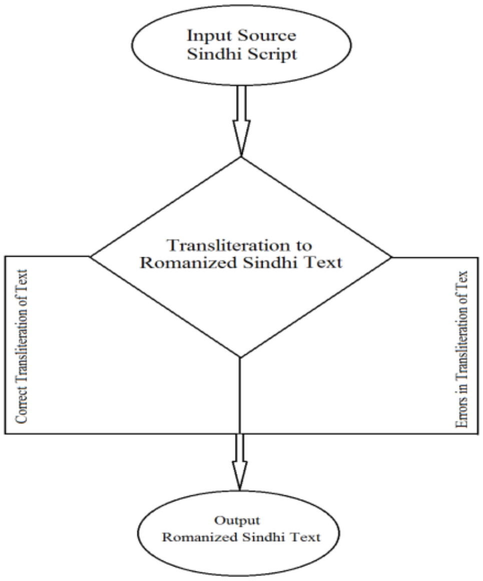


Fig. 2. Research Methodology Diagram.

TABLE. I. Sindhi Data Resources

Resources	Websites
Awami Awaz	https://awamiawaz.pk
Jhoongar	http://dailyjhoongar.com
Sindhi Poetry in roman	http://romansindhi.blogspot.com/2017/
Sindhi Adabi Board	http://www.sindhiadabiboard.org/Catalogue/Poetry/Book92/Book_page19.html
Sindhi Learning	http://sindhila.edu.pk

B. Dots in Sindhi Script

As Sindhi script has fifty two (52) alphabetical Letters and all letters Sindhi words along with Romanized words are given in Table II. In Table II, four letters have no words to start with these letters (ڱ,ڃ ,ڻ ,ڙ) those letters were used in middle or end of words (used for complete the words) as shown in Fig. 3.

Arabic script has 28 characters, Urdu has 39 characters while Sindhi script has 52 characters. Sindhi script is the right handed language as like Arabic and Urdu. The use of Dots is very important in these scripts, otherwise the letters had no any sense and meaning and very difficult to pronounce these letters. Arabic script is mostly used up to three dots, but in Sindhi script four different types of dots (one-1), (Two-2), (Three-3) and (Fourth-4) are used. These dots are placed at different position in letters such as above, below and inside the letter and dots are placed in letters at horizontal or vertical direction. In Sindhi script a few words have many dots in a single word just like (fifar in the Romanized Sindhi text and Lung in English), 12 dots are used in this (fifar) single word of Sindhi script.

Thirty four letters are used with 1 to 4 dots in different axis and positions in Sindhi alphabet. Table III shows details of the Characters of Sindhi script with dots, one letter used small symbol and without dots are presented in this table.

Single dot letters in Sindhi script is shown in Table IV, single dot is placed in letters below, above and inside is used. A total of nine (9) letters having single dot is shown in below Table IV. Double dots or two dots are used in horizontal or vertical direction and placed below, above and inside of the letter. Total eight (8) letters having double dots are used as shown in below Table V.

There are seven (7) letters, which take three dots were used in Sindhi script and placed in three positions below, above and inside of the letter shown in Table VI. Below placement of dots in letter is only (one-1), dots in the horizontal direction three (3) letters are placed, dots in the above position in letters are only (two). The characters having dots in the direction vertical and inside placement is only (1) one. In Sindhi script four dots are also used in characters as shown in Table VII. Total five (5) letters having four dots were used in above (3), below (1) and inside (1) the letter.

ڻ
وڻ
Wanu
Tree

Fig. 3. Sindhi Letter Wanu.

TABLE. II. SINDHI CHARACTERS WITH SINDHI WORDS AND ROMANIZED SINDHI WORDS

S. No.	Sindhi Letters	Word Start with Letter	Romanized Words
1	ا	انب	Unab
2	ب	بدڪ	Badk
3	ٻ	ٻڪ	buk
4	پ	پڳ	pag
5	ڀ	ڀت	bhiti
6	ت	تارو	taro
7	ٿ	ٿلهو	thulho
8	ٽ	ٽبي	ttbi
9	ٺ	ٺونٺ	thonth
19	ث	ثواب	Sawab
11	ج	جبل	jabal
12	جھ	جھرڪي	jharki
13	ڄ	ڄڀ	jibh
14	ڃ	ڃڃ	janj
15	چ	چپ	chupu
16	ڇ	ڇڇ	chahj
17	ح	حجم	hajjm
18	خ	خچر	khchr
19	د	در	daru
20	ڌ	ڌوٻي	dhobi
21	ڊ	ڊيل	Del
22	ڏ	ر ڏيڏ	dedr
23	ڍ	ڍور	dhor
24	ذ	ذرو	zaro
25	ر	رازو	Razo
26	ڙ	رڙ	Radh
27	ز	زمين	Zamen
28	س	ساز	Saaz
29	ش	شخص	Shakis
30	ص	صوف	Soof
31	ض	ضعيف	Zaif
32	ط	طوطو	Toto
33	ظ	ظلم	Zulum
34	ع	عينڪ	Ayeenak
35	غ	غلام	Ghulam
36	ف	فروش	Farosh
37	ڦ	ڦوهارو	Foharo
38	ق	قلم	Kalam
39	ڪ	ڪعبو	Kabo
40	ک	کٿ	Khat
41	گ	گانو	Gano
42	ڳ	ڳئون	Gaon
43	گھ	گھوڙو	Ghoro
44	ڱ	سڱ	Singhan
45	ل	لڪير	Lakeer
46	م	مفهوم	Mafhoom
47	ن	نالو	Nalo
48	ڻ	وڻ	Wandh
49	و	وڏيڪ	Wadheek
50	ھ	هڪڙو	Hikro
51	ي	يگانو	Yagano
52	ء	No start word	Hamzo

TABLE. III. DOTS USED IN SINDHI ALPHABET SYMBOLS

Sindhi Letters	Number of letters	Dots	Symbols	Total letters
ب. ج , جھ , خ , ڊ , ذ , ز , ض , ظ. غ ,ف , ن	12	One dot or Single dot used	Not used	12
ڄ , ڃ , ڌ , ڍ , ق , ,ت , ٽ , ڳ , گ , ي	11	Two dots or double dots used	Not used	11
پ , ٿ , ث , چ , ڏ , ش	6	Three dots or triple dots used	Not used	6
ڀ , ٺ , ڇ , ڙ , ڦ	5	Four dots used	Not used	5
ڻ	1	Not used	(ط)	1
س, ص, ط, ع ,ڪ,ک, گ, گھ,ل, م , و,ه , ء , ا , ح , د , ر,	17	Not used	Not used	17
Total Letters	52			

TABLE. IV. SINGLE DOT LETTER IN SINDHI ALPHABET

S. NO:	Placement	Number of Letter	Placement of Dots
1.	Above or Up	04	خ , ذ , ز , غ
2.	Below or Down	03	ب , جھ , ڊ
3.	Insider or Within	02	ج , ن

TABLE. V. DOUBLE DOTS CHARACTER IN SINDHI ALPHABET

S. NO	Placement of in dots in letters	Direction	
		Horizontal	Vertical
1.	Above letters	ت , ڳ	ٺ
2.	Below letter	ي	ڳ , ٻ
3.	Inside letter	ڃ	ڄ

TABLE. VI. THREE DOTS CHARACTER IN SINDHI ALPHABET

S. NO:	Placement of in dots in letters	Direction	
		Horizontal	Vertical
1.	Above letters	ٿ,ث , ش	ٿ , ش
2.	Below letter	پ	NA
3.	Inside letter	چ	NA

TABLE. VII. FOUR DOTS CHARACTER IN SINDHI ALPHABET

S. NO.	Placement	Letters
1.	Above letters	ڦ, ڙ , ٿ
2.	Below letter	ڀ
3.	Inside letter	ڇ

IV. ISSUES AND CHALLENGES IN ROMANIZED SINDHI TEXT

A. Noise Letters

Use of dots in Sindhi letters shows the appearance and pronunciation of the alphabet. Dots are also used in the meaning of the sentence. Sindhi script contains (Seventeen-17) letters are used without dots and (Thirty Five-35) letters are used with dots. But few letters have an important appearance in words for their correct noise and meaning as shown in Table VIII. When these letters used in Sindhi words, sentences, paragraphs create problem in converting into the Romanized text by using transliterate. Letter ڻ have no doubt used as in letters/words, but a small symbol used in the letters in Sindhi script as shown in Fig. 4. Due to this reason it is very difficult to recognize the use of this letter in transliteration in the Romanized Sindhi text. Above Table II shows there is no any word on Sindhi Script start with ڻ, but it was used in the middle of the words or end of the words.

B. Font of Letters

Font of the letter is imported part of written communication and beautification of letters according to the situation, but in transliteration of text has no any facility available of font family for the users need. The font is one of the main issues in transliteration of Sindhi text into the Romanized Sindhi Text as shown in Table VIII.

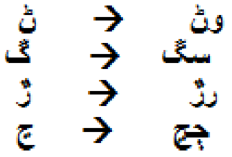


Fig. 4. Noise Letters.

TABLE. VIII. SUMMARY OF ISSUES AND CHALLENGES

S. NO	Issues and Challenges	Explanation
1	Written script with Romanized style as the English style	Left Handed Romanized Script and same as English Example: **Hinaa: aslaamu alekum** **حنا: السلام عليكم**
2	Space Issues in Romanized Script	When no space in a row, so word completely not show in the same row, but the word must be tokenized last letter into next line that's why problem occur in reading text for readers. Example: **نئين مالي سال 20-2019ع جي ساليانی بجيٽ پيش ڪندي چيو** naen maale saala 20-2019a je saalaeunae bjett peshu kande chayo
3	Some Characters are not suitable in Romanized Sindhi	Hamzo and Wao, NN, Dhe, Example: **Asmaa: khhudaa hafiz** **اسماء: خداحافظ**
4.	Paragraph	Not show properly according to paragraph
6	Row	Not Identify either row is complete or next line start.
7	Characters Issues	Sindhi characters have 52 and Roman English have 26 characters so feel difficult to pronunciation of complex word. Example: **توهان→ t~vhaa`n**
8	Punctuation	Comma, Question, Double Quotation and so on show revert in translating text that's why feel difficult to read.
9	Row Break	غريب→ Gwhen this word comes row breakdown and other words shows on the next line.
10	Font Style	No any facility available to change the text in different Font style.

C. Punctuation of Letters

Fourteen-14 punctuation is used frequently in text communication, but much punctuation is changed their positions as well as axis after transliteration Sindhi text into Romanized text. After the transliteration punctuations are not changed, but they are still in the same condition as before transliteration as shown in Table VIII. These issues come in punctuation because translators do not follow the punctuation rules when the Sindhi script is converted into the Romanized script.

In Table VIII, different issues and challenges are presented, these issues and challenges are almost occurs after transliteration of Sindhi text into Romanized Sindhi text by using online converters.

V. Conclusion

Sindhi Script is morphological rich in literature; it is written and spoken by worldwide. Now-a-days Romanized Sindhi text is used as communication purpose. The written style of Sindhi script is right handed as same as Arabic and Urdu written style. In writing of Romanized text there is no use of dots or small symbol, but still limited resources are available for Romanized Sindhi text online conversion. Many issues and challenges were found during Sindhi script translated into Romanized Sindhi Text which are: written script of Romanized and style, space issues with Romanized script, some characters not suitable in Romanized Sindhi, paragraph, rows, character issues, punctuation, row break and font style were observed. This research work may be very helpful for sentiment analysis and summarization of Romanized Sindhi text in the future.

References

[1] M. A. Dootio, & A. I. Wagan, “Development of Sindhi text corpus,” Journal of King Saud University-Computer and Information Sciences, 2019.

[2] A. Pirzado, "Sindhi language and literature (a brief account)," Hyderabad, Sindh: Sindhi language Authority (2009).

[3] M. Leghari, & M. U. Rahman, “Towards Transliteration between Sindhi Scripts by using Roman Script,” In Conference on Language and Technology, 2010.

[4] D. N. Hakro, I. A. Ismaili, A. Z. Talib, Z. Bhatti, & G. N. Mojai, “Issues and challenges in Sindhi OCR,” Sindh University Research Journal (Science Series), 46(2), 143-152, 2014.

[5] Z. Rehman, W, Anwar, & U. I. Bajwa, “Challenges in Urdu text tokenization and sentence boundary disambiguation,” In Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP), pp. 40-45, 2011.

[6] F. H. Khoso, M. A. Memon, H. Nawaz, & S. H. A. Musavi, (2019). To Build Corpus Of Sindhi.

[7] M. U. Rahman, "Towards Sindhi corpus construction," In Conference on Language and Technology, Lahore, Pakistan. 2010.

[8] Z. Bhatti, I. A. Ismaili, W. J. Soomro, & D. N. Hakro, “Word segmentation model for Sindhi text,” American Journal of Computing Research Repository, 2(1), 1-7, 2014.

[9] T, Spilioti, “From transliteration to trans-scripting: Creativity and multilingual writing on the internet,” Discourse, Context & Media, 29, 100294, 2019.

[10] A. Assiri, A. Emam, & H. Aldossari, “Arabic sentiment analysis: a survey,” International Journal of Advanced Computer Science and Applications, 6(12), 75-85, 2015.

[11] A. Ziani, N. Azizi, D. Zenakhra, S. Cheriguene, & M. Aldwairi, “Combining RSS-SVM with genetic algorithm for Arabic opinions analysis,” International Journal of Intelligent Systems Technologies and Applications, 18(1-2), 152-178, 2019.

[12] H. Al Suwaidi, T. R. Soomro, & K. Shaalan, “Sentiment analysis for emiriti dialects in twitter,” Sindh University Research Journal-SURJ (Science Series), 48(4), 2016.

[13] M. Ali, & A. I. Wagan, “Sentiment summerization and analysis of Sindhi text,” Int. J. Adv. Comput. Sci. Appl, 8(10), 296-300, 2017.

[14] A. Z. Syed, M. Aslam, & A. M. Martinez-Enriquez, “Associating targets with SentiUnits: a step forward in sentiment analysis of Urdu text,” Artificial intelligence review, 41(4), 535-561, 2014.

Strategic Planning towards Automation of Fiber To The Home (FTTH) Considering Optic Access Network (OAN) Model

Abid Naeem[1], Shahryar Shafique[2]
Department of Electrical Engineering
Iqra National University, Peshawar, Pakistan

Sheeraz Ahmed[4]
Department of Computer Science
Iqra National University, Peshawar, Pakistan

Zahid Wadud[3]
Department of Computer System Engineering
University of Engineering and Technology
Peshawar, Pakistan

Nadeem Safwan[5]
Department of Management Sciences
Iqra National University
Peshawar, Pakistan

Zeeshan Najam[6]
Department of Electrical Engineering
Ultimate Consultancy, Peshawar, Pakistan

***Abstract*—With the intention to meet the increasing demand of future higher bandwidth applications, fiber based Gigabit Passive Optical Network (GPON) access is considered best resolution to deliver triple play services (voice, data, video). Hence, it becomes obligatory to migrate from traditional copper-based network to fiber-based. Due to rapid technological evolution, tough competition and budget limitation the service providers are struggling to provide a cost effective solution to minimize their operational cost with extra ordinary customer satisfaction. One of the factors that increase the cost of overall Fiber To The Home (FTTH) network is the unplanned deployment resulting in utilization of extra components and resources. Hence, it is imperative to determine a suitable technique, which helps to reduce planning process, required time and deployment cost through optimization. Automation based planning is one of the possible ways to automate the network design at probable lowest cost. In this research, a planning technique for migration from copper to fiber access network with a manageable and optimized Passive Optic Network (PON – FTTx) infrastructure is presented identifying a cost-effective strategy for developing countries.**

Keywords—Fiber To The Home; Passive Optical Networks; GPON; triple play; cost effective; customer satisfaction

I. Introduction

In order to provides a triple-play (voice, data, video) services by high speed collaborating apps, like games that runs online, several telecommunication organizations are considering Fiber based accessed networks as the key resolution. There are two different ways of delivering fiber networks to customer premises, namely Point-to-Point (P2P) and Point to Multi-Point (P2MP). On one hand, P2P [1] fiber networks use a specific fiber constituent to link specific customer sites all the way to exchange. It allows very high bandwidth services (compared with P2MP) to be delivered to businesses or high rise buildings even over a long distance. However, when the number of P2P connections is very high, the installation and maintenance costs can be prohibitively expensive.

On the other hand, the P2MP network based on GPON technology can provide an attractive solution to reduce the overall cost. With the P2MP GPONs, there are no electronic components between an exchange and customer premises. Only optical splitters are used to connect Optical Line Terminal (OLT) equipment at an exchange to a group of premises sharing the same feeder fiber. An Optical Network Unit (ONU) will then be used to convert the optical signal into an electronic signal at the customer's premises.

GPON has a downstream capacity of 2.488 Gb/s and an upstream capacity of 1.244 Gbp/s that is shared among users. GPONs are generally considered to be a more cost effective way of delivering FTTH services with minimum number of fibers and electronics required.

According to the market research division of Light Reading [2], the number of households with fiber-optic network connections was expected to grow by more than 32% worldwide in 2009 and will continue to grow at rates close to 45% a year through 2021. The number of fiber-connected households will reach nearly 130 million globally by that time.

Though, to form a cost effective GPON/FTTHN needs consideration of different factors, such as, locations of splitters, cable assignment of customer sites to splitters and provision of spare capacity for future growth. In addition, all the planning restraints like extreme allowable splitters capability and the extreme distance in OLTs, ONUs, splitters satisfied must.

To plan a GPON/FTTH network manually in a new area, an organizer is usually given a background plan that assisted by exchange. Whole region is subdivided in to small zones that are settled in multiple phases. With given rules of planning and sites of locations, organizer typically positions the optical Splitters (SPs) somewhere in the center of the planned area. Cable Distribution points (CD)s will then be positioned around the SP afterwards. Once the locations of SPs and CDs are specified, the planner will assign cables from each customer premise to a CD and from a CD to a SP based on the shortest distance.

This manual design process is very time consuming. Very often, due to the tight time schedule, when the design proposed by the planner satisfies all the distance and capacity constraints, he/she will submit the design without incorporating much cost optimization or considering the distribution of spare capacity.

A. Components of GPON FTTH Network

A Passive Optical Network (PON) is capable of having P2M (point to multipoint) network with passive components like optical splitter or coupler along the transmission section. It uses active components only at CO and at customer premises. It uses WDM to mix up video signals with the data and voice from OLT. Fig. 1 shows the basic FTTH Network. [3].

1) Optical Line Terminal (OLT): It is the most important part of the network, where the electrical signal from the service provider's equipment are converted into optical signals and given to the feeder network. The mode of transmission from ONT is broadcasting [4] from where it sends GEM frames through the GEM port with GEM port IDs It is capable of having Multi-service chassis for FTTx deployments, Supports a variety of service types, Non-blocking architecture with & Routing within distributed architecture, scalability and line rate performance, Full electrical and optical redundancy Outstanding scalability and line rate performance, Real-time network traffic monitoring and analysis.V8240 GPON OLT is used. Specifications are given in Table I.

2) Optical Network Terminal (ONT): It is an active component used at customer premises which converts optical to electrical signals. ONU/ONT represents the ingle customer where they will get the triple play application. H640 series GPON ONT are used. It is capable of having carrier class VoIP telephony supporting both MGCP and SIP protocols, Flexible VLAN tagging support, QoS for traffic prioritization and bandwidth management, IGMP support for IPTV applications. Its specifications are given in Table II.

3) Splitter: Splitters are used to physically split the fiber to number of fibers; to couple same or different information's to N users. MxN planar splitters are used which is based on planar light wave circuit (PLC) technology and high precision alignment. MxN splitters can split or combine light from one or two fibers into N outgoing fibers uniformly over a wide spectral range with ultra-low insertion loss and low polarization dependent loss. With up to 64 output ports, these splitters are ideal for high density split applications like Fiber To The Home (FTTH) networks, FTTx Deployments Optical CATV Networks, CWDM and DWDM Systems, Passive Optical Networks, Fiber Communication Systems Telecom, LANs. It has the features like Low Insertion Loss, Ultra broadband performance (1260 –1630nm), Low PDL and PMD, Stable towards thermal variations, Superior port to port uniformity. A splitter type is shown in Fig. 2.

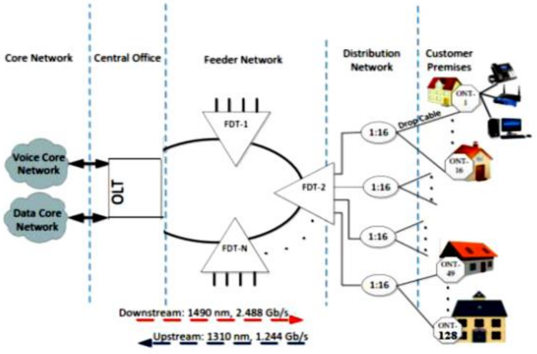


Fig. 1. FTTH Network [3].

TABLE. I. V8240 GPON OLT SPECIFICATIONS

Flash Memory	72 MB
SDRAM	1 GB
Dimensions (W x H x D)	17.1 x 12.2 x 11.2 in (434 x 310 x 285 mm)
Switching Capacity	296Gbps
Power Voltage AC type	100-240VAC, 50/60Hz
DC type	-48/60VDC
Operating Temp	32 to 122°F (0 to 50°C)
SIU (Subscriber Interface Unit)	10 slots
NIU (Network Interface Unit)	2 slots
SFU (Switching Fabric Unit)	2 Slots

TABLE. II. H640 SERIES GPON ONT

Service Interface	4 10/100Base-TX ports (RJ45) 2 POTS ports (RJ11) 1 RF video port (F-connector)
Uplink Interface	1 GPON port (SC/APC type)
Operating Temp	32 to 104°F (0 to 40°C)
Storage Temp	-4 to 140°F (-20 to 60°C)
Input	100-240VAC
Dimensions (W x H xD)	• Excluding bracket:10.24 x 2.05 x 7.87 in (260 x 52 x 200 mm) • Including bracket, wall mounting:10.51 x 2.60 x 7.87 in (267 x 66 x 200 mm) • Excluding bracket, desktop mounting: 10.24 x 2.80 x 7.87 in (260 x 71 x 200 mm)

Fig. 2. PLC Splitter with Ribbon Fiber.

II. Business Model of FTTH Planning

A. Business Model of FTTH Planning

With the intention to meet the increasing demand of future higher bandwidth applications, the fiber based access is considered to be a best resolution to offer triple play services. It is therefore preferred with great need to migrate from traditional capper based network to fiber based access. A business Model of new FTTH network deployment is illustrated in Fig. 3 which consists of some dependent and independent variable.

In Pakistan the broadband growth in wireline is very slow which is very much obvious. In [7], the slow growth is due to some factors that need improvement; these factors include.

- Low literacy rate
- low (level of) consumer awareness
- No coverage of Broadband services
- Traffic reduction in broadband services low computer penetration
- Cost of service (tariff)
- History of market and national regulation

1) Business output: The output of new FTTH Deployment translates into different benefits [20].

- End User/Customer benefits

High bandwidth is the main selling product of FTTH network; it provides the highest available bandwidth in both directions (downstream and upstream). A FTTH user can download data over 10 times quicker than ADSL user. Transfer rate of different content over various types of networks are shown in Table III.

Speed of ADSL over Copper network is inversely proportional to distance from customer end to telephone exchange, while in FTTH network distance does not affect speed. In DSL network signal to noise ratio (SNR), interference and crosstalk during operation also reduces the throughput. Customer satisfaction ratio in FTTH network is above 85%, higher customer satisfaction has a tendency to enhanced customer retention and reduce churn.

- Service provider benefits

✓ The lifespan of Fiber cable is very long as it is more than 30 years therefore FTTH is known as a “future-proof technology”. Fiber cable made of simply plastic and glass, which reduce his lifespan extremely slowly. The fiber cable has almost unlimited capacity and expansion in bandwidth needs only changes to the hardware at the ends of the link.

✓ Operational cost (OPEX) of FTTH networks is very low as compare to existing copper networks. It consumes 20 times less power than other. The operational and maintenance cost can be minimizing by automation control. Maintenance costs can also reduce because there is no active device in the field to maintain, and optical components have better reliability.

✓ Customer satisfaction will reduce the churn value and increase the customer, which also reduce operational cost. To keep the existing customer is so easy as compare to enlist new one. To maintain an existing customer is so easier than to register a fresh customer.

- Community benefits

FTTH enable Communities can get a lot of benefits with a wider range of internet services. Few examples of possible benefits with FTTH networks are as under:

➢ Financial boost with global competition.

➢ Attraction for new businesses.

➢ Provisioning of state of the art services in Education and health sector.

➢ Improving overall quality of life in a community by increasing the opportunities for communication.

➢ Controlling of Road traffic blocking/problem.

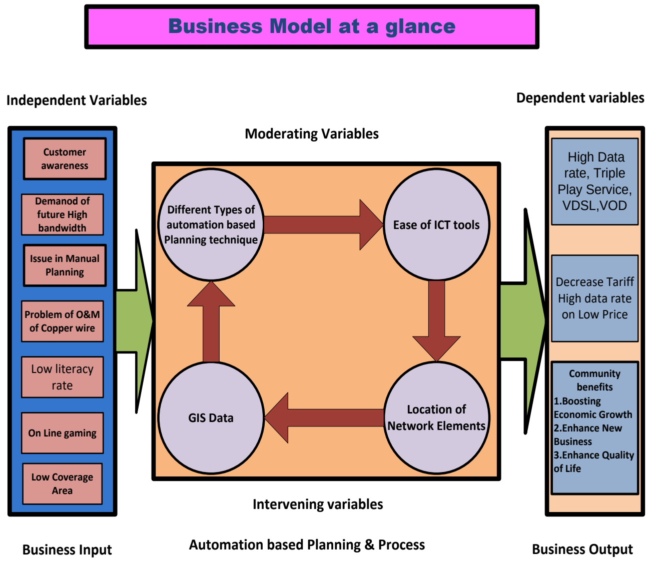

Business Model at a glance
Independent Variables
Dependent variables
Moderating Variables
Customer awareness
Demanod of future High bandwidth
Issue in Manual Planning
Problem of O&M of Copper wire
Low literacy rate
On Line gaming
Low Coverage Area
Different Types of automation based Planning technique
Ease of ICT tools
Location of Network Elements
GIS Data
High Data rate, Triple Play Service, VDSL,VOD
Decrease Tariff High data rate on Low Price
Community benefits 1.Boosting Economic Growth 2.Enhance New Business 3.Enhance Quality of Life
Intervening variables
Business Input
Automation based Planning & Process
Business Output

Fig. 3. Business Model of FTT H Optimization.

TABLE. III. Download/Upload Transfer Rate [20]

Data	FTTH		CATV		DSL	
	Down	Up	Down	Up	Down	Up
Pictures up to 1 GB	1 M 23 S	1 M 23 S	2 M 46 S	13 M 52 S	19 M	2 Hr 32 M
Standard video Up to 5 GB	6 M 31 S	6 M 31 S	13 M 2 S	1 Hr 5 M	1 Hr 29 M	11 Hr 29 M
HD Quality Video up to 25 GB	24 M 40 S	24 M 40 S	1 H 9 M	5 Hr 47 M	7 Hr 55 M	- - - - -

2) Automation based planning and process: To grip complication of FTTHN, an automation scheme is settled by keeping in mind, common optimization framework, which shows in Fig. 4 with different Phases.

a) Input Phase: In this phase various sources can be used to retrieve the data like:

- Geo-graphical Information Systems (GIS)
- Manually formed files

GIS data contains the setup of access network. Usually it used a geographical database with a three-dimensional data structure. To get the quick recovery of information, it associates a wide range of geographic items with a rich set of attributes.

Manually created files are the second source of input data phase. A Map of an exchange or area is typically used by planner. The area further divided into different regions. The planners usually choose a central point of the region for installation of Main Distribution Box (MDB) which consists of various splitters. Moreover, they allocate cables from end user to splitters through distribution Cabinet (DC).

b) Input Analyzer Phase: As the data gathered by different sources are in different format such as DXF, Esri Shape, so this stage is utilized to filter the required data from Input source. The data is first transformed into some matrices which consist of required information to perform optimization procedures efficiently.

c) Business Logic Phase: The comprehensive cost model with engineering rules summarizes in this stage. To reduce the cost of network design problems the engineering rules used for constrains. Different costs of network design like HR cost, ducting costs, cabling and network equipment cost includes in this cost model.

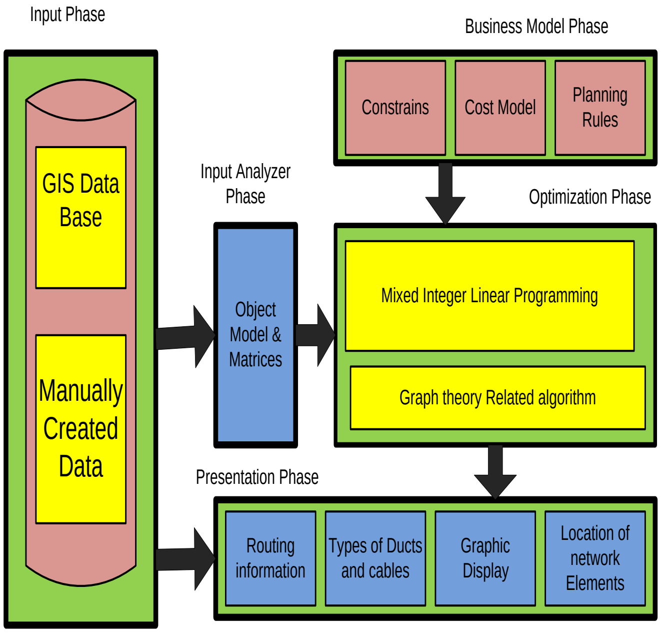


Fig. 4. Automation based Planning Process.

This Phase is especially problem particular as well as frequently modified to meet necessity. Moreover, the verification of this phase can be obtained by manual solution. Calculation of Complete expenditure and design limitation will validate in this phase. The business designer can validate their business model with the proposed Model. After the completion of business model, it can be tested on various optimization methods.

d) Network Optimization Phase: This phase consists of optimization approach based on Mixed Integer Linear Programming (MILP). It is a distinct variation of the linear programming. In MILP few variables have integer values. In MILP, our problem is molded by binary variables [8,11,15].

Prior to executing the MILP-based design tool, we assumed that below given information's are delivered:

- Locations of customer sites.
- Location of one exchange E.
- Possible sites of cable distribution (CD).
- Number of occupancies for all premises that determines number of required PON links.
- Civil layer network which specifies connectivity in different network components.
- Requirements of spare capacity needed to accommodate future network growth.

III. Motivation

In this paper, a network design tool for the GPON/FTTH network is proposed to automate the planning process. Thus, given the locations of customer plots, possible locations of CDs and SPs, the tool decides the optimal or near optimal locations of CDs and SPs taking into account the future growth and spare capacity distribution. In addition, cables are assigned from each plot to a CD and to the selected SP. The solution is optimal in the sense that it leads to the minimum cost of deployment which includes the number of network elements required, the total cabling distance and the installation costs.

By using an automated planning assistant tool, the planner can:

- Minimize network capital expenditure, i.e., installation and materials costs.
- Quickly achieve the network design of a given area.
- Compare what-if scenarios to meet changes in planning requirements.
- Rapidly re-cost networks for contract control and installation.
- Produce cable design and bill-of-materials automatically.
- Specify pre-determined locations for network elements prior to performing the network optimization.

The benefits of automating, network design includes reducing installation and material expenses, decreasing time to make a design from hours to minutes, speedily re-costing networks for different laboring or equipment costs and making network design as well as automatically making bill-of-materials too.

IV. Literature Review

A. Literature Regarding FTTH Planning

The issue of FTTH network planning has been identified in literature. In this section two key methodologies are frequently used. The first one is that meta-heuristic techniques such as genetic algorithms (GA) and exact techniques like use of Mixed Integer Linear programming (MILP) both methodologies have their pros and cons. The MILP can obtain global optimum solution as well as can outclass meta-heuristics for a medium-sized network [5].

According to the authors of [6] presented a solution which employs heuristics, falling in 2nd type of methodologies. Their research emphasized on networks that are multistage splitting. As they have complex problems, the researchers adopted heuristics for reducing computational time. The cluster based and random locations of splitters were taken and performed comparisons. The effect of different localization of splitters on lessening in CAPEX was deliberated. In [7] the authors proposed a model of a real life network. Real data of building sites and streets was mined from a Geographical Information System (GIS) by use of an open source map known as Open Street Map (OSM). This model was adapted to single level passive optical network (PON). These remote nodes are located for serving the customers. For more decrease costs sustained, cables channels shared by different routes of cable.

In [8], the networks that based on MILP, a design tool for GPON/FTTH networks are proposed. This tool automates the planning process of networks. By providing positions of customers and probable positions of SPs and DC’s, it adopts optimum positions of DCs and SPs as well as assignment of cables in network elements and customers. Thus minimizes entire network development cost. The researchers took in to consideration future progression and introduced a technique for planning large networks.

According to [5], an approach is presented that is based on meta-heuristic. This approach used Ant Colony Optimization (ACO). This algorithm achieves cable assignment on a multilevel network which emphasizing on cost minimizing. This algorithm allocates customers to DCs as well as DCs to splitters simultaneously. Authors of [9] proposed a mixture of heuristics and mathematical programming for minimizing deployment cost of a GPON. The methodology of this research work was same to Simulated Annealing by which assignment of cables as well as positions of splitters recursively reallocated till a well cost was found. In [10], a scheme is presented that is based on previous class of methodologies. This scheme is a cross layer optimization. According to this scheme the researchers targets Greenfield network deployment and pursued to produce physical architecture of Wave Length Division Multiplexing (WDM) PON networks. It attains nominal cost for network deployment. This algorithm initially discovers optimum number of clusters for customers and then continues to assign every ONU to a cluster.

Researchers of [10] prolonged their research work in [11] by producing multiple WDM PON networks concurrently. It is attained by searching finest cost effective WDM PON by splitting an area in to sub sections in which every sub region covered by a specific PON. The authors of [12] proposed a tool for semi-automated network planning. It defines a suboptimal route distribution for deployment cost. It utilizes existing cable channels. After clustering customers, the authors of this research utilize GA for route deployment process. The outcomes are compared to network designs attained by manual process which depicted that in most cases this tool generates an inexpensive network.

Though, none of research work considered different types of network elements selection, such as research work in paper [8] emphasized on decisive optimum locations of network elements and cable assigning. They assumed a specific type of network element. Another heuristic approach was recently introduced in [13]. The approach is based on clustering and a Tabu search and has been enhanced with mechanisms handling resiliency issues as presented in [14]. The approach we use in this paper, originally presented in [15], is based on beam search [16]. This approach has been enhanced with mechanisms handling uncertainty issues following those used in [17] and upgraded with the MIP polishing mechanisms of [18]. This last idea to mix MIP methodology with heuristics proved to be very efficient and also has been recently used in [19]. The methodology used in [20, 21] is MIP facilitated by the use of valid inequalities and various algorithmic enhancements. Another recent work by Orange Labs is [22]. It is similar to our research in majority of assumptions and the methodology used. However, it covers only the last access part of FTTH network; thus, the authors do not consider splitting and OLT costs. Still, the detailed view on the fiber splicing problem presented in [22] is definitely worth noting.

B. Literature Regarding GPON Technology

There are two main streams of research focusing on GPON technology: Dynamic bandwidth allocation (DBA) algorithms among OLT and ONUs, which can be found in [23, 24, 25] and optimal network design of the physical layer for GPON deployment. The latter is the one considered in this research and discussed in detail. Using the classical operational research approach, the planning problems can be assumed of as a ordered concentrator network problems. In context of GPONs, the concentrator acts as a splitter to connect several ONUs to an OLT in a star topology. When several splitters are connected to the OLT at different locations, it becomes a double-star topology. Details of the classical access network design approach can be found in [26, 27, 28, 29].

In [30], the authors developed an optimization solution to perform multi-hierarchy PON planning. In their case, upper Optical Branching Devices (OBDs) and lower OBDs were introduced. The upper OBDs were used to connect between OLT and lower OBDs whilst lower OBDs were used to connect between ONUs and upper OBDs.

The locations of OBDs were calculated based on the Max-Min Distance Cluster (MMDC) algorithm which can be found in [10]. Regarding the optimization framework, authors in [31] introduced a segmental framework which primarily emphasized on metaheuristic optimizing approaches. Practical sample from motorized domain comprised for validate how overall problem could be break down in to sub-jobs and controlled through propose framework.

According to [32], the author works on collective deployment of access network architectures such as Fiber To The Node (FTTN), Fiber To The Micro Node (FTTN) and Fiber To The Premise (FTTP) to decrease span of loops of coppers through use of DSL access multiplexer in external cabinet and field micro node which are nearer to subscribers. Several classes of services and subscribers per class per point of demand are considered. The MILP model has been proposed together with a tabular search base process for improving computational time needed for finding best resolution.

V. Automation Leading to Optimization

MIP approach is usually using in different means in FTTH networks designing. According to our research work, we adopt MIP for the improvement of results that are returned by empirical algorithms of optimization framework that are introduced in [33]. Framework used: locations of demand, the available setup, with labor and equipment as well as technology restraints. It returns a complete network planning comprising: the topology of network, OLT, splices, splitters, OLT cards, cables, splices closures as well. By using stated aspects in modeling of a problem would leads to an incredible of variables and restraints making acquired model unsolvable by overall up-to-date MIP solvers. Hence, issue to be shortened for making it amenable. Key supposition was for using this approach is for improved acquired solutions; simplifying of model not be depends on neglecting ostensibly least significant factors, such as like splicing. While apparently additional significant factors, such as the OLT sites, to be static and detached from model. In this research, optimization of capital expenses essential for the placement of FTTH-OAN is addressed which contains one or more OLTs at the CO location and group of point of access that are located in the / or nearby to the CP locations. Networks gratify loads of all point of access take in to account permissible power budget of optical links and split scenarios.

A. Prerequisites Data

The problem that is denoted by Δ needs the given below input data:

- Passive and actives Equipment’s record.
- Infrastructure networks Topology.
- Every distribution and access node infrastructure paths that are selected.
- All access nodes Signal demand.
- The Infrastructure sites that are decided for the installing active, passive equipment.

B. Decision Variables

For optimization, the decision variables are followed:

- Cabinet types utilization (given nodes)
- All types of Splitter with locations utilization
- All types of cables utilization (given topology)
- Splice closure and Splicing locations utilization
- OLT types splice closures (locations are given)

C. Problem Statement

Keeping in mind the structure as well as complexities of complete problem, in our research description of its formulation is split in to four (4) problems that are depicted in Fig. 5 as a square. We focused; these partial problems interlinked. The semantics of each variable that link to specific couples of partial problem is drawn by the shape of oval.

First part of problem is the *bundle layer dimensioning denoted by* Δ^{Bl}. Its purpose is that at assessing number with all types of splitters that are installed and at selection number of OLT cards to be install at Central Office (CO) nodes as well. Delivered resolution promises that all accessing nodes, irrespective of its distance to the CO, delivered with requisite optical signal of appropriate power.

The 2nd part of problem is *cables which is denoted by* Δ^{I}_{cab}. It defines number of cables of all types that are installed at every infrastructure segment. Installed the cables are to provide fibers in numbers sustaining requests of bundle layer dimension problem.

Splices are the 3rd part of problem. *It is denoted by* Δ^{I}_{Spl} which calculates number of optical closures and splices of all types which are essentially be install in every infrastructure node for supporting of solution that assessed in first two problems.

The fourth part of problem is *site dimensioning. It is denoted by* Δ^{I}_{Sit} that’s goals is selecting a site type and number with all types of hardware’s cabinet that are install in each infrastructure location.

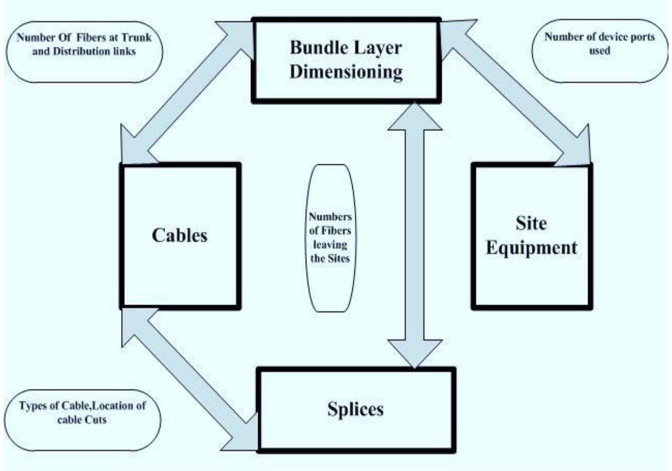

Number Of Fibers at Trunk and Distribution links
Bundle Layer Dimensioning
Number of device ports used
Cables
Numbers of Fibers leaving the Sites
Site Equipment
Types of Cable,Location of cable Cuts
Splices

Fig. 5. Block Diagram of Proposed Structure for OLT.

D. *Problem Objectives*

Key objective of $\mathcal{P}$ problem is to decrease the total cost is expressed by all partial of $\mathcal{P}$:

$$\Delta = \Delta^{Bl} + \Delta^{I}_{cab} + \Delta^{I}_{Spl} + \Delta^{I}_{Sit} \quad (1)$$

In equation (1) ∂ depicting overall cost. The symbols Δ^{Bl}, Δ^{I}_{cab}, Δ^{I}_{Spl}, and Δ^{I}_{Sit} are depicting cost of individual parts of $\mathcal{P}$ that are, bundle layer, cables, splices and site dimensioning respectively each problem.

E. *Description Layout of Problem*

To simplifying the overall layout, we are split problem in to two portions. The first part is the model part that is described in Section 6.1 which presents architectures with important notions that expresses overall organization of FTTH-OAN. The second part is equipment catalogue that is described in Section 6.2. This part describes a catalogue sets which identifies all equipment types. For every type, a set of critical parameters such as capacity, cost.

VI. Model of FTTH- OAN

This section presents a comprehensive model of all resources of FTTH-OANs. The basic FTTH-OAN model shown is in Fig. 6. It is appropriate for the requisites of formulation of the optimization problem. This model shown in Fig. 7 by dividing it into two fragments which are:

- Network
- Equipment

The network Fragment consist of equipment's permissible for installation in infrastructure nodes and segments such as cables, segments preparation types, hardware cabinets, OLT devices and cards, cable closures, splitters, etc.

A. *Network Fragments*

According to telecommunication modeling rules, like [34], we splitting networks fragment in to stack of layers. In these layers every pair neighboring layers, upper layer which is client, gains the benefits of all resource providing by the lowering layer which is server. We differentiate the four (4) layers starting from *signal*, *bundle*, *infrastructure and cable layer*.

Descriptions start from signal layering model. It contains an essential base for our optimizing methodology. It classifies components of signal nodes and links that are needed for delivering signal networks links in ONT and OLT devices. Unluckily, signaling model differentiates each individual component. Its direct application in formulation of optimizing problem leads to unsatisfactorily great optimization problem instances. To overwhelm this issue, we presented a layer of aggregated bundle model. It takes collection of signal nodes and connections as a bundle instead of distinct ones. As there is intricate relation in layers hence we settled their descriptions in an order which is not follows layers' order.

1) Model of infrastructure layer: The infrastructure layer model is illustrated in Fig. 8. This layer characterizes network of channels that are connected through staves which can accommodate optical cables that supports multiple ODNs. Network is needs to be linked therefore at least one link in Central Office and all ONT must be exist node.

Topologies of the infrastructure layers are molded by an undirected graph $G^{IL} = (\mathcal{N}^{IL}, \mathcal{L}^{IL})$ with group of *infrastructure nodes denoted by* $\mathcal{N}^{IL}$ and group of undirected *infrastructure link denoted by* $\mathcal{L}^{IL} \subseteq \mathcal{N}^{IL} \times \mathcal{N}^{IL}$. Infrastructure node that is denoted by n which is belongs to $\mathcal{N}^{IL}$, signifies a position, like a staves or a pole, in which optical wires sacked. The infrastructure link that is denoted by $l \in \mathcal{L}^{IL}$ signifies a place-holder, such as channel in pair of infrastructure node. All links can be accommodating different optical wires.

The set $\mathcal{N}^{ILS} \subseteq \mathcal{N}^{IL}$ of infrastructure sites is differentiated in such a way that nodes furnished to hold either signal splitters or active devices. These locations, depends on theirs positions in service area of networks, are apportioned in to central office sites $\mathcal{N}^{ILSO} \equiv \{n^{ilso}\}$, distribution point sites $\mathcal{N}^{ILSD}$ which is specify by *DP*, point of access locations $\mathcal{N}^{ILSA}$ that is indicated by *AP* while the customer premise sites $\mathcal{N}^{ILSP}$ is depicted through *CP*.

Finally, a set $\mathcal{P}^{IL} \subseteq \mathcal{L}^{IL}$ is described that is the all *infrastructure links* with a suitable subset $\mathcal{P}^{ILS} \subset \mathcal{P}^{IL}$ of *infrastructure trails*; through supposition, for a couple.

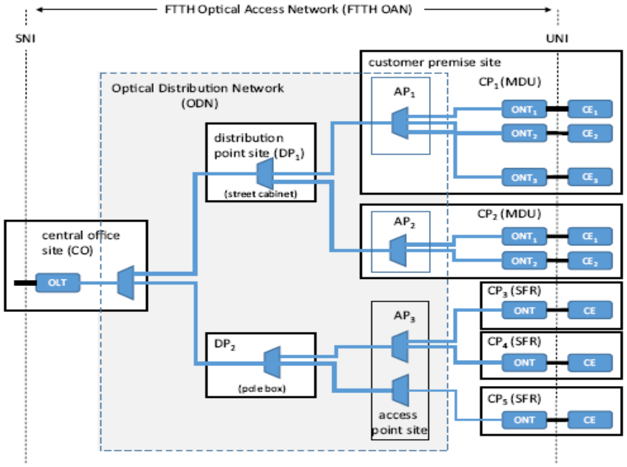


Fig. 6. FTTH –OAN Basic Network [33.]

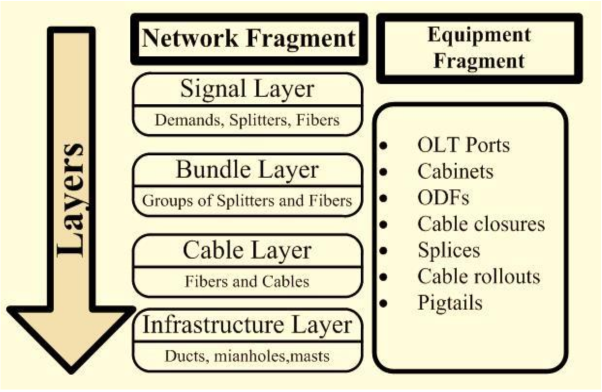


Fig. 7. Model Fragments [34].

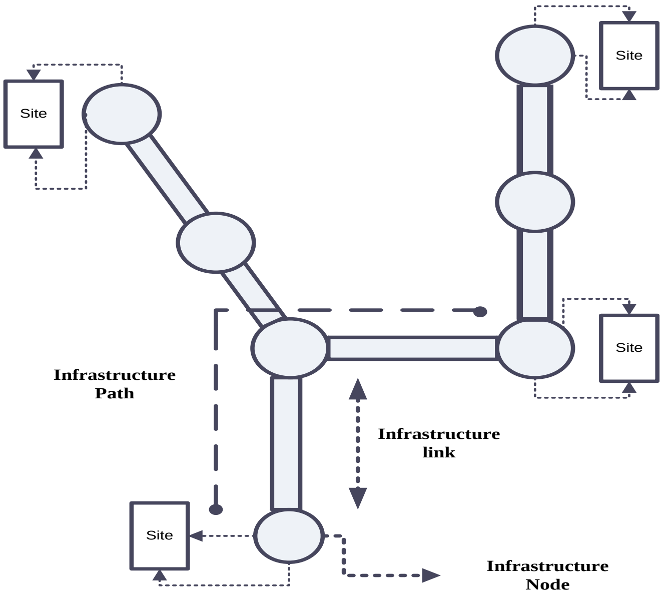


Fig. 8. Infrastructure Layer Topology.

$n_1, n_2 \in \mathcal{N}^{ILS} \times \mathcal{N}^{ILS}$ of infrastructure locations. There is selected maximum one infrastructure link that denoted by $p \in \mathcal{P}^{IL} : p \in \mathcal{P}^{ILS}$ denoted to as infrastructure trail of this pair.

2) *Model of cable layer:* Fig. 9 shows the fiber and cable layer Model. The Every fiber link contains a series of directed segments of *fiber* that are linked permanently or temporarily at the infrastructure node as optical splices or ODFs.

Lastly, optical cables segments are incessant segment of optical cables that is installed in infrastructures route containing single or multiple infrastructure segments.

3) *Model of signal layer:* Signaling layer gives a group of descendent signal network links in OLT devices ports, ONT terminals and group of parallel ascendant network links which links ONTs to OLTs for a single FTTH-OAN. These links are providing through the passive ODNs containing a set of signal splitters connected by dual directional fiber optic routes. The Signal nodes with links model is shown in Fig. 10. Signal layer topology is signified by a graph. This graph is denoted by G^{SL} which is equal to signal node and group of directed signal connections.

Group of signal nodes that are denoted by $\mathcal{N}^{SL}$ and group of directed signal connections denoted by $\mathcal{L}^{SL}$. The signal nodes denote signal transporting functions are done through active as well as passive devices. The signal connections signify in turn distinct fiber optic routes interrelating couples of signal nodes. The $\mathcal{N}^{SL}$ is divided in to set head end nodes that is denoted by $\mathcal{N}^{SLH} \equiv \{n^{slh}\}$, set of access nodes denoted by $\mathcal{N}^{SLA}$, and set of signal distribution points denoted by $\mathcal{N}^{SLD}$. The set $\mathcal{N}^{SLD}$ is also divided into sets of distribution, head end and access signal distribution points that are denoted by $\mathcal{N}^{SLDH}$, $\mathcal{N}^{SLDD}$ and $\mathcal{N}^{SLDA}$ repectively. This division reveals level that engaged by specific splitters within network links. The $\mathcal{N}^{SLA}$ are real source of demanding signal demand of specific access node n *which belongs to* $\mathcal{N}^{SLA}$. The $h^{sl}(n)$ is the number of network links that requires.

A signal node n belongs to $\mathcal{N}^{SL}$ depends on its class either $\mathcal{N}^{SLH}$, $\mathcal{N}^{SLDH}$, $\mathcal{N}^{SLDD}$, $\mathcal{N}^{SLDA}$, or $\mathcal{N}^{SLA}$ is defined a subset of infrastructure location types either $\mathcal{N}^{ILSO}$, $\mathcal{N}^{ILSD}$, $\mathcal{N}^{ILSA}$, or $\mathcal{N}^{ILSP}$ it can be installed in. For viable allocations, we used a function si_site(n) : $\mathcal{N}^{SL} \rightarrow \mathcal{N}^{ILSL}$, that for signal node n ∈ $\mathcal{N}^{SL}$, defines its presenting infrastructure location sl ∈ $\mathcal{N}^{ILS}$. By assumption following assignments are viable:

Single head end signal node n^{slh} that belongs to $\mathcal{N}^{SLH}$ essentially be allotted to single CO site n^{ilso} that belongs to $\mathcal{N}^{ILSO}$, where *si_site* $(n^{slh}) \equiv n^{ilso}$,

Every distribution signal distribution point*n belongs to* $\mathcal{N}^{SLDD}$, can be allotted to CO site n^{ilso} or any DP site *s that belongs to* $\mathcal{N}^{ILSD}$ *that is* $\forall_{n\in} \mathcal{N}^{SLDD}$, *si_site(n)*∈ $\{n^{ilso}\} \cup \mathcal{N}^{ILSD}$,

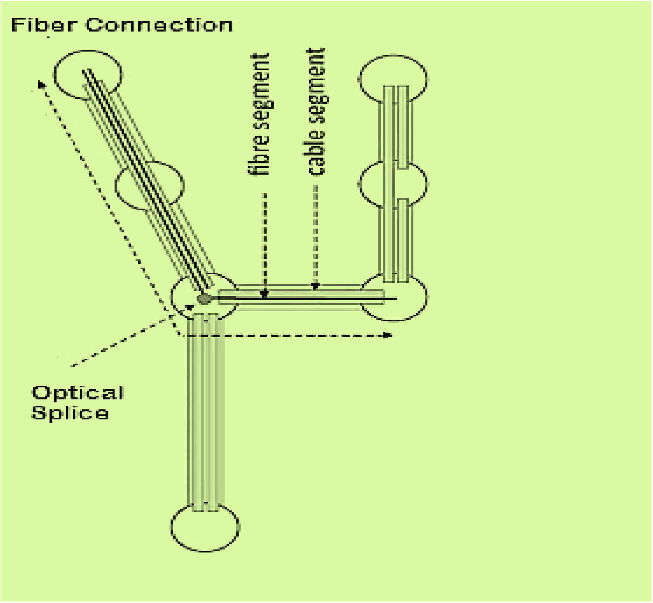


Fig. 9. Fibers and Cable Segment.

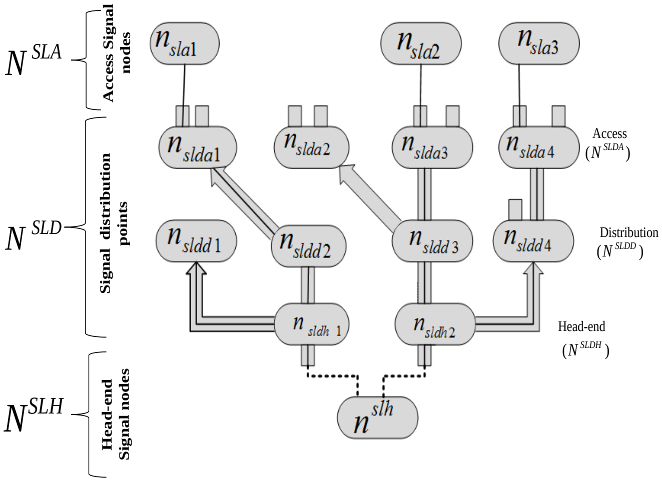


Fig. 10. Signal Nodes with Links and Connections.

Each head end signal distribution point n belongs to $\mathcal{N}^{SLDH}$essentially be allotted to single CO site n^{ilso} , that is $\forall_{n\in}\mathcal{N}^{SLDH}$, *si_site(n)* ≡ n^{ilso} ,

Every access signal distribution point $n\ that\ is\ belongs\ to\ \mathcal{N}^{SLDA}$ can be allotted to any type of an infrastructure site however CP, that is, *si_site(n)*∈ $\mathcal{N}^{ILSO} \cup \mathcal{N}^{ILSD} \cup \mathcal{N}^{ILSA}$,

Lastly, every access signal node *n* that belongs to $\mathcal{N}^{SLA}$ essentially be allotted to any of CP infrastructure locations *s* that is belongs to $\mathcal{N}^{ILSP}$

4) Model of bundle layer: The model of signaling layer recognizes each distinct component must deliver signal network links in OLT and ONT. In our work we presented aggregate model, by the name of *bundle layer.* It deliberates groups of signal nodes and links as a substitute of individual ones.

This model contains directed graph that is denoted by G^{BL} which is equal to $\mathcal{N}^{BL}$,, $\mathcal{L}^{BL}$. *The* bundle nodes denoted by $\mathcal{N}^{BL}$ while $\mathcal{L}^{BL}$ is group of bundle connections. G^{BL} Organizes a contraction of a graph.

G^{SL} = ($\mathcal{N}^{SL}$,$\mathcal{L}^{SL}$) of signaling layer, each bundle node $n_{bl} \in \mathcal{N}^{Bl}$ signifies subset $\mathcal{N}_{bl}^{sl} \subseteq \mathcal{N}^{sl}$ of signaling nodes. All bundle layer links $l_{bl} \in \mathcal{L}^{BL} : l_{bl} \in \delta(n_{bl})$ occurrence of bundle node n_{bl} signifies in turn group.

$\mathcal{L}_{bl}^{sl} = \{ l_s \in \mathcal{L}^{sl} : l_{sl} \in \delta(\mathcal{N}_{bl}^{sl}) \}$ of each signaling link occurrence to this selected subset of signaling nodes. For simplify mapping in signaling and bundle layers model, we present functions *bs_nmap(* n_{bl}): $\mathcal{N}^{BL} \mapsto 2|\mathcal{N}^{SL}|$. This function expresses subset of signaling nodes $\mathcal{N}^{SL}$ amassed to bundles node $n_{bl} \in \mathcal{N}^{BL}$ and function *bs_lmap(*ln_b: $\mathcal{L}^{BL} \mapsto 2|\mathcal{L}^{SL}|$ which states subset of signaling links denoted by a bundle link l_b.

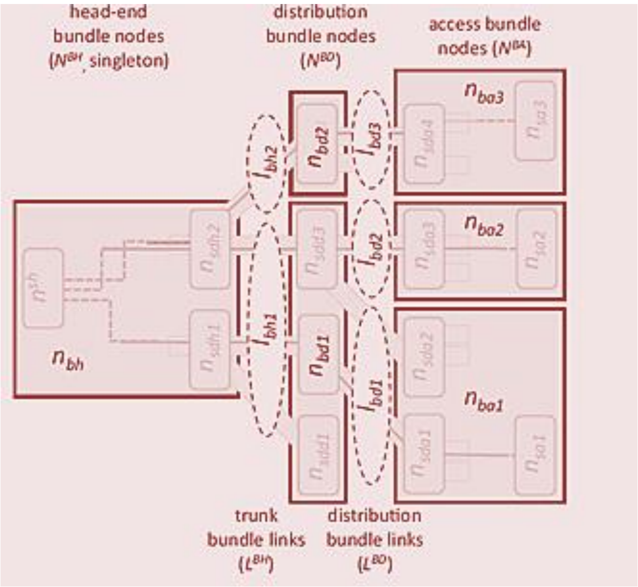


Fig. 11. Bundle Nodes and Bundle Links.

Following is categorization of bundle nodes:

a) Singleton set that represent by $\mathcal{N}^{BLH} \equiv \{n^{blh}\}$ which masses the head end signal node n^{slh} . Each head end signaling distribution point $n_{sldh} \in \mathcal{N}^{SLDH}$, which is placed in CO infrastructure location n^{ilso}. Mentioning to sample network single bundle head-node n^{blh} masses head end signal node n^{slh} and two head end signal distribution points n_{sldh1} and n_{sldh2} .

b) Set $\mathcal{N}^{BLD}$ of distribution bundle nodes. Every distribution bundle node$n_{bld} \in \mathcal{N}^{BLD}$, agrees to CO or a DP infrastructure location $n_{ils} \in \mathcal{N}^{ILSO} \cup \mathcal{N}^{ILSD}$ and groups each distribution signal distribution points $n_{sldd} \in \mathcal{N}^{SLDD}$: *si_site(*n_{sldd} *)* = n_{ils} , positioned in that location. In sample network, there are two distribution bundle nodes n_{bld1} , n_{bld2} that collective, respectively, a subset of distribution signaling distribution points $\{n_{sldd1}, n_{sldd2},\ nsn_{sldd3}\}$ and $\{n_{sldd4}\}$;– set $\mathcal{N}^{BLA}$ *o*f accessing bundle nodes. Every accessing bundle node $n_{bla} \in \mathcal{N}^{BLA}$, related to an infrastructure location $n_{ils} \in \mathcal{N}^{ILS}$, and aggregates every access signal distribution point $n_{slda} \in \mathcal{N}^{SLDA}$: *si_site(*n_{slda}*)* = n_{ils} as well as each accessing signaling node $n_{sla} \in \mathcal{N}^{SLA}$: *si_site(*n_{sla}*)* = n_{ils}*nis* positioned at this location. In sample network as shown in Fig. 11, in which three access bundle nodes represented byn_{bla1}, n_{bla2}, n_{bla3} respectively. We proposed *c(n)* : $\mathcal{N}^{BLD} \mapsto 2|\mathcal{N}^{BLA}|$ function that for every distribution bundle node $n \in \mathcal{N}^{BLD}$ distinguish subset $\{m \in \mathcal{N}^{BLA} : \exists l \in \mathcal{L}^{BLD}$, *b_a(l)* = $n \wedge$ *b_b(l)* = $m\}$ of bundle accessing nodes linked to that distribution node through a bundle connection; referred by *distribution cone* of n distributing nodes.

c) Bundle links that denoted by $\mathcal{L}^{BL}$ further divided into trunk bundle and distribution bundle links that are represented by $\mathcal{L}^{BLH}$ and $\mathcal{L}^{BLD}$ respectively. With support of bb_a(l) : $\mathcal{L}^{BL} \mapsto \mathcal{N}^{BL}$ and bb_b(l) : $\mathcal{L}^{BL} \mapsto \mathcal{N}^{BL}$ functions that recognize, correspondingly, start and end bundle nodes of a directed bundle link denoted by l ∈ $\mathcal{L}^{BL}$, these two can be properly define as, $\mathcal{L}^{BLH}$ = {l ∈ $\mathcal{L}^{BL}$: bb_a(l) ∈ $\mathcal{N}^{BLH}$, bb_b(l) ∈ $\mathcal{N}^{BLD}$} and $\mathcal{L}^{BLD}$ = {l ∈ $\mathcal{L}^{BL}$: bb_a(l) ∈ $\mathcal{N}^{BLD}$, bb_b(l) ∈ $\mathcal{N}^{BLA}$ } respectively,. Sets $\mathcal{L}^{BLH}$ and $\mathcal{L}^{BLD}$ organize the segregating of set $\mathcal{L}^{BL}$.

The actual demands for signals network connections are generates in accessing bundle nodes $\mathcal{N}^{BLA}$; demand h_n^B of every distinct accessing bundle node $n \in \mathcal{N}^{BLA}$ calculated by given below expression:

$$h^B(n) = \sum_{s\in\mathcal{N}^{SLA}:s\in bs\_nmap(n)} h^S(s),\ \ n \in \mathcal{N}^{BLA} \qquad (2)$$

5) Concluding remarks of network fragment: The direct bundle link *l* that is belongs to $\mathcal{L}^{BL}$, of bundle layer maintained through a group of similar fibers links in *si_site(bb_a(l))* and *si_si te(bb_b(l))* infrastructure locations. By supposition, each fiber link used identical infrastructure trail *bi_p(l)* ∈ $\mathcal{P}^{ILS}$ where function *bi_p(l)* : $\mathcal{L}^{BL} \mapsto \mathcal{P}^{ILS}$ expresses an infrastructure trail that is taken by each fiber link supportive bundle link.

Let directed trunk bundle connection from group$\mathcal{L}^{\mathrm{BLH}}$. Each fiber link which supports that link is known as trunk fiber link. It uses a trunk infrastructure trail which contains trunk fiber segments. All cable segments that have trunk fibers denoted as Trunk cable segment. Conferring to stated rules, we introduced group of trunk infrastructures trails representing by $\mathcal{P}^{\mathrm{ILSH}}$and group of distribution infrastructure trails that represented by$\mathcal{P}^{\mathrm{ILSD}}$. We signify group of trunk infrastructure segment and distribution infrastructure segment through, respectively, $\mathcal{L}^{ILH} \subseteq \mathcal{L}^{IL} and \ \mathcal{L}^{ILD} \subseteq \mathcal{L}^{IL}$; we focused on sets $\mathcal{L}^{ILH}$ *and* $\mathcal{L}^{ILD}$ that normally don't organize segregating of set $\mathcal{L}^{IL}$. We assumed that each trunk infrastructure trail from $\mathcal{P}^{ILSH}$ which use single trunk infrastructure segment from $\mathcal{L}^{ILH}$ traverses segment in identical direction. Therefore, trunk and distribution, infrastructure segment can be considered as directed. *ii_a(l)* : $\mathcal{L}^{IL} \mapsto \mathcal{N}^{IL}$ *and ii_b(l)* : $\mathcal{L}^{IL} \mapsto \mathcal{N}^{IL}$, functions expresses respectively, start and ending infrastructure nodes of infrastructure segments. The trunk segment in $\mathcal{L}^{ILH}$ creates a directed tree by root at CO locationn^{ilso}, whereas distribution segments in set $\mathcal{L}^{ILD}$ creates forest of directed tree, that all rooted at a location which hosts distribution bundle node from $\mathcal{N}^{BLD}$ set. Hence, for every trunk segment *l that belongs to* $\mathcal{L}^{ILH}$ there is just one or none predecessor trunk segment *ii_ah(l).* Likewise, for every distribution segment *k that belongs to* $\mathcal{L}^{ILD}$ there is just one or none predecessor distribution segment *ii_ad(k).*

B. Equipment Catalogue Portion

In this section with the help of catalog defining the physical equipment types are acceptable for installation at infrastructure nodes and sites. Every catalog set is denoted by δ with a lowercase upper index; it is also used for individual properties of an example of a particular type. Parameters common to every type, like cost or capacity, are denoted by Greek letters δ and η with appropriate upper indices. A brief list of catalogue sets is listed in Table IV. The detail description of this fragment is beyond the scope of this paper.

TABLE. IV. EQUIPMENT CATALOGUE SETS

Name	Catalogue
δ^{oa}	Optical cables
δ^{ob}	OLT cards
δ^{oc}	Cabinets
δ^{od}	Sites
δ^{oe}	Segment preparations
δ^{of}	OLT devices
δ^{og}	Fiber Splices
δ^{oh}	Optical Splitter
δ^{oi}	Splitter Combinations
δ^{oj}	Joint Closures

VII. CONCLUSION AND FUTURE WORK

The increasing demand for Broad band internet services requires adaption of novel fiber base technologies. To attract new customers, fixed access network operators have to substantially increase the speed and quality of internet services. This can only be achieved by bringing the fiber as close to the customer as possible. This requires extensive planning in term of cost, time and infrastructure. In this paper we have presented a model of automating FTTH planning considering OAN. Different features, planning phases and model fragments have been identified and discussed, both theoretically as well as mathematically. In our future work we will present formulation in term of optimization for FTTH cost effective deployment. Real world experiment will aid in formulation as well as validation.

REFERENCES

[1] J. Prat, Next-generation FTTH passive optical networks: research towards unlimited bandwidth access. Springer, 2008.[Online] Available: http://dx.doi.org/10.1007/978-1-4020-8470-6.

[2] PR Newswire,http://www.prnewswire.com/news-releases-62266647 .html,"April 2011.

[3] M.M.Al-Quzwini,―Design and Implementation of a Fiber To The Home FTTH Access Network based on GPON, in International Journal of Computer Applications, vol.92,no.6, April 2014.

[4] ITU-T G.984 Gigabit Passive Optical Network Specifications

[5] A. Chu, K. F. Poon, and A. Ouali, "Using Ant Colony Optimization to design GPON-FTTH networks with aggregating equipment," in 2013 IEEE Symposium on Computational Intelligence for Communication Systems and Networks (CIComms), April 2013, pp. 10–17.

[6] A. Eira, J. Pedro, and J. Pires, "Optimized Design of Multistage Passive Optical Networks," IEEE/OSA Journal of Optical Communications and Networking, vol. 4, no. 5, pp.

[7] O. Kipouridis, C. Machuca, A. Autenrieth, and K. Grobe, "Street aware infrastructure planning tool for Next Generation Optical Access networks," in 2012 16th International Conference on Optical Network Design and Modeling (ONDM), April 2012, pp. 1–6.

[8] K. F. Poon and A. Ouali, "A MILP based design tool for FTTH access networks with consideration of demand growth," in 2011 International Conference for Internet Technology and Secured Transactions (ICITST), Dec 2011, pp. 544–549.

[9] J. Li and G. Shen, "Cost Minimization Planning for Greenfield Passive Optical Networks," IEEE/OSA Journal of Optical Communications and Networking, vol. 1, no. 1, pp. 17–29, June 2009.

[10] R. Chowdhury and B. Jaumard, "A cross layer optimization scheme for WDM PON network design and dimensioning," in 2012 IEEE International Conference on Communications (ICC), June 2012, pp. 3110–3115.

[11] "A p-center optimization scheme for the design and dimensioning of a set of WDM PONs," in 2012 IEEE Global Communications Conference (GLOBECOM), Dec 2012, pp. 2977–2983.

[12] B. Lakic and M. Hajduczenia, "On optimized Passive Optical Network (PON) deployment," in Second International Conference on Access Networks Workshops, 2007. Access Nets '07, Aug 2007, pp. 1–8.

[13] G. V. Arévalo, R. C. Hincapié, and R. Gaudino, "Optimization of multiple PON deployment costs and comparison between GPON, XGPON, NGPON2 and UDWDM PON," Opt. Switching Netw., vol. 25, no. Supplement C, pp. 80–90, 2017.

[14] G. V. Arévalo and R. Gaudino, “A techno-economic network planning tool for PON deployment including protection strategies,” in 19th Int. Conf. on Transparent Optical Networks (ICTON), July 2017, pp. 1–4.

[15] R. Bisiani, “Beam search,” in Encyclopedia of Artificial Intelligence, Wiley, 1987, pp. 56–58.

[16] M. Żotkiewicz and M. Mycek, “Impact of demand uncertainty models on FTTH network design,” in 18th Int. Conf. on Transparent Optical Networks (ICTON), July 2016, pp. 1–4.

[17] Mycek M, Pióro M, Żotkiewicz M. MIP model for efficient dimensioning of real-world FTTH trees. Telecommunication Systems. 2018 Jun 1: PP 1-20.

[18] F. D’Andreagiovanni, F. Mett, A. Nardin, and J. Pulaj, “Integrating LP-guided variable fixing with MIP heuristics in the robust design of hybrid wired-wireless FTTx access networks,” Appl. Soft Comput., vol. 61, no. Supplement C, pp. 1074–1087, 2017.

[19] Fritzsche, Lutz, Mathias Schweigel, and Rong Zhao. "Integrated Network Planning: A Key Success Factor for Network Operators." In Future Telco, pp. 43-52. Springer, Cham, 2019.

[20] Hervet, C., Faye, A., Costa, M.C., Chardy, M., & Francfort, S. (2013). Solving the two-stage robust FTTH network design problem under demand uncertainty. In Proceedings of the international network optimization conference. Costa Adeje, Spain.

[21] Angilella, Vincent, Matthieu Chardy, and Walid Ben-Ameur. "Fiber cable network design in tree networks." European Journal of Operational Research 269, no. 3 (2018): 1086-1106.

[22] Angilella, V. (2018). Optimal design of Fiber To The Home networks (Doctoral dissertation, Institut National des Télécommunications)

[23] C. Bock, P. Chanclou, J. Finochietto, G. Franzl, M. Hajduczenia, T. Koonen, P. Monteiro, F. Neri, J. Prat, and H. Silva, “Artchitecture of future access networks,” in Next-Generation FTTH Passive Optical Networks: Research towards unlimited bandwidth access, 2008, pp. 5–46.

[24] B. Skubic, J. Chen, J. Ahmed, L. Wosinska, and B. Mukherjee, “A comparison of dynamic bandwidth allocation for EPON, GPON, and next-generation TDM PON,” IEEE Journals, vol. 47, pp. 40–48, 2009.

[25] B. Chen, J. Chen, and S. He, “Efficient and fine scheduling algorithm for bandwidth allocation in ethernet passive optical networks,” IEEE J. Sel. Topics Quantum Elect., vol. 12, no. 4, pp. 653 –660, 2006.

[26] C. Lee, “An algorithm for the design of multi-type concentrator networks,” J. Oper. Res. Soc., vol. 44, pp. 471–482, 1993.

[27] P. McGregor and D. Shen, “An algorithm for the access facility location problem,” IEEE Trans. Commun., vol. 25, pp. 61–73, 1977.

[28] H. Pirkul and V. Nagarajan, “Locating concentrators in centralized computer networks,” Annals of Operations Research, vol. 36, pp. 61–73, May 1992.

[29] S. Narasimhan and H. Pirkul, “Hierarchial concentrator location problem,” Computer Communications, vol. 15, pp. 185–191, March 1992.

[30] M. Lv and X. Chen, “Heuristic based multi-hierarchy passive optical network planning”in Wireless Communications, Networking and Mobile Computing, 2009.WiCom ’09. 5th International Conference on, sept. 2009, pp. 1–4.

[31] M. Lukasiewycz, M. Glaß, F. Reimann, and J. Teich, “Opt4J - A Modular Framework for Meta-heuristic Optimization,” in Proceedings of the Genetic and Evolutionary Computing Conference (GECCO 2011), Dublin, Ireland, 2011.

[32] C. Steve, “Designing low cost access networks with iptv performance constraints,” in Next Generation Internet Networks, 2008. NGI 2008, 2008, pp. 45 –52.

[33] Zotkiewicz, M., Mycek,M., & Tomaszewski, A. (2016). Profitable areas in large-scale FTTH network optimization. Telecommunication Systems, 61(3), 591–608.

[34] ITU-T. (2001). Generic Functional Architecture of Transport Networks. Tech. rep. Recommendation G.805.

Performance Evaluation of Different Data Mining Techniques for Social Media News Credibility Assessment

Sahar F. Sabbeh[1, 2]
College of computer science and engineering, University of Jeddah, KSA[1]
Faculty of computing and information sciences, Banha University, Egypt[2]

***Abstract*—Social media has recently become a basic source for news consumption and sharing among millions of users. Social media platforms enable users to publish and share their own generated content with little or no restrictions. However, this gives an opportunity for the spread of inaccurate or misleading content, which can badly affect users' beliefs and decisions. This is why credibility assessment of social media content has recently received tremendous attention. The majority of the studies in the literature focused on identifying features that provide a high predictive power when fed to data mining models and select the model with the highest predictive performance given those features. Results of these studies are conflicting regarding the best model. Additionally, they disregarded the fact that real-time credibility assessment is needed and thus time and resources consumption is crucial for model selection. This study tries to fill this gap by investigating the performance of different data mining techniques for credibility assessments in terms of both functional and operational characteristics for a balanced evaluation that considers both model performance and interoperability.**

Keywords*—*Data mining; performance evaluation; news credibility; Twitter; social media

I. Introduction

Social media platforms suffer from the lack of supervision over content which can result in the spread of inaccurate (fake) information either unintentionally or intentionally for deceptive purposes. That is why using data mining models for content credibility assessment has become an important practice in the context of social media. To date, the bulk of the work in the literature focused on identifying the most informative features for higher precision credibility. Features are extracted at different levels (user, topic and propagation) level [1]-[9]. User-related features such as account/profile, demographics, age, account age, followers, photo, behavior (tweeting, retweeting) can be extracted and used to evaluate source credibility as inaccurate news can probably be created and spread by automated software agents or fake accounts created only for this sake. Topic related features can content related or contextual related. Content - related features include visual and textual features which can be collected and analyzed using standard NLP and text analysis techniques (i.e. images included, Hash-tags, URLs, sentiments/subjective content, etc.). Contextual information includes topic headlines, users' comments, rates, likes, emotional reactions, number of shares, etc. For example, topic/post headline may be misleading (known as "clickbait") which implies non-credible content or at least irrelevant content.

The extracted features are fed into different classification models which are then evaluated to identify the best performance given those set of features. The used techniques include: Logistic Regression (LR) [9]-[12], Decision Trees [1], [3], [6], [13]-[17], [19], Artificial Neural Networks [20],[21],[22], Support Vector Machines(SVM) [6], [13], [14], [15], [17]-[21], [23] Random Forest (RF) [13], [15], [18], [24], Naïve Bayesian (NB) [6], [16]-[19], [21] and K-nearest Neighbor (KNN) [17], [20], [21]. SVM and Decision Trees are the most known and widely used models. Very few works tried to use other models such as Linear Discriminant analysis (LDA) [21] and Adaptive Boosting (Adaboost) [23]. The performance of data mining techniques for credibility analysis included only the most well – known techniques disregarding more advanced techniques that may better utilize the extracted features such as bagged and boosted ensemble models. Moreover, the results of the performances are difficult to compare as each study recommends a different model and therefore, no general agreement can be reached.

Additionally, those studies focused only on the functional capabilities of the models by evaluating their predictive power, which, despite being important, is not enough. Operational characteristics are as important as functional capabilities. These include evaluating time and memory usage during both training and runtime. That is, measuring the amount of time and memory during model training and the amount needed to classify new data. As long classification time or excessive memory usage may mean that the model is unsuitable for real-time environments. Thus, a benchmark or empirical analysis that focuses only on the predictive performance will be insufficient to evaluate models operability.

This study tries to fill the gap in the research by focusing on news credibility assessment on Twitter as a case study. The used dataset for this study is publically available at GitHub1, it contains a set of 9252 Twitter news related to US election 2016 represented by 23 mixed features (numerical and binary). A set of 13 chosen models that represent different learning models were used in this study to provide an empirical analysis of different models and to identify the extent to which they are applicable for credibility assessment. LDA was selected as a

[1] https://github.com/marianlonga/FakeNews

linear learning model, mixture discriminant analysis (MDA), SVM, KNN, and NB. Both Multi-layer perceptron (MLP) and learning vector quantization (LVQ) were selected as ANNs. CART and C50 represent tree-based models and finally, Bagging CART (BaggedCart), ADAboost, Gradient boosted machine (GBM) and RF represent ensemble learning models. The selected models are evaluated based on accuracy, precision, recall, F-measure and computational time (processing and classification) and memory usage.

This paper is organized as follows: in Section II, a review of the previous empirical analysis of different data mining models in credibility assessment is presented. Section III provides a step-by-step description of the study methodology. Experimental results are discussed in Section IV. And Section V concludes the study and sheds light on study limitations and possibilities for future work.

II. Data Mining for Credibility Assessment in Social Media

Data mining is a process that aims to analyze, identify hidden patterns, and discover knowledge from large volumes of data. Classification techniques are supervised techniques that classify data item into predetermined classes. These techniques construct models using the labeled data to predict the label of unknown data sets.

The data mining process begins by applying data preprocessing (i.e. data transformation, cleaning, feature selection, etc.) is applied to improve the classification efficiency of the algorithm. The data set contains each tuple is labeled to belong to a predefined class. Part of the tuples is used for model construction (training dataset). The models are represented as classification rules or mathematical formulae and are tested using a set of independent data samples/tuples (test dataset) otherwise overfitting may occur. Finally, accuracy rate of the model is calculated as the percentage of test set tuples that are correctly classified by the model. Data mining techniques have been used for assessing the credibility of both information content and source. Credibility is assessed in terms of multiple features that are related to the news source, content and propagation medium. Data mining techniques use the features at one or more levels to label information content and/or source as credible/non-credible or fake/real. The comparisons summarized in Table I were performed among different models have conflicting results regarding their relative performance to one another. In the work [6], [19], DT achieved higher performance than SVM while in [15] SVM achieves better performance than DT. In [17], two different datasets were used and DT achieved the highest performance among other models given the first dataset while KNN was the best given the second dataset. In [20] LR model outperformed more sophisticated non-linear models such as ANN, DT, and SVM. However, ANN proved higher performance in [21]. Ensemble models RF in [13], [18] and Adaboost in [23] proved higher performance over SVM.

In conclusion, there is a need for a unified study that analyzes the performance of different models and evaluates their performance and applicability for credibility assessment.

TABLE. I. Summary of Empirical Studies of Data Mining Models for Credibility Assessment

Study	Models	Best performance
[1]	• (SVM) • Decision trees • extremely randomized trees (ERT) • Naive bayes	**ERT**
[6]	• SVM • Decision trees • Bayes networks	**Decision tree**
[13]	• Decision trees • Random Forest • SVM	**Random Forest**
[14]	• Decision trees • SVM	**Decision tree**
[15]	• Decision tree • SVM • Random Forest	**SVM**
[16]	• Decision tree • Naïve Bayes	**Decision tree**
[17]	• SVM • Naïve Bayes • KNN • decision trees	• **Decision tree for 1st dataset** • **KNN for the 2nd dataset**
[18]	• Naïve Bayesian • SVM • Random forest	**Random Forest**
[19]	• Decision tree • Naïve Bayesian • SVM	**Decision tree**
[20]	• Logistic Regression (LOG) • SVM • KNN • ANN • Decision trees	**Logistic Regression**
[21]	• ANN • KNN • SVM • Naive Bayes • Linear discriminant analysis (LDA)	**ANN**
[23]	• SVM • Adaboost	**Adaboost**

III. Methodology

A. Dataset

The used dataset contains twitter news related to US elections 2016. The dataset contains 9252 Twitter news represented by 22 explanatory variables and one response variable. The predictors are related to both news content and source. The target variable labels each tweet to be fake/non-fake represented by (True/False) variable. The dataset contains 254 instances labeled “unknown” and 2749 with no label. For this study unlabeled observations and noisy/unknown ones were disregarded.[2] The remainder 5598 include approximately 87% labeled false/ to indicate non-fake/real news or other type of news (i.e. comment, etc.), where 13% are labeled True to indicate fake news. Dataset metadata is presented in Table II.

[2] Dealing with missing and noisy labels are out of the scope of this study.

B. Data Preprocessing

1) Data transformation

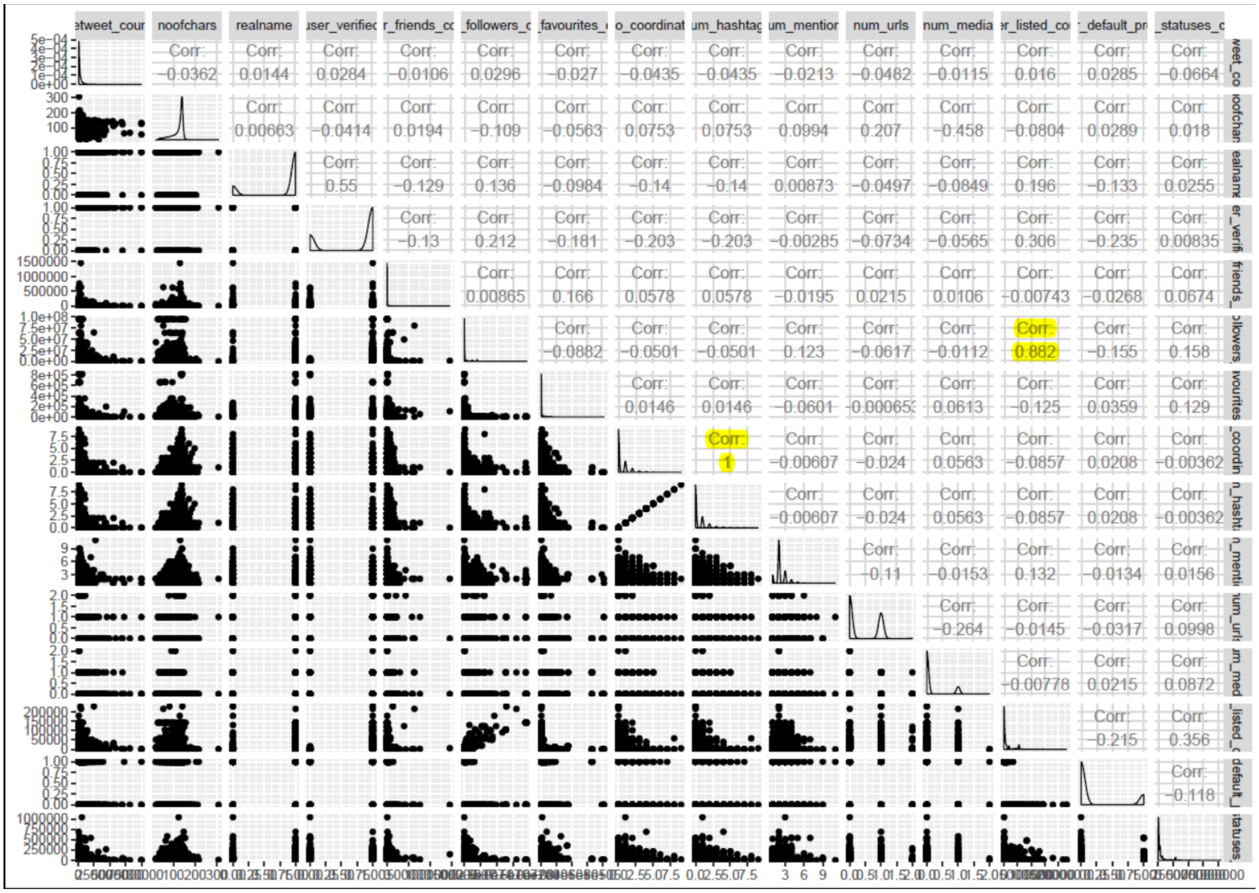
- The variable “Tweet Id” was removed for its irrelevance to the problem.
- The variable “text” was used to derive a new variable “noofchars” to indicate the number of characters in each tweet.
- The “Description” variable was removed and instead, a Boolean variable was added to indicate whether or not user profile has a description.
- “text”, “description”, “Tweet_id”, “created_at” and “tweet_source” were removed from the dataset as they are considered unrelated to the classification problem.
- The variables “user_name” and “user_screen_name” were removed and instead a new derived variable that indicates whether or not the account has a real/nick name is added.
- Binominal variables (user_verified, isfake with (true/false) values was transformed into binary (0/1).

2) Explanatory data analysis: The Purpose of exploratory analysis is to discover patterns or correlation between explanatory variables. The correlation matrix for the variables in the dataset is calculated. The pair-wise correlation among variables indicated low correlation among most of the variables except for the pairs: 1) “user_listed_counts-user_followers_count” and 2)“geo_coordinates–num_ hashtags”) as shown in Fig. 1(a) and correlation matrix in Fig. 1(b).

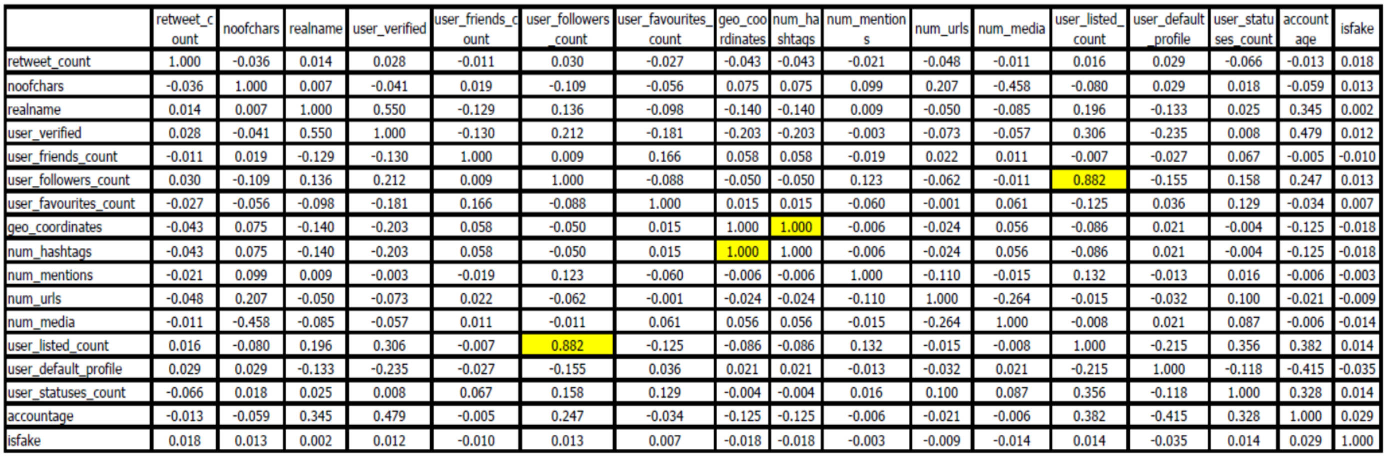
Strong correlation between explanatory variables (collinearity) can result in limitations of the analytical models. Variance inflation factors (VIF) test [25] was applied on data to verify collinearity among explanatory variables. VIF measures **the variance** between two variables **when correlated compared to variance when they are uncorrelated.** VIF value can indicate the degree of collinearity, where, VIF = 1 means variables are not correlated, 1 < VIF < 5 means moderately correlated and VIF >=5 indicates highly correlated variables. Results of VIF test indicated high VIF value for the variables user_followers_count =5.5, user_listed_count=6.77 and “infinity” for the variables (num_hashtags and geo_coordinates) as shown in Table III(a).

TABLE. II. DATASET METADATA

Variable	Type	level	
			Description
tweet_id	Integer	Content	Id for each Tweet.
created_at	Date/Time	Content	Date at which tweets had been created
retweet_count	Integer	Context	Number of time news had been retweeted
Text	Text	Content	The textual content of the news tweet
num_hashtags	Integer	Content	Number of hashtags included in the tweet.
num_mentions	Integer	Content	Number of users who are mentioned in the tweet.
num_urls	Integer	Content	Number of URLs included in the tweet.
num_media	Integer	Content	Number of images/videos included in the tweet.
user_screen_name	Text	User	Account display name
user_verified	Boolean	User	Whether or not the Twitter account is verified.
user_friends_count	Integer	User	The number of friends of the author
user_followers_count	Integer	User	The number of followers the user has.
user_favourites_count	Integer	User	The number of tweets the user has favorited.
tweet_source	Text	User	URL of the tweet
geo_coordinates	Integer	User	The geographic location of the Tweet as reported by the user or client application.
user_default_profile_image	Boolean	User	Whether or not the user uses the default profile image or his account
user_description	Text	User	Description included in the profile
user_listed_count	Integer	User	The number of public lists that the user is a member of.
user_name	Text	User	User’s unique name.
user_profile_use_background_image	Boolean	User	Does the profile has a background image
user_default_profile	Boolean	User	Is this the default user account??
user_statuses_count	Integer	User	Number of tweets issued by the user
isfake	Boolean	Content	Label each tweet as fake/real.



(a) Pairwise Correlation Matrix.



	retweet_count	noofchars	realname	user_verified	user_friends_count	user_followers_count	user_favourites_count	geo_coordinates	num_hashtags	num_mentions	num_urls	num_media	user_listed_count	user_default_profile	user_statuses_count	accountage	isfake
retweet_count	1.000	-0.036	0.014	0.028	-0.011	0.030	-0.027	-0.043	-0.043	-0.021	-0.048	-0.011	0.016	0.029	-0.066	-0.013	0.018
noofchars	-0.036	1.000	0.007	-0.041	0.019	-0.109	-0.056	0.075	0.075	0.099	0.207	-0.458	-0.080	0.029	0.018	-0.059	0.013
realname	0.014	0.007	1.000	0.550	-0.129	0.136	-0.098	-0.140	-0.140	0.009	-0.050	-0.085	0.196	-0.133	0.025	0.345	0.002
user_verified	0.028	-0.041	0.550	1.000	-0.130	0.212	-0.181	-0.203	-0.203	-0.003	-0.073	-0.057	0.306	-0.235	0.008	0.479	0.012
user_friends_count	-0.011	0.019	-0.129	-0.130	1.000	0.009	0.166	0.058	0.058	-0.019	0.022	0.011	-0.007	-0.027	0.067	-0.005	-0.010
user_followers_count	0.030	-0.109	0.136	0.212	0.009	1.000	-0.088	-0.050	-0.050	0.123	-0.062	-0.011	0.882	-0.155	0.158	0.247	0.013
user_favourites_count	-0.027	-0.056	-0.098	-0.181	0.166	-0.088	1.000	0.015	0.015	-0.060	-0.001	0.061	-0.125	0.036	0.129	-0.034	0.007
geo_coordinates	-0.043	0.075	-0.140	-0.203	0.058	-0.050	0.015	1.000	1.000	-0.006	-0.024	0.056	-0.086	0.021	-0.004	-0.125	-0.018
num_hashtags	-0.043	0.075	-0.140	-0.203	0.058	-0.050	0.015	1.000	1.000	-0.006	-0.024	0.056	-0.086	0.021	-0.004	-0.125	-0.018
num_mentions	-0.021	0.099	0.009	-0.003	-0.019	0.123	-0.060	-0.006	-0.006	1.000	-0.110	-0.015	0.132	-0.013	0.016	-0.006	-0.003
num_urls	-0.048	0.207	-0.050	-0.073	0.022	-0.062	-0.001	-0.024	-0.024	-0.110	1.000	-0.264	-0.015	-0.032	0.100	-0.021	-0.009
num_media	-0.011	-0.458	-0.085	-0.057	0.011	-0.011	0.061	0.056	0.056	-0.015	-0.264	1.000	-0.008	0.021	0.087	-0.006	-0.014
user_listed_count	0.016	-0.080	0.196	0.306	-0.007	0.882	-0.125	-0.086	-0.086	0.132	-0.015	-0.008	1.000	-0.215	0.356	0.382	0.014
user_default_profile	0.029	0.029	-0.133	-0.235	-0.027	-0.155	0.036	0.021	0.021	-0.013	-0.032	0.021	-0.215	1.000	-0.118	-0.415	-0.035
user_statuses_count	-0.066	0.018	0.025	0.008	0.067	0.158	0.129	-0.004	-0.004	0.016	0.100	0.087	0.356	-0.118	1.000	0.328	0.014
accountage	-0.013	-0.059	0.345	0.479	-0.005	0.247	-0.034	-0.125	-0.125	-0.006	-0.021	-0.006	0.382	-0.415	0.328	1.000	0.029
isfake	0.018	0.013	0.002	0.012	-0.010	0.013	0.007	-0.018	-0.018	-0.003	-0.009	-0.014	0.014	-0.035	0.014	0.029	1.000

Fig. 1. (b) Correlation Matrix between Dataset Explanatory Variables.

The variable with the highest VIF value is removed from the dataset and the VIF test is repeated as values may change after each variable is removed. Results after removing “geo_coordinates” variable and repeating the test for the 2nd time indicated low VIF for “num_hashatgs” while both “user_followers_count” and “user_listed_count” still have high VIF values as shown in Table III(b).

The variable with the highest VIF value “user_listed_count” was removed and the test was repeated. Results of the 3rd test indicated low VIF value for all the variables as shown in Table III(c).

3) Variable selection: An important step before model training is to select the features with the highest predictive p*ower*. For this study, features are evaluated and ranked based on the model in [27]. The model measures the effect of each variable on the target via an iterative variables' permutations process. The model calculates the mean decrease importance of each variable based on which variable is confirmed or rejected. Results of the feature selection model confirmed all the selected variables as shown in Fig. 2 and Table IV.

TABLE. III. (A) VIF VALUES FOR THE 1ST TEST

Variables	VIF
retweet_count	1.013285
noofchars	1.366300
realname	1.467026
user_verified	1.833684
user_friends_count	1.058622
user_followers_count	5.362508
user_favourites_count	1.143870
geo_coordinates	Inf
num_hashtags	Inf
num_mentions	1.060546
num_urls	1.144046
num_media	1.375882
user_listed_count	6.621056
user_default_profile	1.239349
user_statuses_count	1.549332
Accountage	1.776116
Isfake	1.004115

(B) VIF VALUES FOR THE 2ND TEST

Variables	**VIF**
retweet_count	1.014
Noofchars	1.360
Realname	1.456
user_verified	1.852
user_friends_count	1.056
user_followers_count	**5.435**
user_favourites_count	1.109
num_hashtags	1.065
num_mentions	1.055
num_urls	1.148
num_media	1.382
user_listed_count	**6.689**
user_default_profile	1.228
user_statuses_count	1.564
Accountage	1.774
Isfake	1.003

(C) VIF VALUES FOR THE 3RD TEST

Variables	**VIF**
retweet_count	1.013
Noofchars	1.364
Realname	1.474
user_verified	1.815
user_friends_count	1.051
user_followers_count	1.142
user_favourites_count	1.088
num_hashtags	1.062
num_mentions	1.055
num_urls	1.149
num_media	1.380
user_default_profile	1.227
user_statuses_count	1.233
Accountage	1.768
Isfake	1.005

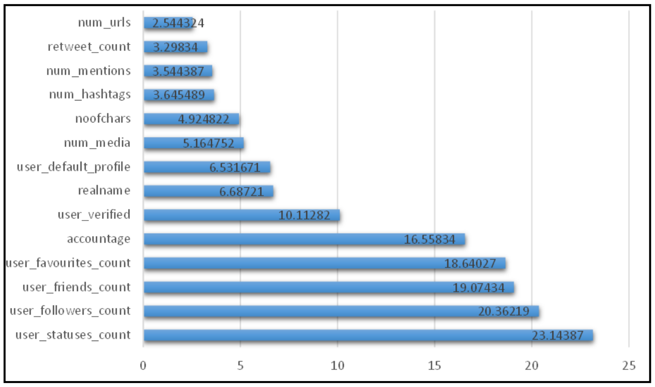


Fig. 2. Mean Importance of the Explanatory Variables.

TABLE. IV. MEAN IMPORTANCE OF THE EXPLANATORY VARIABLES

Variable	**meanImp**	**decision**
user_statuses_count	23.14387	Confirmed
user_followers_count	20.36219	Confirmed
user_friends_count	19.07434	Confirmed
user_favourites_count	18.64027	Confirmed
accountage	16.55834	Confirmed
user_verified	10.11282	Confirmed
realname	6.68721	Confirmed
user_default_profile	6.531671	Confirmed
num_media	5.164752	Confirmed
noofchars	4.924822	Confirmed
num_hashtags	3.645489	Confirmed
num_mentions	3.544387	Confirmed
retweet_count	3.29834	Confirmed
num_urls	2.544324	Confirmed

C. *Analytical Models*

A set of the most known and most widely used models for fake news detection in the literature were selected for this study. The selected models cover different learning models (linear, non-linear, tree-based and ensemble).

1) *Linea- learning models*

- LDA: LDA is a linear learning model that tries to find for a grouping of predictors that can discriminate two targets. LDA is related to regression as they both try to express the relationship between one dependent response variable and a set of independent variables. However, LDA uses continuous independent variables and a categorical dependent variable. The label for the new instance is estimated by the probability that inputs belong to each class and the instance is assigned the class with the highest probability calculated based on Bayes Theorem [28].

2) *Non-linear learning models*

- Mixture Discriminant Analysis (MDA): MDA is an extension of LDA that models the within-group multivariate density of predictors through a mixture (i.e., a weighted sum) of multivariate normal distributions [29]. In principle, this approach is useful for modeling multivariate non-normality or nonlinear relationships among variables within each group, allowing for more accurate classification. to determine whether underlying subclasses may be present in each group.

- SVM: A supervised learning model that analyses data in order to identify patterns. Given a set of labeled training data, SVM represents instances in the dataset as points in a high-dimensional space and tries to identify the best separating hyperplanes between different classes. New instances are represented in the same space and are classified to a specific class based on their closeness to the separating gap [30].

- NB: Naïve Bayesian (NB) is a classification technique that is based on Bayes' theorem [31]. It assumes complete variables independence, as the presence/absence of one variable is unrelated to the presence/absence of any other feature. It considers that all variables independently contribute to the probability that the instance belongs to a certain class. NB bases its predictions for new observations based on the analysis of previous observations. NB model usually outputs a probability score and class membership.

- KNN: KNN is an Instance-based or memory-based learning, labeling new instances is based on in-memory instances stored in advance. In KNN, no internal model is constructed, and computations are performed at classification time. KNN only stores instances of the training data in the features space and the class of an instance is determined based on the majority votes from its neighbors. The instance is labeled with the class most common among its neighbors. KNN determines neighbors based on distance using Euclidian, Manhattan or Murkowski distances for continuous variables and hamming for categorical variables. Calculated distances are used to identify a set of training instances (k) that are the closest to the new point and label is assigned based on them [32].

3) ANNs: ANNs try to mimic the performance of the biological neural network of the human brain. ANNs are adaptive, fault tolerant and can learn by example. An ANN is composed of a set of connected neurons organized in layers. The input layer communicates with one or more hidden layers, which in turn communicates with the output layer. Layers are connected by weighted links. Those links carry signals between neurons usually in the form of a real number. The output of each neuron is a function of the weighted sum of all its inputs. The weights on the connection are adjusted during the learning phase to represent the strengths of connections between nodes. ANNs come with many structures. The most common structures are feed-forward neural network (single and multi-layer) and recurrent neural nets. Multilayer perceptron (MLP) is a feed-forward ANN that contains at least one hidden layer. Neurons in each layer use supervised learning techniques [33]. LVQ is also a feed-forward ANN that is based on the winner – takes – all learning approach. In this approach, the distance is measured between each data point and the output. The smaller distance indicates a winner which is then adopted by adjusting its weights. It's as if, the prototype is moved closer if it correctly classifies the data point or moved away if otherwise [34].

4) Tree-based learning: Tree-based learning makes use of decision trees as a predictive model. Items are represented in a tree structure. In such structure, nodes represent test points for variables, leaves represent class labels and branches represent a combination of variables that lead to class labels [35]. Two popular implementations of DTs are a) CART [36] and C50 [37]. CART is a binary DT that can be used for classification and regression. For classification, CART used Gini index function to indicate the purity of the leaf nodes. C5.0 algorithm is used to build decision tree or a rule set. It works by splitting the sample based on the field that provides the maximum information gain. It uses subsamples based on a variable and iteratively split data until subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed/pruned.

5) Ensemble learning: Ensemble learning trains multiple models using the same learning algorithm and set learners to solve the problem. The main causes of error in learning are due to noise, bias, and variance. Ensemble minimizes these factors and may produce a more reliable classification than a single classifier. Bagging (i.e. Bagging CART, Random Forest) and Boosting (i.e. Ada Boost and Stochastic Gradient Boosting) get N learners by generating additional data in the training stage. N new training data sets are produced by random sampling with replacement from the original set. By sampling with replacement, some observations may be repeated in each new training data set. In the case of Bagging, any element has the same probability to appear in a new data set. However, for Boosting the observations are weighted and therefore some of them will take part in the new sets more often. Both are good at reducing variance but only Boosting tries to reduce bias, Bagging may solve the over-fitting problem, while Boosting can increase it [38].

D. Performance Evaluation Metrics

The performance of the selected models' predictive power is evaluated based on accuracy, precision, recall, and F-measure (F1).

1) *Accuracy:*

1) Accuracy: Indicates the ability of the model to differentiate the fake and real instances correctly. It's the proportion of true positive (TP) and true negative (TN) in all evaluated news:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

where

TP: is the total number of tweets correctly identified as fake.

FP: is the total number of tweets incorrectly identified as fake.

TN: is the total number of tweets correctly identified as real.

FN: is the total number of tweets incorrectly identified as real.

2) Precision and recall: Precision and recall can give a better insight into the performance as they do not assume equal misclassification costs. Precision indicates is the fraction of tweets correctly classified as fake among all classified instances, while recall is the fraction of tweets correctly classified as fake over the total number of fake tweets. relevant instances.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

3) F-measure: F-measure (F1) is calculated based on a combination of both precision and recall to provide a better evaluation of predictive performance.

$$F_1 = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \quad (4)$$

IV. Model Training and validation

Model training is an important step, as based on which models will behave. During this step, models are fed with labeled training dataset. Dataset was split into 80% for training and 20% for testing. For model training, 5 x 2-fold cross-validation was applied as recommended by [26]. Initial parameters are tuned via grid search during the training stage. The optimal parameter values are selected based on cross-validated accuracy as shown in Table V and the mean accuracy achieved by the models during the cross-validation is shown in Fig. 3.

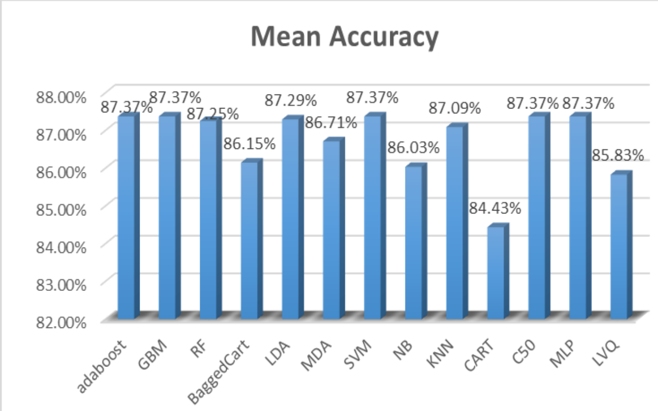


Fig. 3. Mean Accuracy Values During Cross Validation.

TABLE. V. PARAMETERS VALUES

Model	Parameters	Tuning values
AdaBoost	Number of iterations	[**50**,100,150]
	Maximum depth	[**1**,2,3]
	Weight update coefficient	**Beriman**
C50	Trials(number of iterations)	**1**
	Model(tree-based or rule-based)	**Rules**
	Winnow(use feature selection?)	**True**
DT	Complexity parameter	[0.0005500550, 0.0007616146, **0.0012376238**]
GBM	Number of trees	[**50**,100,150]
	Interaction depth(number of splits)	[**1** , 2 , 3]
	Shrinkage(learning rate)	0.1
	Min observations in node	10
KNN	K	[5 , 7 , **9**]
LVQ	Learning capacity(size)	[6, **9** , 12]
	K	[1 , 6 , **11**]
MDA	Subclasses(#*Subclasses* Per Class)	[**2**, 3 , 4]
MLP	Learning function	Std_Backpropagation
	Maximum iterations(maxit)	100
	Initial weight matrix (initFunc)	Randomized_Weights
	number of units in the hidden layer(size)	[**1**, 3 , 5]
SVM	δ	0.08984069
	C (cost of penalty)	[**0.25**, 0.50 , 1.00]

V. Results and Discussion

The experiment was carried out on an Acer machine with 64-bit Windows 10 OS, Intel® Core™ i7 – 7500U CPU @ 2.70GHZ and 8 GB Memory using R language. In order to test the performance of the selected models, unlabeled 20% of the dataset was used as an input the trained models for performance evaluation. Results of testing are used to compare the models based on a) predictive performance in terms of the selected metrics, and b) amount of time and memory usage during processing and classification time.

A. *Predictive Power Evaluation*

Results in Table VI show that linear-based learning model LDA achieved high performance compared to other models with 86.41% accuracy, 86.41% precision, 100% recall and 92.71% F1. Within the non-linear classifiers, SVM outperformed other non-linear models with 86.41 % accuracy, followed by MDA with 86.07%, KNN with 85.82% where NB achieved the lowest accuracy of 85.57%. SVM also outperformed the other non-linear models in both recall and F1 values followed by MDA, KNN and finally NB. However, KNN outperformed all non-linear models with precision of 86.45% followed by SVM with 86.41%, MDA with 86.36% and finally NB with 86.35%. It worth noting that NB model works well only with categorical data and cannot perform on continuous data. Thus, discretizing the continuous data may lead to better performance of this model.

For ANNs – despite achieving 86.41% accuracy during training, LVQ accuracy dropped to 82.49% to achieve approximately 4% lower accuracy, 6% lower recall and 2% lower F1 compared to MLP where precision of the two models is almost the same with 86.41% for MLP and 86.36% for LVQ. For tree-based learning models both CART and C50 trees achieved the same performance over all metrics with 86.41% accuracy, 86.41% precision, 100% recall and 0.9261 F1. For ensemble learning – based models boosted models (GBM and AdaBoost) showed higher accuracy, recall, and F1 compared to bagged models (BaggedCart and RF) with 86.41% accuracy, 100% recall, and 92.71% F1. However, BaggedCart achieved 86.59% precision which outperforms all the ensemble-learning models. Comparison between different models is shown in Fig. 4.

B. *Operational Characteristics Evaluation.*

Beside their predictive capabilities, operational characteristics in terms of runtime and memory usage were tested for each model during both processing and classification as shown in Table VII. the running statuses of each model was obtained using "profvis" profiling tool in R. Results show variation in time and memory consumption as Adaboost has the maximum processing time which is much longer than all other models recording *1 hour 48 seconds and 350 milliseconds* while, the processing time of all other models ranged from 350 milliseconds for LDA (lowest processing time) to 42 seconds, 450 milliseconds for RF. For non-linear models, KNN achieved the lowest processing time during training, followed by MDA, SVM, and finally NB while MDA achieved the lowest classification time followed by KNN, SVM, and NB. For memory usage, KNN had the minimum usage during training and classification followed by MDA, NB, and finally SVM. For tree-based models, C50 outperformed CART in training time while they both achieved the same classification time. For memory usage, CART had the lowest memory usage. For ANNs, LVQ outperformed MLP with lower time and memory usage in both phases. For ensemble learning models BaggedCart achieved the lowest processing time while GBM achieved the lowest classification time and memory usage during both training and classification among the rest of the models. It is worth noting that despite their high processing time, AdaBoost achieved reasonable classification time in relevance with the ensemble learning models. The best classification time was achieved by GBM and LVQ (10 milliseconds), followed by CART, C50 and MDA (30 milliseconds). LDA had the lowest memory usage during classification, followed by KNN they both had less than 200 MB memory usage. In runtime, MLP and LVQ achieved the lowest memory usage followed by KNN and LDA.

A comparison between the models based on time and memory usage is found in Fig. 5(a,b,c,d).

Choosing the suitable model has to balance between high predictive performances, low classification time and memory usage. That's why LDA and CART can be recommended as they provide high predictive power with low time and memory usage compared to other models. GBM is recommended too as it gives a good balance with the same performance with lower classification time and memory but higher processing time and memory.

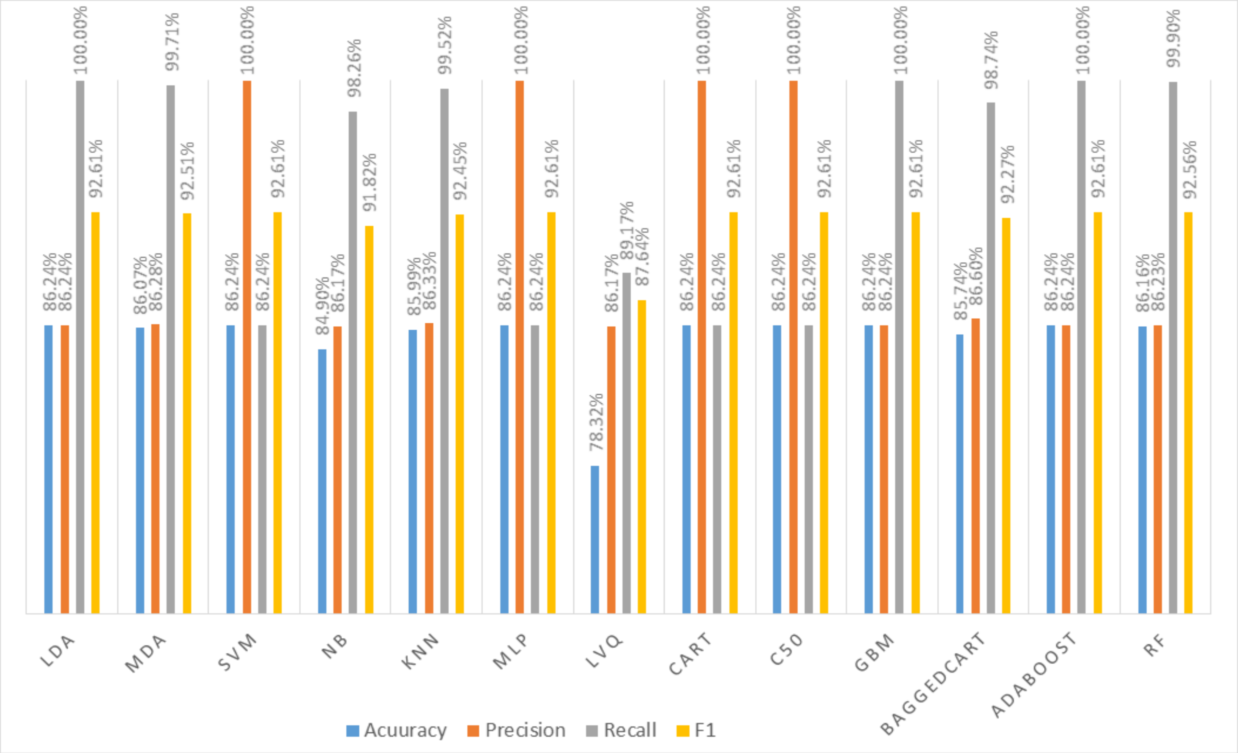


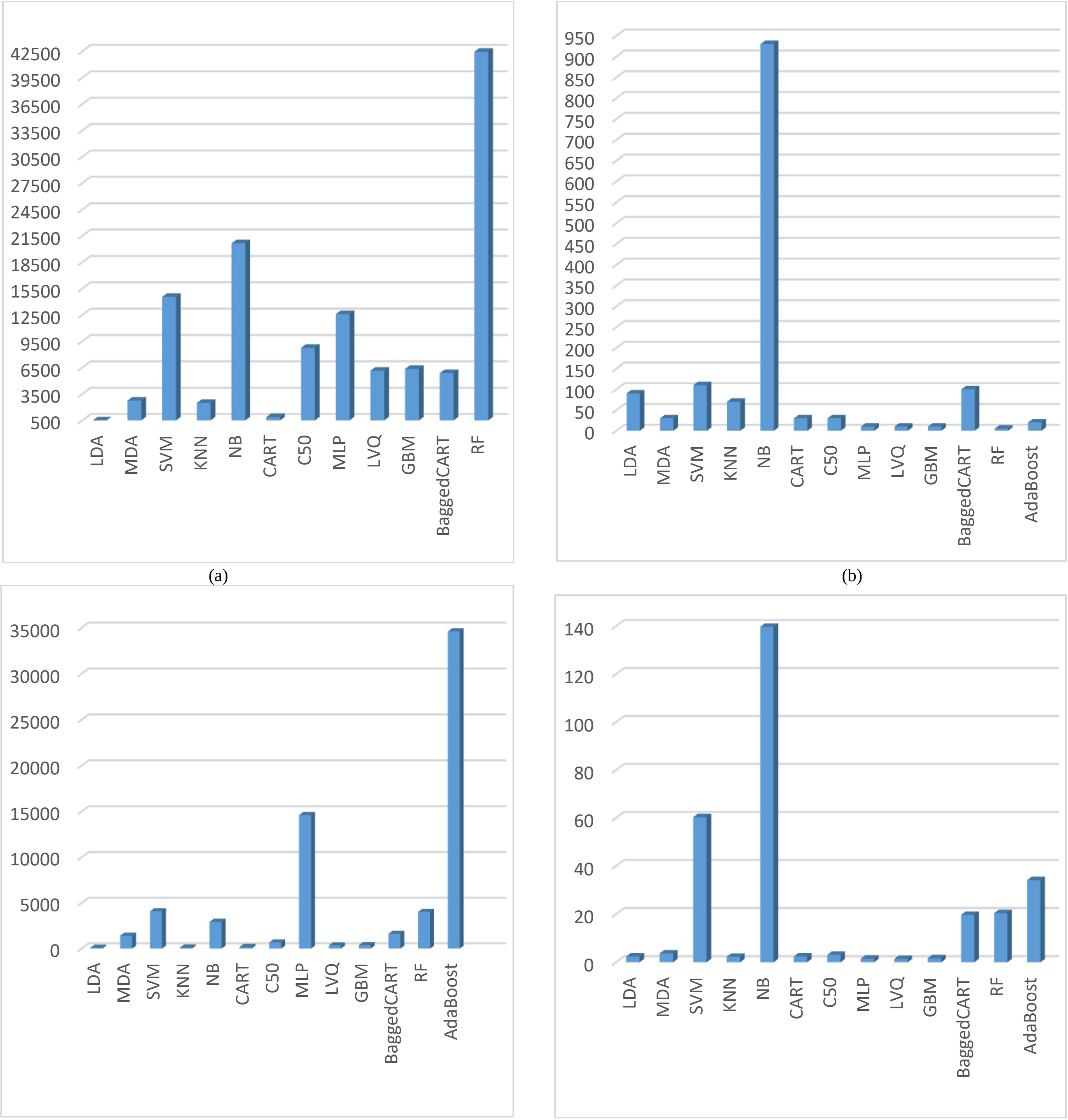
Fig. 4. Performance Comparisons of the Models.

TABLE. VI. MODELS' PERFORMANCE

Model	**Accuracy**	**Precision**	**Recall**	**F1**
Linear Based learning				
LDA	**0.8641**	0.8641	1	0.9271
Non-Linear Learning				
MDA	0.8607	0.8636	0.9961	0.9251
SVM	**0.8641**	0.8641	**1**	**0.9271**
NB	0.8557	0.8635	0.9894	0.9222
KNN	0.8582	**0.8645**	0.9913	0.9236
ANN				
MLP	0.8641	0.8641	1	0.9271
LVQ	0.8249	0.8636	0.9469	0.9003
Tree-based Learning				
CART	0.8641	0.8641	1	0.9271
C50	0.8641	0.8641	1	0.9271
Ensemble learning				
GBM	**0.8641**	0.8641	**1**	**0.9271**
BAGGEDCART	0.8549	**0.8659**	0.9846	0.9214
AdaBoost	**0.8641**	0.8624	**1**	**0.9271**
RF	0.8632	0.8645	0.9981	0.9265

TABLE. VII. TIME AND MEMORY CONSUMPTION OF THE MODELS

Model	**Processing**		**Classification**	
	Time(ms)	**Memory(MB)**	**Time(ms)**	**Memory(MB)**
Linear Learning				
LDA	**00:00:00.350**	**104**	**00:00:00.090**	2.4
Non-linear Learning				
MDA	00:00:02.750	1419.4	**00:00:00.030**	3.7
SVM	00:00:14.550	4097.7	00:00:00.110	60.3
NB	00:00:20.650	2941.5	00:00:00.930	139.7
KNN	**00:00:02.480**	**119.6**	00:00:00.070	**2.3**
Tree-based Learning				
CART	00:00:00.870	**180**	**00:00:00.030**	2.4
C50	**00:00:08.750**	690	00:00:00.030	3.1
ANN				
MLP	00:00:12.580	14629.4	00:00:00.010	1.5
LVQ	**00:00:06.150**	**355.3**	**00:00:00.010**	**1.4**
Ensemble Learning				
GBM	00:00:06.360	**380.4**	**00:00:00.010**	**1.7**
BaggedCART	**00:00:05.880**	1624.6	00:00:00.100	19.7
AdaBoost	01:00:48.530	34620.7	00:00:00.110	4.1
RF	00:00:42.450	4032.2	00:00:00.050	20.4



(a)

(b)

(c)

(d)

Fig. 5. (a) Processing Time of Models3. (b) Classification Time of Models. (c) Processing Memory usage. (d) Memory usage During Classification.

[3] AdaBoost Processing time is not included due to its large value compared to other models (3648.53 second).

VI. Conclusion and Future Work

This study tries to present an evaluation of the performances of different data mining models for credibility assessment in the context of social media. This study focused on Twitter news credibility assessment as a case study. The bulk of works in the literature focused on identifying the most informative features, feed those features into different models to select the model with higher predictive power and all of them disregarded time and memory consumption during both processing and runtime. Results of these studies contrast each other and cannot give a unified decision. This study tries to address this limitation by benchmarking different data mining models for news credibility assessment on Twitter. Models are evaluated in terms of their predictive performance using Accuracy, Precision, Recall and F-measure and time and memory usage during both processing and prediction.

However, the study still has some limitations and future research opportunities. First, the results on Twitter data may not be applicable on different social media contexts (i.e. blogs, Facebook, etc.). One possible future research shall utilize different datasets in different contexts for the evaluation. Another possible future work can be to explore the performance of other models including the less well known models and deep learning models. Performance can be evaluated with missing and noisy labels.

References

[1] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, Karl Aberer, “Web Credibility: Features Exploration and Credibility Prediction”, in the proceedings of European Conference on Information Retrieval. ECIR 2013:Advances in Information Retrieval pp 557-568, 2013.

[2] John ODonovan, Byungkyu Kang, Greg Meyer, Tobias Hollerer, Sibel Adal, “Credibility in Context: An Analysis of Feature Distributions in Twitter”, In the proceedings of the International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, 2012.

[3] A. A. A. Mansour, "Labeling Agreement Level and Classification Accuracy," 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Naples, 2016, pp. 271-274. doi: 10.1109/SITIS.2016.51

[4] Dana Movshovitz-Attias, Yair Movshovitz-Attias,Peter Steenkiste, Christos Faloutsos, “Analysis of the Reputation System and User Contributions on a Question Answering Website: StackOverflow”. In the Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining Pages 886-893. 2013.

[5] Ruohan Li, Ayoung Suh , “Factors Influencing Information credibility on Social Media Platforms: Evidence from Facebook Pages”, In the proceedings of the 3rd Information Systems International Conference (ISICO2015), 2015.

[6] C. Castillo, M. Mendoza, and B. Poblete, “Information credibility on twitter”, In the Proceedings of the 20th international conference on World wide web, Hyderabad, India, 2011.

[7] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, Julia Schwarz, “Tweeting is Believing? Understanding Microblog Credibility Perceptions”, CSCW 2012, USA.

[8] Kanda Runapongsa Saikaew, Chaluemwut Noyunsan, “Features for Measuring Credibility on Facebook. Information”. In the proceedings of the XIII International Conference on Computer Science and Information Technology (ICCSIT 2015),Thailand, 2015.

[9] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro, “Some Like it Hoax: Automated Fake News Detection in Social Networks”, CoRR,abs/1704.07506, 2017.

[10] Mehrbod Sharifi, Eugene Fink, and Jaime G. Carbonell. “Detection of Internet scam using logistic regression”. Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, pages 2168–2172, 2011.

[11] James Fairbanks, Natalie Fitch, Nathan Knauf, Erica Briscoe, “Credibility Assessment in the News: Do we need to read?”, MIS2’18, Feb 2018, Los Angeles, California USA.

[12] William Ferreira and Andreas Vlachos. “Emergent: a novel dataset for stance classification”. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016.

[13] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media", in Proceedings of International Conference on Data Mining, pp. 103-1108, 2013.

[14] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. “Enquiring minds: Early detection of rumors in social media from enquiry posts”. In Proceedings of the 24th International Conference on World Wide Web . ACM, 1395–1405, 2015.

[15] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. “Detect rumors using time series of social context information on microblogging websites”. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 1751–1754. 2015.

[16] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. “Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy”. In Proceedings of the 22nd international conference on World Wide Web. ACM, 729–736. 2013.

[17] Manish Gupta, Peixiang Zhao, Jiawei Han, “Evaluating Event Credibility on Twitter”, Proceedings of the 2012 SIAM International Conference on Data Mining, pages = 153-164.

[18] Rim El Ballouli, Wassim El-Hajj, Ahmad Ghandour, Shady Elbassuoni, Hazem Hajj and Khaled Shaban, “CAT: Credibility Analysis of Arabic Content on Twitter”, Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 62–71, 2017.

[19] Sahar. F. Sabbeh, S. Batawah, Arabic news credibility on Twitter: An Enhanced Model using Hybrid Features", Journal Of Theoretical And Applied Information Technology , Vol 96 April 2018.

[20] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, Fabiana Zollo, “Polarization And Fake News: Early Warning Of Potential Misinformation Targets”, Arxiv:1802.01400v1 [Cs.Si] 5 Feb 2018.

[21] R.Deepa Lakshmi , N.Radha , “Supervised Learning Approach for Spam Classification Analysis using Data Mining Tools “, (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 09, p= 2783-2789, 2010.

[22] Marin Vuković ,Krešimir Pripužić, Hrvoje Belani, “An Intelligent Automatic Hoax Detection System”, In Knowledge-Based and Intelligent Information and Engineering Systems , pages 318–325. Springer, Berlin, Heidelberg, September 2009.

[23] Benjamin Markines, Ciro CaŠuto, and Filippo Menczer, “Social spam detection. In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web . ACM, 41–48, 2009.

[24] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, Benno Stein” A Stylometric Inquiry into Hyperpartisan and Fake News”, arXiv:1702.05638, 2017.

[25] James Gareth, Witten Daniela, Hastie Trevor, Tibshirani Robert, “ An Introduction to Statistical Learning (8th ed.)”. Springer Science+Business Media New York. ISBN 978-1-4614-7138-7, 2017.

[26] Dietterich, T. G. “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms”. Neural Comput, 10(7):1895–1923, 1998.

[27] Miron B. Kursa, Aleksander Jankowski, Witold R. Rudnicki, "Boruta – A System for Feature Selection", Fundamental Informaticae volume101, pages:271–285, 2010.

[28] McLachlan, G. J. Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience. ISBN 0-471-69115-1. MR 1190469. 2004.

[29] Fraley, C., & Raftery, A. E. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 97(458), 611-631.2002.

[30] Cortes, Corinna; Vapnik, Vladimir N. "Support-vector networks". Machine Learning. Volume: 20 No:3: p:273–297. doi:10.1007/BF00994018, 1995.

[31] Russell, Stuart; Norvig, Peter. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955. 2003.

[32] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185. doi:10.1080/00031305.1992.10475879.

[33] Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (editors);

[34] Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press, 1986.

[35] T. Kohonen, "Learning vector quantization", in M.A. Arbib, The Handbook of Brain Theory and Neural Networks, Cambridge, MA: MIT Press, pp. 537–540, 1995.

[36] Rokach, Lior; Maimon, O. “Data mining with decision trees: theory and applications”. World Scientific Pub Co Inc. ISBN 978-9812771711. 2008.

[37] Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J., “Classification and regression trees”. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8. 1984.

[38] Rokach, L. (2010). "Ensemble-based classifiers". Artificial Intelligence Review. 33 (1-2): 1–39.

AUTHOR’S PROFILE

SAHAR F. SABBEH earned her B.Sc. degree in information systems from the Faculty of computers and information technology, Mansoura university, Egypt in 2003. She earned her M.Sc. also in information systems from the same department in 2008 and earned her Ph.D. degree in 2011 She has been a member of the IEEE since 2017. Dr. Sabbeh worked in Alzarka High Institution for management information systems from 2004 to 2009. She worked at Misr Higher Institution of Engineering and Technology, Mansoura, Egypt from 2009 till 2011. She worked with the Faculty of Computers and Information Technology, Banha University, Egypt during the period from 2011 - 2018 as an assistant professor. She also worked part time as an assistant professor in several reputable private universities in Cairo, Egypt. She worked as an Associate professor in the faculty of computers and information technology, King Abdul-Aziz university, KSA during the period from 2016–2018. Currently, she is an associate professor at the Faculty of Computers and Information Technology, Banha University, Egypt and an associate professor in the computer science and engineering, university of Jeddah, KSA. She supervised 5 M.Sc. and one Ph.D. students.

Rule-based Emotion AI in Arabic Customer Review

Mohamed M.Abbassy[1]
Information Technology Department
Faculty of Computers and Artificial Intelligence
Beni-Suef University, Beni-Suef, Egypt

Ayman Abo-Alnadr[2]
Information System Department
Higher Institute of Management and Information
Technology, Kafer el Shekh, Egypt

***Abstract*—The e-commerce emotion analysis is notable and the most pivotal advance since it catches the customer emotion in a product, and emotions with respect to product to decide if the customer attitude is negative, positive, or neutral. Posting on the customer's reviews have turned into an undeniably famous path for individuals to share with different customers their emotion and feelings toward a product. This review has a significant impact on sales in the future. The proposed system utilizes mixed word from an adjective (adj) and adverb to improve the emotion analysis process utilized a rule-based emotion analysis. The system extracts an Arabic customer review and computes the frequency of each word. At that point, it computes the emotion and score of each customer review. The system likewise computes the emotion and score of straightforward Arabic sentence.**

Keywords—Component; rule-based; emotion; customer review; Arabic

I. Introduction

With the rapid development of web applications, social network and online shopping, there moved toward becoming audits, comments and feedback generated by customers. These emotions can be about essentially anything, including products, politics issues, news and service. All of which should be handled and broke down to get a good estimation of what the customer thinks and feels. Before the accessibility of automatic emotion analysis tools, the way toward getting customer surveys was an incredibly cumbersome and time-consuming task [1].

Numerous emotion AI were created for English, yet in this paper are attempting to break another ground in this field and concoct a high accuracy Arabic based emotion analysis tool which isn't influenced by the utilization of vernaculars; a tool that enables Arab customers to analyze the e-commerce shopping, enabling them to know the general feeling about products being talked about. The Arabic language has numerous lingos that ought to be considered, wherein every vernacular implications of words can be very surprising. Arabic is a morphologically rich language and this can raise issues for any programmed content examination instrument [2, 3].

The enormous increment in e-commerce shopping in middle east especial Egypt, gulf countries, that made customer reviews significant in decision making procedure of a customer. The quantity of reviews for a product can be very high, particularly for a most prevalent product. A significant number of customers is interested on emotion of a product, so for this reason, they should initially read all the reviews to reach a resolution. What's more, since perusing countless is a dreary procedure and may create upsets in basic decision making [4].

In this way, an effective method for showing the general emotion of a product dependent on customer reviews is required. This paper inquire about in the investigation of product reviews are worried about ordering the general emotion for a specific product. As a customer review does not have a standard structure and may incorporate spelling blunders and equivalent words for the product features, emotion classification per feature can be troublesome.

Emotion AI is a procedure of extracting information from users' assessments. The decisions of the people get influenced by the conclusions of other individuals. Today, if any person needs to purchase an item then the person will initially look through the surveys and emotion about that item via an online shopping, a social network like Twitter, Facebook, and other user forums, at that point recognizable proof of assumption, turns out to be extremely troublesome from this colossal information physically. Thus, there is a need for a computerized emotion analysis system. The fundamental goal of this paper is to perform emotion AI for Arabic sentences.

II. Related Work

Elhawary and Elfeky [5] utilized that gathering Arabic business reviews, and dedicating 80% of the gathered business reviews to prepare their classifier which is utilized to recognize review's records. They developed various Arabic vocabularies used to investigate distinctive Arabic reviews and emotion. The extremity of every Arabic business review whether it is: positive, negative, neutral or mixed is made a decision about dependent on the assembled dictionaries.

Diverse strategies were utilized by El-Halees [6] to decide the extremity of various Arabic `s. The extremity of the entire Arabic comment is resolved first utilizing the vocabulary-based technique, where the output from the primary strategy (dictionary based) is considered as a preparation set for greatest entropy strategy, which is utilized to order these comments.

Another methodology has been proposed depends on translating the source Arabic emotions into English and after that utilization the equivalent relevant procedures to examine the came about English emotions. Almas and Ahmad in [7] utilized machine translation systems to translate the source comment or review from Arabic to English language before passing them to an English based emotion analysis system. The issue of this methodology was the loss of nuance after translating the source to English.

Rushdi-Saleh et al. [8] utilized another methodology was machine learning algorithm to arrange the extremity of Arabic reviews extricated from specific Web pages identified with motion movies. Inui et al. [9] think about receive making an interpretation of suppositions from English to Japanese, trailed by emotion analysis. They applied sentiment-oriented sentence filtering strategy to alleviate numerous interpretation mistakes that happen as a reaction of interpretation to decrease the impact of interpretation blunders in multilingual comment level review.

Choi et al. [10] presents a structure for emotion analysis, focus around the feeling piece of information that is identified with a supposition theme, for example, company or individual. They utilize a domain-specific sentiment classifier for every domain with the recently totaled signs (for example a subject or the theme of the emotion) in light of a proposed semi-supervised strategy. Yi et al. [11], Kim et al. [12], Choi et al. [10] extricate emotion about a subject spotlight on the estimation piece of information that is identified with a conclusion theme. This is characterized as an essential subject of supposition articulation in a sentence, for example, organization, individual or occasion.

III. Material and Methods

To perform emotion AI, basic Arabic content record, tweets or comment in online shopping are inputted by the client. At that point, the system takes a shot at it and figures its emotion and score. The design appeared in Fig. 1 show the working of rule-based emotion AI system.

A. Tokenizer

The system takes a product review in the Arabic language as an input; the input sentence is part into tokens through tokenizer. A token is a piece of an arrangement of characters in content that are combined together as an important semantic unit for handling. The tokenizer changes over a sentence into word level tokens comprising of a word, accentuation marks, and different symbols.

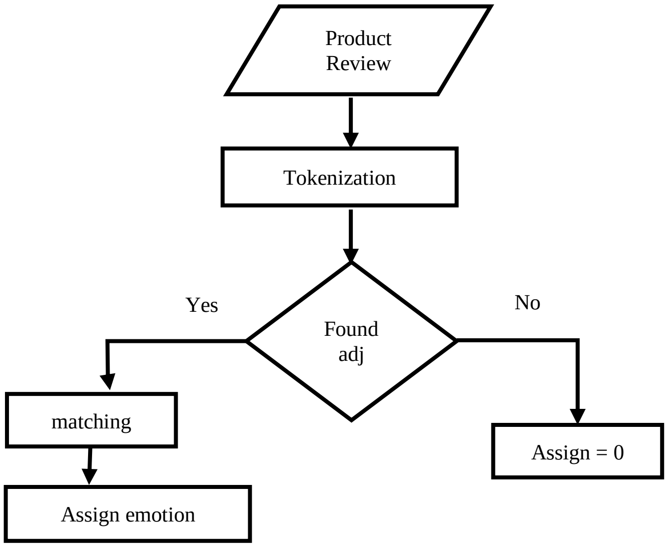


Fig. 1. Proposed Architecture of Rule-based Emotion AI System.

B. Procedure of Final Emotion and Score

Most of the work use adjectives only for emotion analysis, and some of them use nouns, verbs, adverbs or a combination of them. This proposed model use emotion adj and nouns of adj words because that some emotion do exclude any adj, however, express a negative or positive sentiment such as; The vast majority of the work use adjectives just for emotion investigation, and some of them use noun, verb, adverbs or a mix of them. utilize emotion adjectives and noun of adjectives since they noticed that some emotion statements do exclude any adjectives, however, express negative or positive emotion, for example, “رائع ”, “جامد جدا”, “ممتاز ”, “جميل جدا ”, “حلو مرة ”. That means “Wonderful”, “Excellent”, “Very beautiful”, “Very nice”. All of them are adjectives in this example, although it expresses positive emotions.

IV. Applying Rule

An algorithm is proposed to extend and recognize the emotion AI automatic of new feeling words utilizing total procedure and free online Arabic word references and vocabularies and calculate the frequency of each emotion word from a dataset. The scores of adj are appointed between -1 and 1. In the event that any of the token matches with terms of adj and adverb score content then their comparing scores are processed. At that point, the last emotion is assigned to content as positive, negative or natural utilizing calculation has been proposed below. Be that as it may, if any of the tokens in the content does not match with terms of adj score. At that point, 0 scores are assigned to that content lastly no emotion is assigned to that content. For example: “هذا المنتج جميل جدا”. In this example, the process of this content will be split into tokens as “هذا”, “المنتج ”, “جميل”, “جدا”. Presently, these tokens are coordinated with adj and adverb score content. Here, “جميل ”, “جدا” are found in adverb and adj word score content separately. Along these, relating scores of “جميل”, “جدا” are appointed as 0.1 and 1 individually. And afterward, the last score and emotion are registered by a system utilizing the proposed model in this paper.

A. Role of Score Adjective and Adverb

The score document contains the rundown of pre-processed emotion scores as appeared in the table below. Each line of the scoring document contains a word or expression alongside its emotion score. In the event that a word or expression which is found in content yet not found in score document at that point word or expression is given an emotion score of 0.

At that point, the emotion of content based on assumption scores of the terms in the content is registered. The emotion of content is equivalent to the whole of the assumption scores for each term in the content.

Grammatical forms information is most generally used in linguistic tasks. It is used to disambiguate sense which subsequently is used to coordinate component decision [13]. Researchers basically use adjective (adj) words and adverb as highlights to discover the emotion in content. Adjective (adj) words are most normally utilized as highlights among all grammatical features. There is a solid connection amongst adj and subjectivity of content. Indeed, even every one of the grammatical features assumes a critical role, yet just adj words as highlights feature the emotions with high exactness. An exactness of around 82.8% has been accomplished in film survey spaces by utilizing adj words just as highlights [14]. A few instances of positive and negatives have appeared in Table I.

B. Role of Score Content

1) Calculate strong function: If the score of an adj is greater than 0 then adj is positive and adv has a place with positive. For example, “جميل” has meant in English "beautiful" is a positive adj and “جداً” mean in English "Very" has a place with a strong adverb.

2) Calculate weak function: If the score of an adj is less than 0 then is a negative emotion and adverb has a place with is negative then score of both adj and adverb will be less than 0. For example, “سئ” that means in English “bad” is negative adj and “جدا” that mean in English “very” is also a negative adverb. The model has been proposed to calculate strong and weak functions in this work as shown in Fig. 2.

TABLE. I. SOME WORD OF POSITIVE AND NEGATIVE ADJECTIVE

Positive adjective					**Negative adjective**			
جميل	رائع	جيد	عالى	ماركة	ردئ	سئ	محلى	مستهلك
ممتاز	مبهر	ناجح	اقتصادى	فل	صعب	بطئ	اوفر	عادى
اجمل	مميز	مذهل	صاروخ	فل الفل	ممل	مقزز	تقليدى	زفت
تحفة	حلو	خامة	ملفت	ياعينى	وحش	غالي	صينى	ز الزفت
راض	كويس	قوى	عالمى	ياجمالوا	سلبي	بيئة	تقليد	اطران
ممتع	انبسطت	عملى	براند	ياحلوته	فاشل	كوبى	ضعيف	هباب

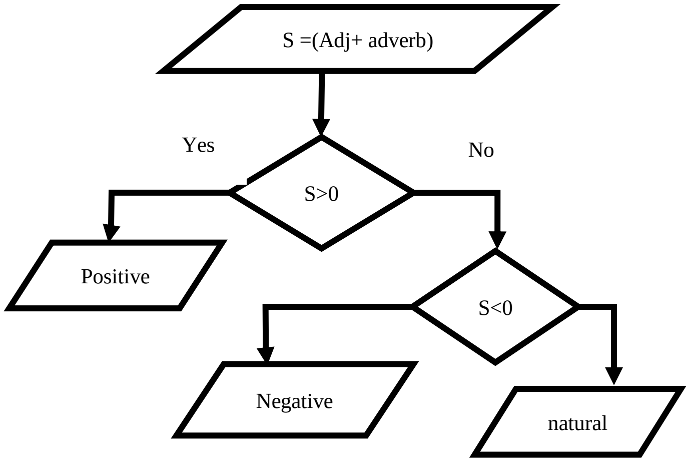


Fig. 2. The Rule-Based Model is to Calculate the Score of Arabic Sentences Whether Strong or Weak.

V. DESCRIPTION OF THE PROPOSED ALGORITHM

First step is to take an Arabic sentence as output for a system, the second step is the identification of relations and widespread words, third step is assigning a score to emotion according to pre-processed emotion scores as appeared in table previous, fourth step calculation a final score according to degree of strong or weak function according to relation identification as shown below.

If score > 0

Then Arabic sentence show positive emotion,

else if score < 0:

then the Arabic sentence show negative emotion,

else: the Arabic sentence shows no emotion or normal emotion.

VI. CONCLUSION

In this paper proposed technique to extract emotion focus from Arabic customer review from online shopping. In this paper, proposed Adjective and Adverb together are considered for performing emotion analysis as it gives preferable result. The purposed system helps in computing the emotion and score of customer review.

The proposed system computes the emotion and score of basic Arabic sentences. These sentences are part into tokens by tokenization process. At that point, with the assistance of adjective and adverb score records; emotion and score of these sentences are found by creating database for both negative and positive word used by customer.

REFERENCES

[1] Waters, John K. The Everything Guide to Social Media: All you need to know about participating in today's most popular online communities. Adams Media, 2010.

[2] Chiang, David et al. "Parsing Arabic dialects." Proceedings of the European Chapter of ACL (EACL) 2006: 112

[3] "Morphology (linguistics)", Wikipedia, the free encyclopedia." Last Accessed, 4 May 2019 .http://en.wikipedia.org/wiki/Morphology _(linguistics)

[4] S.Pednekar,K.Patil,R. Sawant,T.Shah, ”Sentiment Analysis On Online Product Reviews”, International Education & Research Journal [IERJ], pp.130-131,Mar 2017.

[5] Elhawary, Mohamed, and Mohamed Elfeky. "Mining Arabic Business Reviews." Data Mining Workshops (ICDMW), 2010 IEEE International Conference on 13 Dec. 2010: 1108-1113.

[6] A. El-Halees, “Arabic Opinion Mining Using Combined Classification Approach,” In: Proceedings of the International Arab Conference on Information Technology, Zarqa, Jordan, 2011.

[7] Y. Almas,K. Ahmad, “A note on extracting ‘sentiments’ in financial news in English, Arabic & Urdu,” In: The Second Workshop on Computation, al Approaches to Arabic Script-based Languages, Linguistic Society of America 2007 Linguistic Institute, Stanford University, Stanford, California., Linguistic Society of America, pp. 1–12, 2007.

[8] M. Rushdi-Saleh, M. Teresa Martín-Valdivia, L. Alfonso UreñaLópez,J. M. Perea-Ortega, “Bilingual Experiments with an ArabicEnglish Corpus for Opinion Mining. Language,” In: Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria. pp. 740-745, 2011.

[9] T. Inui, M. Yamamoto “Applying Sentiment-oriented Sentence Filtering to Multilingual Review Classification,” In: Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP, Chiang Mai, Thailand, pp. 51–58, 2011.

[10] Y. Choi, Y. Kim,S-H. Myaeng, “Domain-specific Sentiment Analysis using Contextual Feature Generation,” In: Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement(TSA'09), Hong Kong – China, pp. 37-44, 2009.

[11] S-K. Kim, E. Hovy, “Determining the sentiment of opinions,” In: Proceedings of the 20th international conference on computational linguistics (COLING 2004), Geneva, Switzerland. pp. 1367–1373, 2004.

[12] M. Elhawary,M. Elfeky, “Mining Arabic Business Reviews,” In: Proceedings of the2010 IEEE International Conference on Data Mining Workshops; pp. 1108-1113, 2010.

[13] Pang B, Lee L 2008 Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2):1-135.

[14] Pang B, Lee L, Vaithyanathan S 2002 Thumbs up? Sentiment classification using.

[15] Machine learning techniques. Proc. ACL-02 Conf. on Empirical methods in natural language processing, 10: 79-8.

Microcontroller-based Vessel Passenger Tracker using GSM System: An Aid for Search and Rescue Operations

Joel I. Miano*[1], Ernesto E. Empig[2], Alexander R. Gaw[3], Ofelia S. Mendoza[4], Danilo C. Adlaon[5]
Sheena B. Cañedo[6], Roan Duval A. Dangcal[7], Angelie S. Sumalpong[8]
Computer Applications Department, College of Computer Studies
Mindanao State University-Iligan Institute of Technology
Iligan City, Lanao Del Norte, Philippines 9200

Abstract—**The Maritime Transport industry in the Philippines has been growing through the years and has been a catalyst in the industrial development of the country. Although the maritime transport sector is one of the largest industries in the country, the safety devices and technology used are sluggish phase to change. The natural hazards and human error are main cause of maritime incidents, resulting to multiple casualties and missing persons every year of which this study seek to address the problem of safety in the maritime transport industry. The study aims to design and develop a system that will locate an overboard[1] passenger whenever a vessel is in distress. The Floating Overboard Accident Tracking System (FLOATS) was conceptualized by combining the Search Theory, Theory of Planned Behavior (TPB) and Disaster Preparedness, and the increasing availability of tracking device and monitoring technologies and the advancement of communication systems. The system consists of the Global Positioning System (GPS) for location data, Global System for Mobile (GSM) communications for the transmission and reception of emergency messages, Arduino-Nano microcontroller to handle the processing, the used of an inflatable life jacket with signal light and a rescue update display using an organic light emitting diode (OLED) for the search and rescue operations. Tests and surveys established the functionality, reliability, and acceptability of the system, which will greatly benefit maritime incident responders by securing vessel passengers from hazards and reducing the time allotted through speedy search and rescue operations.**

Keywords—Global Positioning System (GPS); Global System for Mobile communications (GSM); Organic Light Emitting Diode (OLED); Arduino-Nano microcontroller; tracking system; life jacket; life jacket light

I. INTRODUCTION

Maritime transport is the foundation of globalization and is the center of cross-border networks of transport that aid supply chains and empower international trade [1]. The Philippines having composed of 7,641 islands makes its shipping transport industry a vital part of the economic growth and development. One may assume that because of the archipelagic nature of the Philippines, the maritime industry is traditionally anchored in its economy [2]. The country's archipelagic setting requires an efficient maritime transport infrastructure and systems of safety aids. According to a statistical report compiled by the Philippine Maritime Industry Authority (2016), there is an increasing rate of passenger traffic yearly [3]. In 2017, a total of 72,438,609 passenger traffic based on the total embarking and disembarking data was reported. Alongside the volume of passengers are risks regarding their safety. The Philippines being situated near the Pacific Ocean makes it one of the most vulnerable countries in the world to weather-related extreme events. The Philippines placed second among 171 countries ranked on their risk level to disasters the report added that the country lacks 80.03% of coping capacity to minimize the negative consequences of natural hazards and climate change through direct action and the resources available [4]. In 2012, there were 610 reported persons killed or missing (lives lost) worldwide [5]. In addition to this, it was also stated that human errors and fatigue were featured eminently in these accidents. Over the recent years, most casualties recorded from maritime incidents are from passenger or roll-on/roll-off (RO-RO) ships and general cargo ships. There are 185 average number of deaths due to maritime accidents in the country yearly [6]. In addition to this, a report released by MARINA in 2016 states that there were 707 Search and Rescue (SAR) missions, 211 casualties, and 216 persons missing related to maritime incidents on that year. Several memorandum and circulars were made to increase maritime safety and resilience to hazards [7]. An example of a provision under a memorandum is to upgrade the maritime safety infrastructures like the navigation aids, lighthouses, vessel monitoring services and systems, and other maritime ancillary services like the weather bureaus. Memorandums and circulars were also made to prevent maritime accidents. While strict regulations are implemented, the number of search and rescue operations as stated earlier in a report by MARINA in 2016 is relatively higher compared to 126 number of maritime search and rescue operations in 2015. Moreover, the increasing number of passenger traffic yearly, implies the need for an appropriate technology to counter the consequences of maritime incidents and promote disaster resilience particularly in the maritime transportation sector.

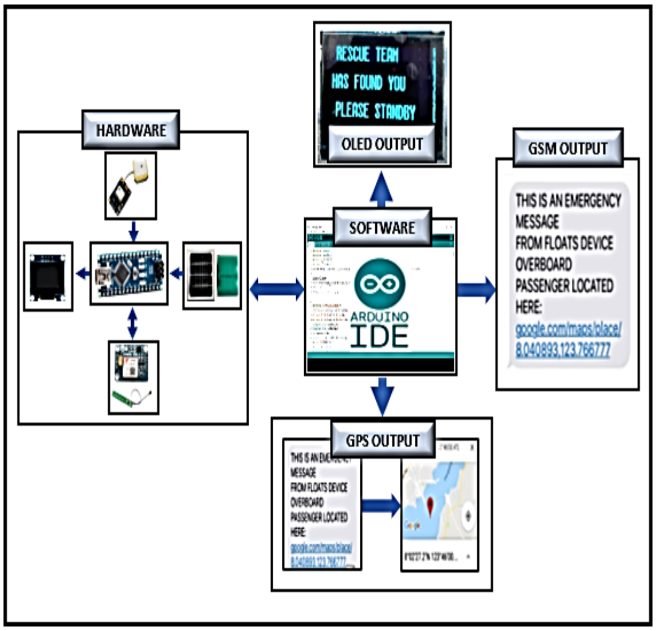
This study provides the design and development an appropriate technology to be used in times of maritime incidents. The technology is known as Floating Overboard Accident Tracking System (FLOATS) was integrated in a life jacket with an extend battery life span that sustains the

* Corresponding Author

tracking device[1] through a solar panel therefore assuring a higher survivability, reducing the exposure of passengers involved in a maritime incident/accident or natural hazards and minimizing the time allocated in locating the strayed passengers aiding the authorities in search and rescue mission. Next section describes the system design model of FLOATS.

II. System Design

Fig. 1 shows design model of the FLOATS prototype components hardware and software developing the tracking device that is integrated to the life jacket, increasing the probability of success of a search and rescue operation through GSM, GPS, and OLED output.



RESCUE TEAM
HAS FOUND YOU
PLEASE STANDBY

THIS IS AN EMERGENCY
MESSAGE
FROM FLOATS DEVICE
OVERBOARD
PASSENGER LOCATED
HERE:
google.com/maps/place/
8.040893,123.766777

Fig. 1. Design Model of the FLOATS.

A. Hardware

1) Arduino Nano Boar[2] – Arduino Nano is a small, complete, and breadboard-friendly board based on the ATmega328P. It has 22 Digital I/O pins and 8 Analog IN pins. The microcontroller board is used to handle the processes needed by the tracking device.

2) SIM800L GSM/GPRS Modul[3] – SIM800L is a GSM Module that features a complete Quad-band GSM/GPRS solution in a LGA type. The module can transmit Voice, SMS, and data information with low power consumption. The tracking device's SMS interfacing is made possible by SIM800L.

3) Neo M8N GPS Module[4] – Neo M8N is a GPS Module used to retrieve location and time information from GPS satellites. The Neo M8N GPS Module is a significant component of the tracking device that receives location information of the overboard passenger.

4) OLED Display[5] – Organic Light Emitting Diode (OLED) Display is an efficient and thinner display that can light up individual pixels when necessary and to different degrees. The maximum resolution of the OLED Display is 128x64. The OLED display shows the rescue updates sent by the search and rescue authorities.

5) Ni-MH Battery[6] – the battery used in the tracking device has a voltage of 7.2V with a capacity of 2200 mAh.

6) Solar Panel[7] – used in addition to mains-supply chargers for energy saving during the daytime with a maximum voltage of 17.5 V, current of 0.57A and maximum power of 10W.

B. Software

Arduino Integrated Development Environment[8] (IDE) – Arduino IDE is an open-source software program that allows user to write and upload code within a real-time work environment. The IDE had been instrumental in writing, compiling, and uploading codes to the Arduino Board. Writing the individual codes for each module and integrating all of them were vital in developing the firmware for the tracking device.

III. Methodology

Fig. 2 illustrates the comprehensive flow of the whole research process to guide the researchers in creating a prototype with an appropriate technology namely the Floating Overboard Accident Tracking System (FLOATS).

A. Analysis and Data Gathering

In this stage, the researchers started gathering data and relevant information, conducting preliminary investigation and interviews. Gathering and supporting facts about the existing problem to be able to design and develop the FLOATS. The researchers investigated first the scale of the problem in the Philippines. Having read articles and releases by the World Risk Report of the United Nations University Institute for Environment and Human Security, the researchers found out that the Philippines placed second among the 171 countries ranked on their risk level to disasters. The Philippine lacks eighty-percent 80% of coping capacity to minimize the negative consequences. With the increasing number of passenger traffic in the maritime transport industry of the country, the researchers saw the importance of safety in this particular sector. Secondary data from the Philippine Maritime Industry Authority (MARINA) and the Philippine Coast Guard (PCG) further supports the claim that there is an increasing number of search and rescue operations regarding maritime incidents in the Philippines. Researchers conducted a survey at the City Disaster Risk Reduction Management Office and Philippine Coast Guard Iligan Station concerning the existing tracking devices [12]. Based on the interviews conducted with the authorities, the researchers found out that there were no tracking devices used to locate or track overboard passengers when doing search and rescue operations.

[1] https://www.merriam-webster.com
[2] https://store.arduino.cc/usa/arduino-nano
[3] https://lastminuteengineers.com/sim800l-gsm-module-arduino-tutorial/
[4] https://www.u-blox.com/en/product/neo-m8-series
[5] https://www.thefreedictionary.com/OLED
[6] https://en.wikipedia.org/wiki/Nickel%E2%80%93metal_hydride_battery
[7] https://www.conserve-energy-future.com
[8] https://en.wikipedia.org/wiki/Arduino_IDE

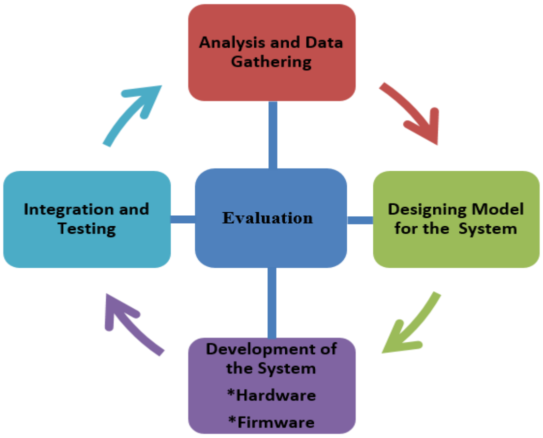

Analysis and Data Gathering
Designing Model for the System
Evaluation
Integration and Testing
Development of the System
*Hardware
*Firmware

Fig. 2. Research Methodology.

To understand the underlying technical and conceptual aspects of a search and rescue operation, the researchers came across with the Search Theory [8]. As the foundation of modern search and rescue operations, Search Theory became a pillar of this study. Sweep width, an indicator whether search conditions are ideal or not is heavily considered in the study. Sweep width is a key factor in determining the Probability of Success (POS) of a search and rescue operation.

B. Designing Model for the System

Developing the FLOATS, the researchers designed a model (Fig. 1) for the system with the following brief operation procedure:

a) When the life vest is submerged in the water, the life vest will automatically inflate. The victim must press the button to activate the device. An initialized message will be display on the OLED indicating that the device is activated. After 30 seconds, it will then send a message to the designated receiver's number via SMS (short message service).

b) The receiver will receive a message containing an overboard passenger's location converted to hyperlink including latitude and longitude coordinates that will point to the victim's exact location.

c) When the hyperlink is tapped, it will take the receiver to the web mapping service Google Maps. Then the device sends the first location message in 30 seconds, and the succeeding location messages are sent every one minute.

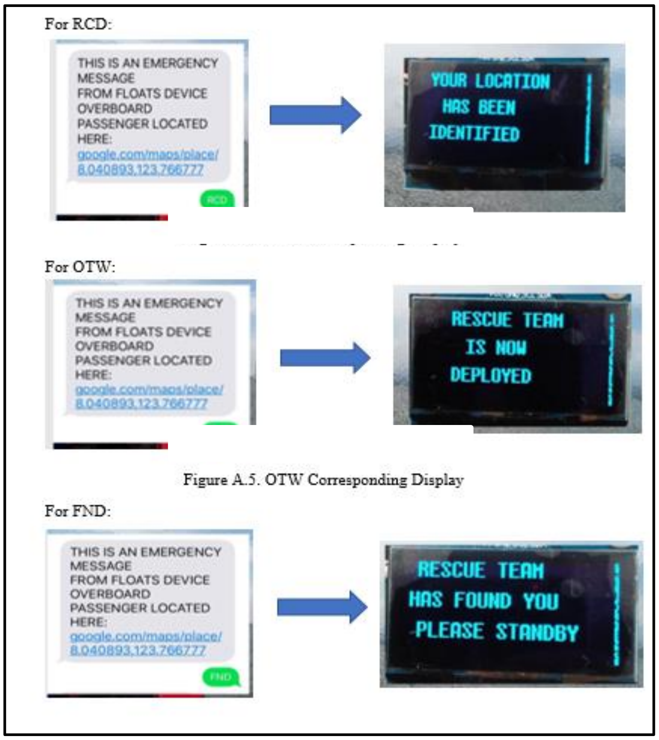
d) The rescuer can also send pre-configured messages to the victim, namely, 'RCD', 'OTW'; 'FND' as an assurance to the victim, the pre-configured message will be displayed to the OLED in Fig. 3. The device's reliability depends on the signal strength of the chosen network of mobile communications.

C. Development of the System

From the gathered data the researchers came up with a suitable design for the hardware that is fit for the maritime environment. The use of a Global Positioning System (GPS) device in the system corresponds to the Search Theory where the Probability of Detection (POD) is important especially during the planning phase of a search and rescue mission. The importance of POD also propelled the users to use the Global System for Mobile communication (GSM) as the transmission and reception means of the tracking system [10]. The use of a highly-reflective color of life jacket was also considered in the development of the device. Search objects can be easily detected when they contrast their background. In case of a night time incident, the use of lifejacket light is employed. An OLED display in Fig. 4 was also considered for the search and rescue updates.

The gathered data and the application of theories, models and frameworks were analyzed. The hardware needed such as the components and modules used were gathered. Each component was tested Fig. 5 integrating them all in a single circuit.

A circuit design was developed and all hardware components were integrated into a single circuit. Multiple simulations were executed on the tracking device to test if the system operates on its defined operation. One of the largest obstacles the researchers faced during the hardware development stage is the integration of the GSM and GPS modules. By troubleshooting and through continued tests, all the hardware components were successfully integrated.



For RCD:

THIS IS AN EMERGENCY MESSAGE
FROM FLOATS DEVICE
OVERBOARD
PASSENGER LOCATED
HERE:
google.com/maps/place/8.040893,123.766777

RCD

YOUR LOCATION
HAS BEEN
IDENTIFIED

For OTW:

THIS IS AN EMERGENCY MESSAGE
FROM FLOATS DEVICE
OVERBOARD
PASSENGER LOCATED
HERE:
google.com/maps/place/8.040893,123.766777

RESCUE TEAM
IS NOW
DEPLOYED

Figure A.5. OTW Corresponding Display

For FND:

THIS IS AN EMERGENCY MESSAGE
FROM FLOATS DEVICE
OVERBOARD
PASSENGER LOCATED
HERE:
google.com/maps/place/8.040893,123.766777

FND

RESCUE TEAM
HAS FOUND YOU
PLEASE STANDBY

Fig. 3. Web Mapping Service Google Maps.

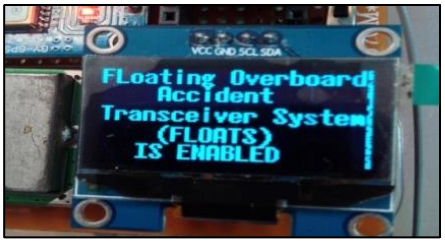
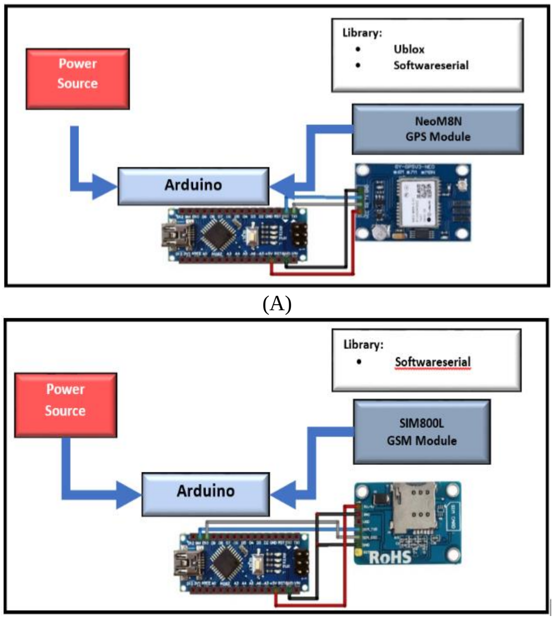


Fig. 4. OLED Display.



(A)

(B)

Fig. 5. (A). Arduino Board and GPS Module Block Diagram, (B). Arduino Board and GSM Module Block Diagram.

In the same manner as the hardware development of the system, gathering data and analyzing it is the foundation of all the processes. To make sure all components are functional, before doing so, the researchers tested the components individually. This includes GSM module, GPS module, OLED, solar panel and the battery [11]. All these components were then connected to the microcontroller board known as Arduino Nano. Before a final code was written, individual codes were developed according to each component of the tracking device. The codes are written, compiled, and uploaded to the Arduino Nano board through the Arduino IDE. The loaded program of each component was uploaded and simulated. At times, the researchers would notice that the device was not working based on its defined function. Each respective code would then be debugged to find and resolve code problems. After testing each code, a final code was written derived from each component's code. The researchers integrated the final code to the system, combining the components as a whole.

D. Integration and Testing of the System

In this phase, the researchers tested the hardware and firmware used in making the device and integrated it as a whole. Troubleshooting, debugging and real-time tests are employed in this phase. Each major component of the Floating Overboard Accident Tracking System (FLOATS) was tested with their respective codes. Troubleshooting and debugging were important processes in achieving the correct operation of the system. Lastly, all of the major components were eventually integrated. The evaluation stage is a multi-cross stage throughout the whole study. The hardware and firmware used for example are evaluated if its necessity is high through previously discussed theories, models and international standards. However, the evaluation stage is also where the final prototype is assessed and recorded based on its functionality, reliability and acceptability.

IV. System Results and Discussion

The importance of Search Theory, the Theory of Planed Behavior and Disaster Preparedness makes the researcher's concerned with the appropriate technology aspect of the study, on such theories helped the researchers how to counter the problem in a conceptual manner. Sweep width, which is the main indicator of Search Theory has been helpful to the researchers by considering its factors. The researchers applied the factors to practical specifications of an object to be used during a search and rescue operation. Table I points some concepts of a sweep width and the actual specification of an object used in the FLOATS.

In this study, the device integrates both floatation and tracking attached to the overboard vessel passenger. In order to cater the different weight classification of passengers, the researchers chose a life jacket that can handle a passenger with weight not exceeding 150 kilograms. A person is perceived as a large object from afar through the help of the automatic inflation feature of the life jacket. Thus, increasing the visual detection of the overboard vessel passenger. In addition to this, the color of the life jacket also improved the probability of detection of a passenger as it contrasts the background of the search area which is the blue sea/ocean water. On the other hand, a night time search and rescue operation requires a search object to have illumination. Therefore, the researchers attached an automatic light to the life jacket. Meanwhile, the reflectivity factor was provided by the Safety of Life at Sea (SOLAS)-approved reflective tape of the life jacket. Since the system used electronic means to locate an overboard passenger, sweep width heavily considers the signal strength which in the study is specified as the signal strength of the mobile communications network used.

TABLE. I. FLOATS Actual Specification

Parts	Device Specification
Life Jacket	a. Type – 150 kg maximum weight b. Size – Inflatable c. Color – Bright Yellow Orange d. Illumination – Automatic Light e. Reflectivity – SOLAS (*Safety of Life at Sea Treaty*) Reflective Strap
Tracking System	a. Signal – GSM (*Global System for Mobile Communications*) b. Location Tracking– GPS (*Global Positioning System*) c. Display – OLED *(Organic Light Emitting Diode)* Screen Display
Supply	a. Type – Lithium Ion Battery b. Solar Panel Type – Mini Polycrystalline Solar Panel

The researchers also considered the behavioral concepts of a passenger bound to be exposed to hazards caused by a maritime incident. The Theory of Planned Behavior stated that attitude, social pressure, and control are the factors that determine an intention to engage in a behavior at a specific time and place [9]. As to the problem concerning this study, the researchers considered the place where a person engages a specific behavior which is in a maritime incident setting. When the FLOATS is implemented during a maritime incident, the attitude of a vessel passenger that is bound to be distressed might change from a state of apprehension to a state of assurance. Social pressure might inflict panic among vessel passengers. However, the system might reduce the level of it.

The study emphasized the development and integration of the tracking device with solar panel to the life jacket. After the components were individually tested and integrated to form the tracking device, the problem on whether the tracking device was possible to be fully-integrated to the life jacket surfaced. Fig. 6 shows the final integration of the prototype.

Since the researchers have used prototyping modules to make the tracking device possible, it was difficult to fully integrate the tracking device to the life jacket. The researchers view the size of these prototyping modules as the main contributory factor to the obstacle of the integration. In addition to this, altering the original design of the life jacket might dispute its original function and damage it in the long run. To deal with this obstacle, the researchers used a carabiner and attached the tracking device to the strap of the life jacket.

The functionality of the FLOATS is evaluated by recording the results throughout the whole system operation specifically the response time of the device. Table II exhibits the response time of the tracking device SMS receive from each trial.

In the overall response time test result, five trials were executed by the researchers. The time was tracked using a smart phone's stopwatch. Through the results shown on the table, each operation time (in seconds) is averaged. It would take an average of 4.698 seconds for the initialization message to be displayed. On the other hand, the SMS sent display would take an average of 9.798 seconds. From the passenger's location, it would take an average of 8.288 seconds to reach the rescuer from the time the SMS sent display is flashed on the tracking device. And, the rescue update sent by a rescuer to the passenger in average would take 3.98 seconds to be displayed on the OLED of the tracking device. The results imply that by using the FLOATS tracking device, in a matter of seconds whenever vessel passengers are forced overboard, rescuers can already receive location data and start tracking the passengers.

This would significantly decrease the time allotted for searches and uplift the feeling of assurance for the vessel passengers. The researchers conducted a real-time test of the tracking device on the surface waters of the Panguil Bay. Two members of the research team were on the field to test the tracking device. Meanwhile, the third member was assigned to gather the location data sent by the tracking device. The third member stays in the Disaster Risk Reduction Management Office (DRRMO) of the Municipality of Tubod, Lanao del Norte. The two members on the field were accompanied by two rescuers from Tubod-DRRMO to guarantee the safety of the testing. In every location point, the boat used by the research team will turn to a halt until 5 location data are obtained. The trial lasted for two hours. The areas included in the testing of the tracking device include the municipalities of Tubod and Baroy in the province of Lanao del Norte and Barangay Silanga of Tangub City. Table III exhibits all the location points gathered during the testing of the tracking device.

Trial 1 of the first location had the location result of 8.078998 and 123.795753. The first set of numbers, specifically 8.078998, indicates the latitude coordinate of the location data. The latter indicates the longitude coordinate. Both are vital for its plotting on a web-based mapping service. By using a web-based mapping service called Google Maps and its counterpart Google Earth, the researchers plotted all the location data resulting to the image (Fig. 7).

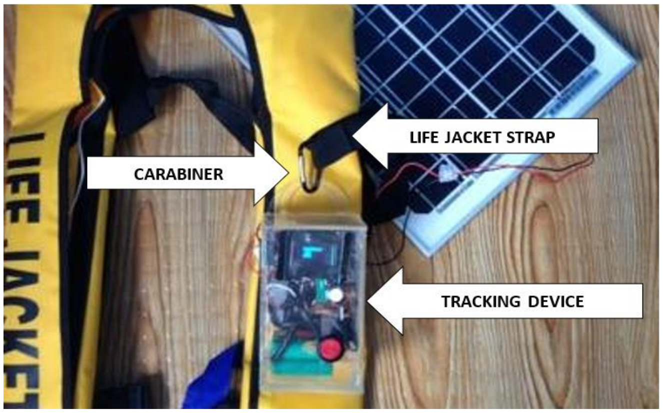


Fig. 6. Integration of the Prototype.

TABLE. II. FUNCTIONALITY OF THE FLOATS

No. of Trials	Initialization Display (sec)	SMS Sent Display	Location Message Receive (✓,×) (Sec)		Rescue Update Display (sec)
Trial 1	4.82	9.58	✓	8.57	5.41
Trial 2	3.44	9.74	✓	7.76	3.22
Trial 3	4.90	10.09	✓	8.38	4.40
Trial 4	4.81	9.82	✓	8.68	3.30
Trial 5	5.52	9.76	✓	8.05	3.57
Average	4.698	9.798	ok	8.288	3.98

TABLE. III. FLOATS TESTING OF THE FLOATS TRACKING DEVICE

Location Area (Latitude, Longitude)	**No. of Trials**				
	Trial 1	*Trial 2*	*Trial 3*	*Trial 4*	*Trial 5*
Tangueguiron, Tubod (*Somewhere Alim Shrine*) 8.080128 , 123.796287	8.078998 , 123.796287	8.078836 , 123.795768	8.078648 , 123.795799	8.078311 , 123.795799	8.078311 , 123.795799
Tubod Port 8.05708 , 123.782714	8.056834 , 123.782714	8.055989 , 123.782196	8.055922 , 123.782096	8.055735 , 123.781997	8.055574 , 123.781898
Sagadan, Tubod (*MCC HOTEL*) 8.048721 , 123.775749	8.047289 , 123.77414	8.045537 , 123.772735	8.04344 , 123.772544	8.045242 , 123.772415	8.044876 , 123.772117
Baroy, Tubod *(Seaside Cuzina Bar)* 8.033973 , 123.758476	8.033945 , 123.75846	8.033936 , 123.758201	8.033894 , 123.758102	8.033854 , 123.757987	8.033809 , 123.757843
Silanga, Tangub City *(Silanga Port)* 8.064160 , 123.77534	8.064497 , 123.775749	8.064688 , 123.776054	8.064755 , 123.776222	8.064462 , 123.776809	8.064395 , 123.777114

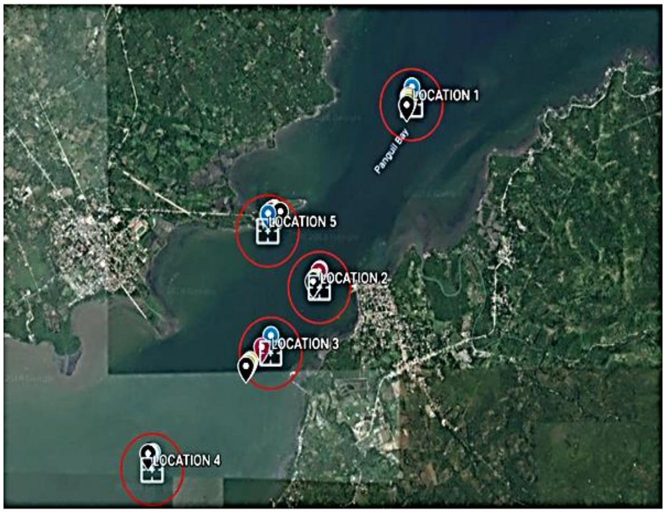


Fig. 7. Google Maps Location Data Result.

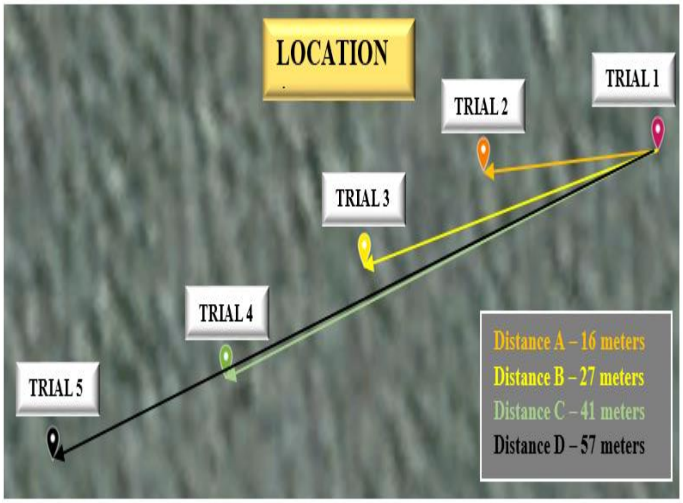


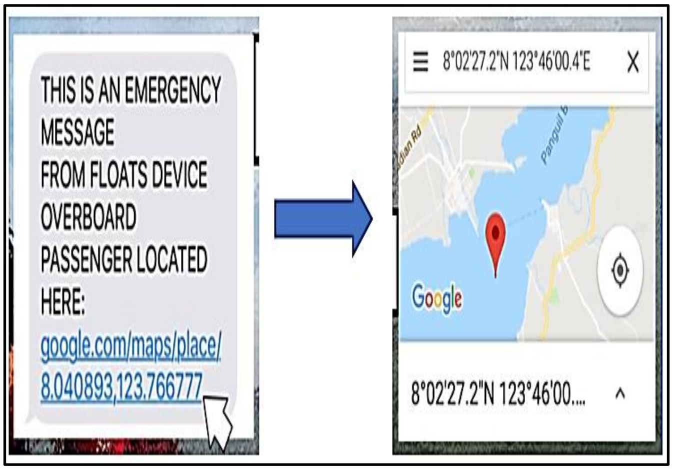
Fig. 8. Location Data Result.

Table IV shows the distance between other four trials (2 to 5) from the reference location to trial 1, using the distance ruler attributes in the Google map application the outcome presented in Fig. 8 location points gathered during the testing of the tracking device.

Fig. 9 shows the actual test floats tracking device receive SMS from the rescuer's smart phone enable the passenger to be located via hyertxt link of google maps from the tracking device.

TABLE. IV. DISTANCE BETWEEN LOCATION FROM TRIAL 1 FLOATS TESTING

Distance Between Location From Trial 1 (meters)	
Distance A *(Trial 1 - Trial 2)*	16 m
Distance B *(Trial 1 - Trial 3)*	27 m
Distance C *(Trial 1 - Trial 4)*	41 m
Distance B *(Trial 1 - Trial 5)*	57 m



THIS IS AN EMERGENCY MESSAGE FROM FLOATS DEVICE OVERBOARD PASSENGER LOCATED HERE:
google.com/maps/place/8.040893,123.766777

8°02'27.2"N 123°46'00.4"E

8°02'27.2"N 123°46'00....

Fig. 9. Actual Test FLOATS Tracking Device.

TABLE. V. ACCEPTABILITY OF THE FLOATS TRACKING DEVICE

Statement	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1. The life jacket is user-friendly.				3	15
2. The device does not interfere with simple movements while the user is in the water.			2	3	10
3. The placement of the device is comfortable while the user is still in the water.				9	9
4. The OLED display is readable			1	4	13
5. The rescue details displayed give the user confidence towards the rescuers.			1	6	11
6. The device is helpful in locating an overboard passenger			2	6	10
7. The device will be helpful in decreasing the number of missing passengers in a case of maritime accident.				2	16
8. The device can be made as an 'essential item' to daily water transport.			1	7	10
9. The application of the device is not only limited to vessel passengers.				7	11
10. Would you recommend this system to be applied in the transportation industry of the country?			1	4	13
TOTAL			8	54	118
Percentage			**3.73%**	**30.34%**	**66.29%**

Shown in Table V is the survey questionnaire result regarding how the general public accepted the device.

In general (Fig. 10), majority of respondents, specifically 66.29% strongly agreed to most of the statements in the survey questionnaire. 30.34% merely agreed and 3.37% were tallied for neutral ratings. Moreover, the most agreed statement in the survey questionnaire is statement number seven, emphasizing the level of confidence the respondents have for the system's potential in minimizing the number of missing persons caused by a maritime incident. The results regarding the survey questionnaire implies that there is a high acceptability of the Floating Overboard Accident Tracking System.

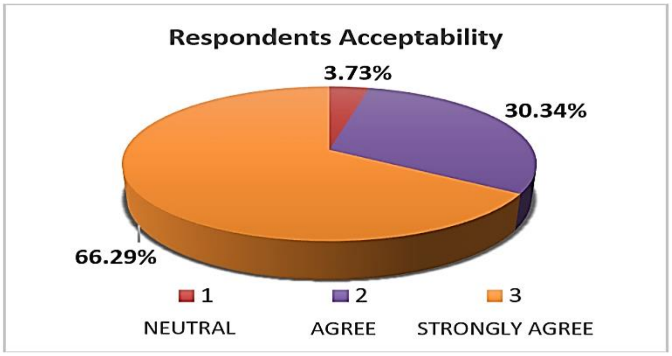


Fig. 10. Respondents Acceptabilty Graph of FLOATS Tracking Device.

V. CONCLUSION

Natural hazards and human error are inevitable factors in the maritime transportation sector of the country. With the increase of passenger traffic yearly, protecting the human resource of the country should be a priority, specifically those who frequently use maritime transport because of economic and efficiency factors. To counter the risks of passenger's exposure to the hazards caused by a maritime incident, the researchers designed and developed the Floating Overboard Accident Tracking System (FLOATS): (a) The concept of making one-sided searches into two-sided searches by means of designing and developing a system that the passengers and rescuers can use simultaneously was made possible. (b) The Theory of Planned Behavior on the other hand gave the researchers idea on how safety devices and tracking devices are perceived by passengers in an actual maritime incidence. (c) The researcher's successfully designed and developed a system that was intended to locate overboard passengers in a maritime incident. (d) Using GSM module, the tracking device can send its first location message to a rescuer's mobile phone in an average time of 8.28 seconds. (e) The rescuer's search and rescue update response can be sent and displayed to the passenger tracking device's OLED display in an average of 3.98 seconds.

This confirmed the system's high response level. Thus, it was designed and developed successfully. The reliability factor on the other hand was indicated by the choice of a mobile network and its signal strength. The acceptability part of survey results show that the tracking system was met with high acceptance with 66.29% strongly agree the use of FLOATS.

This study effectively develops a tracking device with the appropriate technology that can be used by passengers during vessel capsizing at sea. The researchers emphasize the importance of safety in the maritime transportation industry and by applying the system to the current situation of the industry, the risks regarding passenger safety while at sea can be reduced.

ACKNOWLEDGMENT

This Research work is done by the undergraduate student of Bachelor of Science in Electronics and Communication Technology with the advice and support of the faculty of Computer Application Department.

REFERENCES

[1] United Nations Conference on Trade and Development (UNCTAD) (2016). Review of Maritime Transport: Fostering the transparency of maritime markets and analyzing relevant developments. Retrieved May 23, 2018, from https://unctad.org

[2] Lara Richter (2016). The Impact of the Maritime Industry on the Philippine Economy. German-Philippine Chamber of Commerce and Industry, Inc. Makati City 1234, Philippines. Avilable in https://philippinen.ahk.de/fileadmin/AHK_Philippinen/Publications/Maritime_Industry_in_the_Philippines__GPCCI_2016_.pdf

[3] Maritime Industry Authority (MARINA) (August 2018). Annual Report on Basic Maritime Statistics: the available maritime and maritime-related statistical information from the year 2012 up to the year 2016. Retrieved October 8, 2018, from http://marina.gov.ph

[4] United Nations University Institute for Environment and Human Security (UNU-EHS) (2014). The World Risk Report 2014. Available: online. https://i.unu.edu.

[5] Stipe Galic, Zvonimir Lusic, and Ivica Skoko (2014, April 28-29). 6th International Maritime Science Conference (IMSC): The Role and Importance of Safety in Maritime Transportation. Availble in https://bib.irb.hr

[6] Martinez RE, Go JJ and Guevarra J (2016). Epidemiology of drowning deaths in the Philippines, 1980 to 2011. Western Pac Surveill Response J. 2016 Nov 8;7(4). doi:10.5365/wpsar.2016.7.2.005. Available: online. http://ojs.wpro.who.int

[7] Orlando S. Dimailig, Jae-Yong Jeong, and Chol-Seung Kim (June 2011). Marine Transportation in the Philippines: The Maritime Accidents and their Causes. Available: online. https://www.researchgate.net

[8] U.S. Coast Guard Research and Development Center. Review in Search Theory: Advances and Applications to Search and Rescue Decision Support. Report No. CG-D-12-01, Washington ,DC. National Technical Information Service, SpringField, VA 22161. Available: online. https://apps.dtic.mil/dtic/tr/fulltext/u2/a397065.pdf

[9] Icek Ajzen (1991). The theory of planned behavior. Organizational behavior and human decision processes 50 (2), 179-211

[10] Arwa Masoud Hamza El-Nasri (June 2011). Design and Implementations of GPS Mobile Tracking System: Overall Tracking Centre Design. Retrieved December 12, 2018, from http://khartoumspace.uofk.edu

[11] Ralph H. Balingasa , Maria Tricia Camille R. Bilog , Jonnelle Klenn D. Castillo , Jerome M. Perez , Agnes F. Terrible , & Rionel B. Caldo. Distress Signal Tracker Using GPS and SMS Technology: A Prototype. Retrieved December 12, 2018, from http://lpulaguna.edu.ph

[12] Ernesto Empig, Joel Miano, Harreez Villaruz, Nieva Mapula, et.al, 2015, Development of Digital Human Body Tracker Alarm System Using GPS and Transceiver for Catastrophic Events Rescue Operation (DHBT AS), 8th AUN/SEED-Net Regional Conference on Electrical and Electronics Engineering. https://uyr.uy.edu.mm/handle/123456789/371.

Utilizing Feature Selection in Identifying Predicting Factors of Student Retention

January D. Febro
Department of Information Technology
MSU–Iligan Institute of Technology, Iligan City, Philippines

***Abstract*—Student retention is an important issue faced by Philippine higher education institutions. It is a key concern that needs to be addressed for the reason that the knowledge they gain can contribute to the economic and community development of the country aside from financial stability and employability. University databases contain substantial information that can be queried for knowledge discovery that will aid the retention of students. This work aims to analyze factors associated with student's success among first-year students through feature selection. This is a critical step prior to modelling in data mining, as a way to reduce computational process and improve prediction performance. In this work, filter methods are applied on datasets queried from university database. To demonstrate the applicability of this method as a pre-processing step prior to data modelling, predictive model is built using the selected dominant features. The accuracy result jumps to 92.09%. Also, through feature selection technique, it was revealed that post-admission variables are the dominant predictors. Recognizing these factors, the university could improve their intervention programs to help students retain and succeed. This only shows that doing feature selection is an important step that should be done prior to designing any predictive model.**

Keywords—Educational data mining; feature selection; data preprocessing; knowledge discovery; student retention

I. Introduction

Universities have continuously experience challenges in retaining students. Accordingly, about 40% of students in tertiary will not graduate on time [1]. This has been a pressing problem in universities around the world. As 'higher education enrolments have increased in recent decades, dropping out of university has become a common experience' [2]. Like in the Philippines, Commission on Higher Education (CHED) records show that there has been a 4.1 million to 3.6 million total number of dropped out students between academic year 2015-2016 and 2016-2017 [3]. Further, according to the survey, "only 23% of Filipinos finish college" [4]. Undergraduate college enrolments have grown increasingly but with less graduates. Yet, few researchers in Philippine-educational-community have addressed attrition and retention problems.

First-year is regarded in this study considering that it has high attrition rates [5]. It has been affirmed in the study of Garett, Bridgewater and Feinstein [6] that first year is vital in indicating academic success and considered very important at many educational institutions [7]. Thus, the assistance and monitoring of first-year students should be regarded because universities can respond to these students through intervention programs. According to Seidman [8], the "formula for student success is: Retention = Early Identification and Early Intensive Continuous Intervention".

Educational data mining (EDM) can be used to resolve this student retention problem. EDM 'refers to a method for extracting information from large collection of data in educational institutions through data mining (DM) techniques to extract useful knowledge to help decision makers' [12]. Records of students can be queried as an attribute dataset, such as admission test scores and socio-demographic attributes. These can be utilized as predictors for the prediction model for knowledge discovery in databases (KDD).

The two of the three most popular model used in extracting knowledge from data are KDD process model (shown in Fig. 1) and Cross-Industry Standard Process for Data Mining (CRISP-DM) model [9]. Both models contain data preprocessing phase, which is crucial and tedious. In fact, performing the tasks in this phase can consume considerable amount of time. This includes data cleaning, data transformation, and data reduction. An overview of common DM preprocessing steps will be discussed in details in the succeeding section.

However, this paper will only use filter selection feature methods: Correlation Feature Selection, Information Gain Ratio, and Chi-Square analysis. To sought if the results of these selection methods will vary, it will be tabulated and ranked according to feature importance, and will be compared.

This study also aims to cite evidence in support of feature selection method as part of preprocessing step to increase the classification accuracy of a predictive model which has been omitted in some DM studies; like in the following similar studies [10],[11], and [12]. In view of this, two predictive models using classification technique with different feature datasets is proposed– model 1 will used all the dataset attributes queried from the university database and model 2 will used the ranking of important features. Moreover, feature selection method in this study is utilized to identify the possible factors instrumental to student retention and as part of data reduction phase. The significance of this result affects the student and society, along with financial consequences for the institution.

The structure of this paper is as follows: Section 2 reviews some similar works of this study and presents feature selection methods used in this study. Then in Section 3, presents the

The author has been supported by the Commission on Higher Education, Individual Research Grants (CHED M.O. No. 14-2016).

methodology while Section 4 discusses the results. Finally, Section 5 provides conclusion and future work.

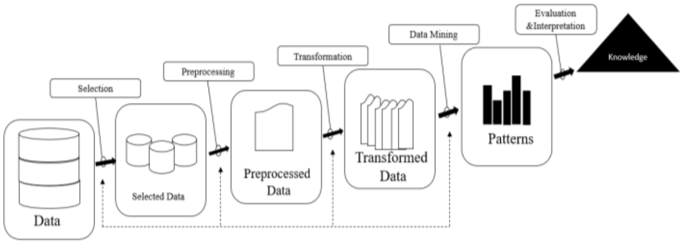


Fig. 1. KDD Process Model [8].

II. Literature Review

Because of the importance of student success in any universities around the world, there were a number of studies in EDM that studied student retention, attrition or academic performance. Most works presented were developing predictive models. The usual conceptual framework for EDM is shown in Fig. 2. Generally, student data were large datasets collected from university databases and normally, data attributes is selected manually.

A. Preprocessing Techniques

Data preprocessing steps is one of the major activities to perform to turn the collected data in an appropriate format for DM algorithms. Its purpose is to remove noise, handle missing data, normalize, select attributes or features, discretize and reduce dimensionality. This section will focus more on the feature selection methods while data cleaning, data transformation and data reduction are discussed briefly below, based on the review of [13].

1) Data cleaning. Row data can have incomplete or irrelevant records that needs to be done. To handle missing data, it can be ignored for large datasets but for small datasets values must be fill manually, fill with global constants or probable value. For noisy data, it can be handled through binning, regression or clustering methods.

2) Data transformation. Data are transformed to suitably fit DM algorithms. This may involve normalization, smoothing, aggregation and generalization.

3) Data reduction. To improve efficiency in large amount of data, data reduction technique is utilized. Certain steps may be performed, these are: data cube aggregation, attribute subset selection or feature selection, numerosity reduction, and dimensionality reduction.

And most frequently, feature selection is disregarded in data preprocessing phase despite that it is a substantial technique for it has been very effective to improve accuracy results. In the following studies presented, the researchers did not perform feature selection as part of data pre-processing step prior to modelling.

In the [11] study, the researchers used the dataset from Prince of Songkla University to predict dropout. They had collected four academic year of student data from Faculty of Science. Tree model and Rule-induction was used and compared in creating the model. The parameters used were pre-academic data and GPA. The model JRip rule induction has the highest accuracy result of 77.30%. Data transformation and Data cleaning were the only pre-processing steps made.

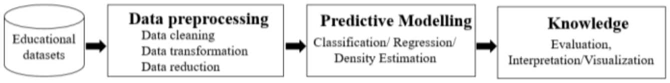


Fig. 2. A Typical EDM Conceptual Framework.

In the study of [12], researcher collected an 8-year-period record that contains student demographics, family background information and academic records for predicting academic performance of the Computer Science students. They used DT, Naïve Bayes and Rule based classification to create a predictive model. The best model in their study was Rule Based with 71.3% accuracy value. In their study, preprocessing steps performed were Data Selection, Data Cleaning and Normalization. The selected nine parameters on the study were based on their literature review.

In the study of [13], the researchers used regression in analyzing the academic performance using the academic subject data of the graduating students of Computer Science students of New Era University and in calculating accuracy they used Mean Absolute Percentage Error. In their study, they performed Data Selection, Data Cleansing and Data Transformation as a preprocessing step. The factors selected were the course subjects of students and the GPA.

B. Feature Selection Methods (FSM)

FSM has been proven to reduce dimensionality, remove noise and unimportant data from thus improving accuracy result [14]. Moreover, feature selection is substantial for data mining algorithms for many reasons such as generalization performance, running time requirements and constraints. There are three types exists for machine learning these are, filters, wrappers and embedded. The difference between filter method and wrapper methods is that the first calculates the number of features based on the common features of the data utilizing heuristics while the latter evaluate the number of features employing the learning algorithm [15]. Embedded methods on the other hand searches the best subset of parameters that is that is embedded in the classifier construction [16].

As a pre-processing step to DM, filter method has an advantage, for example filters execute faster than wrappers and it does not need re-execution on different learning algorithms. Hence, this study is focused on the filter method specifically Correlation Features Selection, Chi-Square Analysis, and Gain Ratio.

C. Correlation-based Feature Selection (CFS)

CFS, "ranks feature subsets as per correlation based heuristic evaluation function in which numeric features are first discretized to gauge correlation between nominal features". This algorithm seeks for features that are particularly correlated with the explicit class [15]. CFS formula is given in equation (1):

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \qquad (1)$$

where r_{zc} in the equation is the interrelatedness of scored features, k is the sum of features, r_{zi} is the mean of the correlations relating to the class variable, and is the mean of inter-correlation between features [15].

D. Chi-Squared

Chi-squared is a frequently used method for feature assessment by calculating the value of chi-squared statistic with regard to the class [17]. The formula is provided below.

$$X^2 = \frac{\sum (o-e)^2}{e} \tag{2}$$

where o is observed frequencies and e is expected frequencies. This method is used to identify whether a distribution of observed frequencies varies from the supposed expected frequencies.

E. Informaton Gain Ratio (IGR)

IGR method computes the importance of the features using information gain and give weights to them accordingly even if it applied to features that have dissimilar value using the equation below [18].

$$GR(att) = \frac{IG(att)}{H(att)} \tag{3}$$

where equation (4)

$$H(att) = \sum_j - P(v_j) \log_2 P(v_j) \tag{4}$$

where $P(v_j)$ corresponds to the chances of having v_j by providing general values for an attribute j.

III. Methodology

Fig. 3 illustrates the activity in this study. In the data pre-processing feature selection is emphasized.

A. Dataset Collected

The records used in this study were real records of five academic years queried from a university database. These records contain information about the entrance result, grades, and among others. The data for this research was inputted in a data mining tool. The dataset is comprised of 7, 936 records with 29 features.

The potential predictor variables queried fall into two categories: pre-college data and post-admission data. Pre-college data are records prior to admission, it includes admission test scores and socio- demographic attributes. The pre-college dataset features examined in this study is grouped into two: demographic and socio-economic (gender, blood type, skills, sports, musical instrument, province of origin, parents educational background, parent’s income, parent’s tribe, religion, number of brothers, number of sisters and rank in family) and academic potential (admission test score in Math, Language Usage, Aptitude, and Science). On the other hand, post-admission data are educational achievement indicators such as course, scholarship status, grades in Math and English subjects, and grade point average of first semester.

B. Data Pre-Processing

Prior to modelling and to improve the input data quality and suitability, data pre-processing is needed. For this study, identifying noise data, missing values, irrelevant and redundant data and removing outliers were crucial steps. Data cleaning is done using a data mining tool.

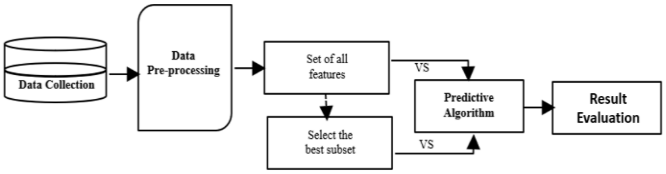


Fig. 3. Methodology Flow.

To remove the irrelevant data and noise from the dataset, the following steps were carried out.

1) Load data collected
2) Integrate collected data
3) Filter data by removing missing records
4) Remove duplicates
5) Do normalization
6) Detect Outlier

Careful data integration is done to reduce and avoid redundancies and inconsistencies. Redundant data were carefully examined; same attributes were not included in this study.

Data cleansing steps is performed to remove the incomplete data. A list-wise deletion method is adopted to delete the entire record from the analysis if any variable in the model has a missing value. Missing data is ignored to avoid adding bias and distortion to the dataset. Removing a few records will not impede the results of the model since this study contains large dataset. Finally, to handle outliers, local outlier factor (LOF) is executed.

C. Feature Selection

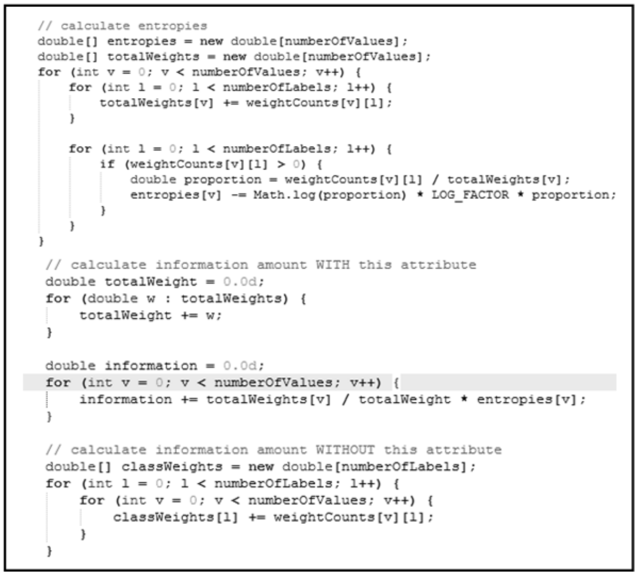
One of the main goals in this study is to identify what dominant variable or combination of variables collected can be used as predictors of first year student success. In this study, filter model using feature rankings are used, namely, the Info Gain Ratio, the Correlation Feature Selection, and the Chi Square, to identify the dominant variables. The significance of using filter model method is that it separates feature selection from learning [19]. Thus, no bias towards any learning algorithm.

During the feature selection process, no specific form of relationship is assumed. The outcome of the feature selection is list of predictors ranked according to their importance.

1) Information Gain Ratio (IGR): The first FSM employed is the IGR. It calculates the entropies in class and resolves the vulnerability of IG. Fig. 4 shows the code snippet of the method used in this study.

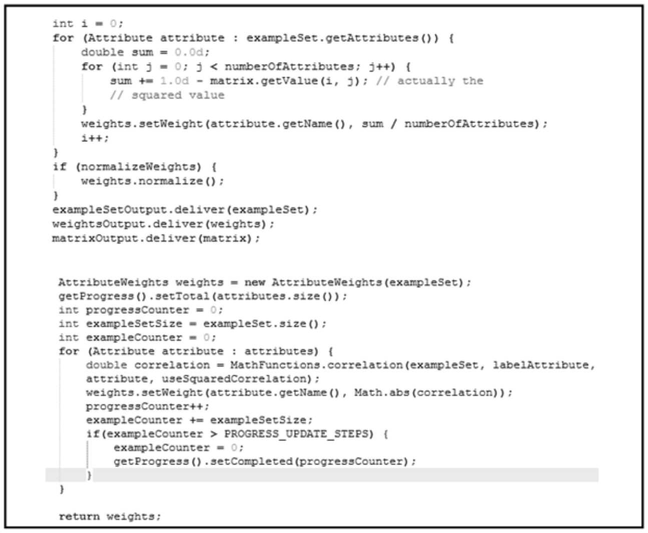
2) Correlation feature selection: CFS finds attribute that are highly related with the specific groups but still have at least inter-correlation amongst the attributes themselves. Fig. 5 shows the code snippet of the method used in this study.

3) *Chi-Square:* The chi-square statistics use nominal data and is utilizes to identify if a distribution of the stated frequencies varies from the actual expected frequencies. Fig. 6 shows the code snippet of the method used in this study.



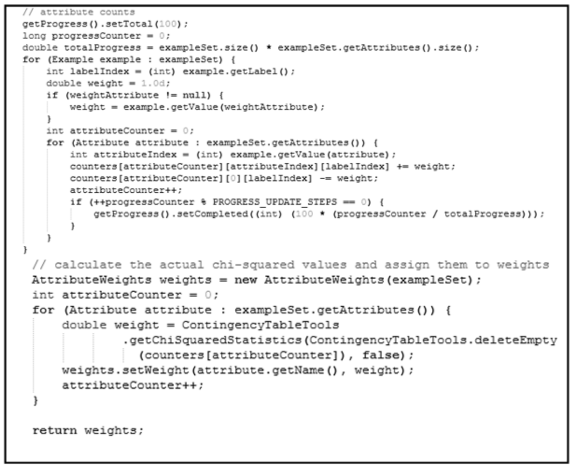
```
// calculate entropies
double[] entropies = new double[numberOfValues];
double[] totalWeights = new double[numberOfValues];
for (int v = 0; v < numberOfValues; v++) {
    for (int l = 0; l < numberOfLabels; l++) {
        totalWeights[v] += weightCounts[v][l];
    }

    for (int l = 0; l < numberOfLabels; l++) {
        if (weightCounts[v][l] > 0) {
            double proportion = weightCounts[v][l] / totalWeights[v];
            entropies[v] -= Math.log(proportion) * LOG_FACTOR * proportion;
        }
    }
}
 // calculate information amount WITH this attribute
 double totalWeight = 0.0d;
 for (double w : totalWeights) {
     totalWeight += w;
 }

 double information = 0.0d;
 for (int v = 0; v < numberOfValues; v++) {
     information += totalWeights[v] / totalWeight * entropies[v];
 }

 // calculate information amount WITHOUT this attribute
 double[] classWeights = new double[numberOfLabels];
 for (int l = 0; l < numberOfLabels; l++) {
     for (int v = 0; v < numberOfValues; v++) {
         classWeights[l] += weightCounts[v][l];
     }
 }
```

Fig. 4. IGR Code Snippet.



```
int i = 0;
for (Attribute attribute : exampleSet.getAttributes()) {
    double sum = 0.0d;
    for (int j = 0; j < numberOfAttributes; j++) {
        sum += 1.0d - matrix.getValue(i, j); // actually the
        // squared value
    }
    weights.setWeight(attribute.getName(), sum / numberOfAttributes);
    i++;
}
if (normalizeWeights) {
    weights.normalize();
}
exampleSetOutput.deliver(exampleSet);
weightsOutput.deliver(weights);
matrixOutput.deliver(matrix);


 AttributeWeights weights = new AttributeWeights(exampleSet);
 getProgress().setTotal(attributes.size());
 int progressCounter = 0;
 int exampleSetSize = exampleSet.size();
 int exampleCounter = 0;
 for (Attribute attribute : attributes) {
     double correlation = MathFunctions.correlation(exampleSet, labelAttribute,
     attribute, useSquaredCorrelation);
     weights.setWeight(attribute.getName(), Math.abs(correlation));
     progressCounter++;
     exampleCounter += exampleSetSize;
     if(exampleCounter > PROGRESS_UPDATE_STEPS) {
         exampleCounter = 0;
         getProgress().setCompleted(progressCounter);
     }
 }

 return weights;
```

Fig. 5. CFS Code Snippet.



```
// attribute counts
getProgress().setTotal(100);
long progressCounter = 0;
double totalProgress = exampleSet.size() * exampleSet.getAttributes().size();
for (Example example : exampleSet) {
    int labelIndex = (int) example.getLabel();
    double weight = 1.0d;
    if (weightAttribute != null) {
        weight = example.getValue(weightAttribute);
    }
    int attributeCounter = 0;
    for (Attribute attribute : exampleSet.getAttributes()) {
        int attributeIndex = (int) example.getValue(attribute);
        counters[attributeCounter][attributeIndex][labelIndex] += weight;
        counters[attributeCounter][0][labelIndex] -= weight;
        attributeCounter++;
        if (++progressCounter % PROGRESS_UPDATE_STEPS == 0) {
            getProgress().setCompleted((int) (100 * (progressCounter / totalProgress)));
        }
    }
}
 // calculate the actual chi-squared values and assign them to weights
 AttributeWeights weights = new AttributeWeights(exampleSet);
 int attributeCounter = 0;
 for (Attribute attribute : exampleSet.getAttributes()) {
     double weight = ContingencyTableTools
             .getChiSquaredStatistics(ContingencyTableTools.deleteEmpty
               (counters[attributeCounter]), false);
     weights.setWeight(attribute.getName(), weight);
     attributeCounter++;
 }

 return weights;
```

Fig. 6. Chi-Square Code Snippet.

D. Data Modelling

A prediction model for EDM can be developed using EDM techniques but will heavily depend on the type of datasets. In this study, logistic regression method is used.

The dataset is partitioned into training and validation subsets. Two predictive models were created, for the first model all the features will be inputted. On the second model, only the significant variables assessed by feature selection techniques were the final parameters in creating the model. 70% of the dataset is used in training and the remaining 30% is used as a test-set for both models and are tested for accuracy using 10-fold cross-validation.

E. Result Evaluation

The performance of the two models is evaluated by its accuracy and precision which are computed using the equation below.

$$Accuracy = \frac{\text{True Positive+True Negative}}{\text{True Positive+True Negative+False Positive+False Negative}} \tag{5}$$

The accuracy is computed by the actual instance of correct classification (True Positive + True Negative) over the total instances of that class.

$$Precision = \frac{\text{True Positive}}{\text{True Positive+False Positive}} \tag{6}$$

Precision is computed by the positive predicted instances over the total predicted instances.

IV. Results and Discussions

A. Results of Feature Selection

Fig. 7 shows the result of IGR. The result is based on the upmost Gain ratio ranked by their importance. Any information gain above zero shows some type of significance. Factors like English status, Math status, family income and college entrance score for language usage, math, aptitude and science largely influence the result of student's retention.

Fig. 8 shows the features and Correlation-based Feature Selection scores ranked in ascending order of importance.

Among the highest ranked by CFS are English status, gross income and math status.

Fig. 9 results show features that were highly influential or with high chi-square values. These values are displayed in ascending order.

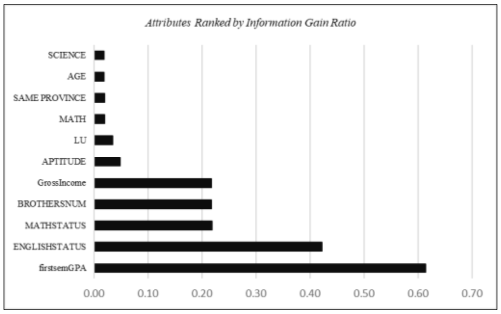


Fig. 7. Attributes Ranked by Information Gain Ratio.

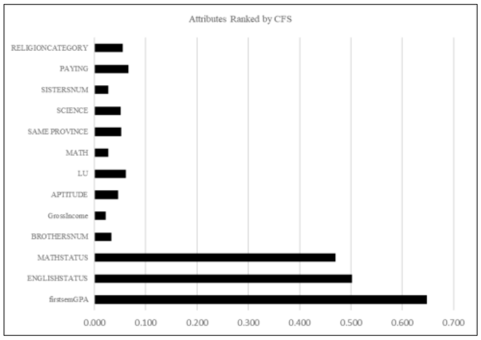


Fig. 8. Attributes Ranked by CFS.

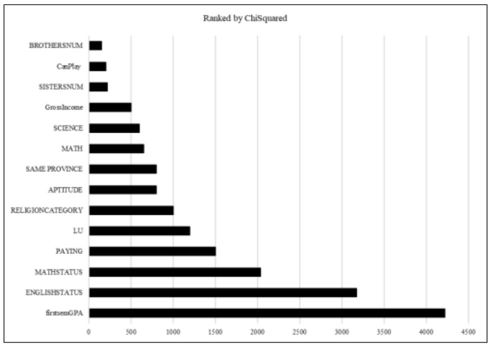


Fig. 9. Attributes Ranked by ChiSquared.

From the results, it can be concluded what dominant predictors affects student retention. These predictors in Table I are utilized in building a predictive model. From 29 possible predictor variables, only 14 predictor variables are used in the second model.

B. *Modelling*

In this study, logistic regression is used. As mention, the Model 1 used all 29 predictor variables and the Model 2 used only 14 predictor variables which were ranked by the filter model during the selection feature analysis. Both models tested for accuracy using 10-fold cross-validation.

TABLE. I. RANKED OF PREDICTOR VARIABLES IN ASCENDING ORDER

CFS	Info Gain Ratio	ChiSquared
firstsemGPA	firstsemGPA	firstsemGPA
ENGLISHSTATUS	ENGLISHSTATUS	ENGLISHSTATUS
MATHSTATUS	MATHSTATUS	MATHSTATUS
PAYING	BROTHERSNUM	PAYING
LU	GrossIncome	LU
RELIGIONCATEGOR Y	APTITUDE	RELIGIONCATEGOR Y
SAME PROVINCE	LU	SAME PROVINCE
SCIENCE	MATH	APTITUDE
APTITUDE	SAME PROVINCE	SCIENCE
BROTHERSNUM	SCIENCE	MATH
MATH	SISTERSNUM	GrossIncome
SISTERSNUM	RELIGIONCATEGOR Y	CanPlay
GrossIncome	PAYING	SISTERSNUM
CanPlay	CanPlay	BROTHERSNUM

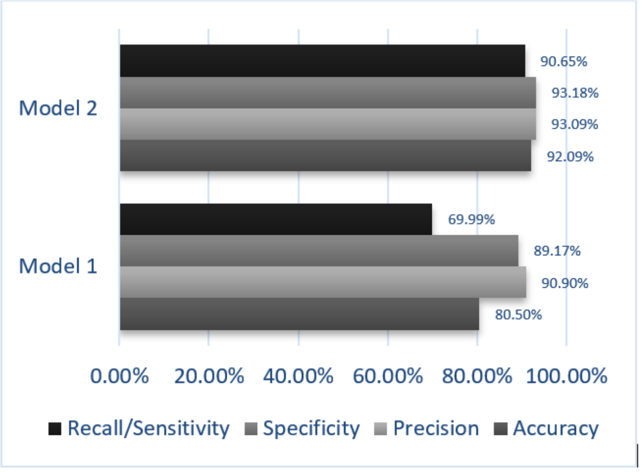


Fig. 10. Result Comparison of the Two Models.

Fig. 10 shows the results of the two predictive models. The results presented, indicates that the accuracy result of model 2 jumps to 92.09% from 80.50%. This only tells that doing feature selection is vital as part of preprocessing.

V. CONCLUSION AND FUTURE WORK

Early detection of potential student leavers is favorable for both students and institutions. In this paper, 14 features from 29 predictor variables have identified to have importance by performing filter model FSM. Based on the feature selection result, it was found that aside from first semester gap, students retaining in university was positively correlated with the following predictors, namely, college entrance exam score (math, language usage, aptitude and science category), number of siblings, family income, English grade, and math grade. The generated information will be quite useful for the university management to develop policies and strategies for better planning and implementation to increase the retention rate in HEIs.

In future, the study can be enhanced by applying few hybrid feature selection algorithms on student datasets in order to predict student retention. A web-based system will be developed that helps to monitor students and accurately predict student retention and attrition.

ACKNOWLEDGMENT

The author would like to thank the Commission on Higher Education (CHED) for supporting this study under the Individual Research Grants (CHED M.O. No. 14-2016).

REFERENCES

[1] National Center for Education Statistics, "The Condition of Education 2016." (NCES 2016-144), Undergraduate Retention and Graduation Rates, 2016.

[2] A. Norton, and I. Cherastidtham, "Dropping out: the benefits and cost of trying university", Grattan Institute, 2018.

[3] Commission on Higher Education (CHED), "2018 Higher Education Facts and Figures," 2018.

[4] Philippine News Agency, "Only 23% of Filipinos finish college," BusinessMirror, (April 27, 2017).

[5] Australian Government Department of Education and Training, "Improving retention completion and success in higher education," Higher Education Standards Panel Discussion Paper, June 2017.

[6] N. Garett, M. Bridgewater and B. Feinstein, "How Student Performance in First-Year Composition Predicts Retention and Overall Student Success," Retention, Persistence, and Writing Programs, Louisville, CO: University Press of Colorado, 2017.

[7] P. Van der Zanden, E. Denessen, A. Cillesen and P. Meijer, "Domains and predictors of first-year student success: A systematic review," Educational Research Review, 23 57-77, 2018.

[8] A. Seidman, "College student retention: formula for student success," Westport, CT: ACE/Praeger, 2005.

[9] U. Shafique, and H. Qaiser, "A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)," International Journal of Innovation and Scientific Research, Vol. 12 No. 1 Nov. 2014.

[10] P. Ramya, K. Gudlavalleru and M. Kumar, "Student Performance Analysis Using Educational Data Mining," International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, 2016.

[11] J. Pattanaphanchai, K. Leelerpanyakul, & N. Theppalak, "The Investigation of Student Dropout Prediction Model in Thai Higher Education Using Educational Data Mining: A Case Study of Faculty of Science, Prince of Songkla University," Journal of University of Babylon for Pure and Applied Sciences, Vol.(27), No.(1): 2019.

[12] F. Ahmad, N. Ismail, and A. Aziz, "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques," Applied Mathematical Sciences, vol. 9, no. 129, pp. 6415-6426, 2015.

[13] W. Bhaya, "Review of Data Preprocessing Techniques in Datamining," Journal of Engineering and Applied Sciences, 12 (16): 4102-4107, 2017.

[14] A. Algarni, "Data Mining in Education," (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016.

[15] M. Paraiso, H. Torres, et al., "Data Mining Approach for Analyzing Graduating Students' Academic Performance of New Era University – Bachelor Science in Computer Science". International Journal of Conceptions on Computing and Information Technology. Vol. 3. Issue 3, 2015.

[16] M. Hall, and L. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper," Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, 1999.

[17] Sheena, K. Kumar, and G. Kumar, "Analysis of Feature Selection Techniques: A Data Mining Approach," International Journal of Computer Applications, 2016.

[18] J. Novaković, P. Strbac, and D. Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," Yugoslav Journal of Operations Research 21, 2011.

[19] M. Trabelsi, N. Meddouri, and M. Maddouri, "A New Feature Selection Method for Nominal Classifier based on Formal Concept Analysis," Procedia Computer Science, 2017.

An Enhanced Deep Learning Approach in Forecasting Banana Harvest Yields

Mariannie A Rebortera[1]
Graduate Programs
Technological Institute of the Philippines
Quezon City, Philippines

Arnel C Fajardo[2]
Dean, School of Graduate Studies
Manuel L. Quezon University
Diliman, Quezon City Philippines

***Abstract*—This technical quest aspired to build deep multifaceted system proficient in forecasting banana harvest yields essential for extensive planning for a sustainable production in the agriculture sector. Recently, deep-learning (DL) approach has been used as a new alternative model in forecasting. In this paper, the enhanced DL approach incorporates multiple long short term memory (LSTM) layers employed with multiple neurons in each layer, fully trained and built a state for forecasting. The enhanced model used the banana harvest yield data from agrarian reform beneficiary (ARB) cooperative of Dapco in Davao del Norte, Philippines. The model parameters such as *epoch, batch size* and *neurons* underwent tuning to identify its optimal values to be used in the experiments. Additionally, the root-mean-squared error (RMSE) is used to evaluate the performance of the model. Using the same set of training and testing data, experiment exhibits that the enhanced model achieved the optimal result of 34.805 in terms of RMSE. This means that the enhanced model outperforms the single and multiple LSTM layer with 43.5 percent and 44.95 percent reduction in error rates, respectively. Since there is no proof that LSTM recurrent neutral network has been used with the same agricultural problem domain, therefore, there is no standard available with regards to the level of error reduction in the forecast. Moreover, investigating the performance of the model using diverse datasets specifically with multiple input features (multivariate) is suggested for exploration. Furthermore, extending and embedding this approach to a web-based along with a handy application is the future plan for the benefit of the medium scale banana growers of the region for efficient and effective decision making and advance planning.**

Keywords—Yield forecasting; Deep Learning; Long short-term memory; Banana harvest yield forecasting

I. INTRODUCTION

Deep learning (DL) is a method that has been enticing attention in recent years of machine learning and its continuous growth gains more popular among researchers in diverse disciplines [1] where advancement and progression are fast and incremental. Frequently, development takes place in a well-resourced area (e.g. medical, security) and this budding application of DL is dispersed to the agricultural sector [2]. In agriculture, production is an essential phenomenon in natural aspect where progressions fuse in multifaceted ways and production patterns are specifically influence by market factors and relentlessly affected by extreme events (i.e. floods, droughts) and revealed to deteriorate and decreases its yield. Thus, management technologies and timely interventions should be put in place. Otherwise, it would remain unmitigated or even intensified and could suffer shortfalls and will continue to exert pressure on agricultural produce.

Yields from crops play a noteworthy role in the economic progression. Among the major fruit crops, banana (*Musa sp*.) is one of the important tropical fruit crops and part of the rising economy of many developing countries like Philippines and the world’s most important goods following rice, wheat and maize. Hence, yield assessment of banana production is essential for policy decisions regarding procurement, distribution, buffer stocking, import-export, price-fixation and marketing [3]. In view of that, more accurate forecasts of the harvest yields and crop production provides an aid to an effective and efficient decision making using timely information. It is a significant phase for an emerging economy so that adequate planning is undertaken for sustainable growth [4] and for the overall development of the country. However, studies have shown that agricultural problems like forecasting yields remain difficult due to the lack of the necessary infrastructures and there is no proof of optimal model to handle time series (TS) data to be used in forecasting such as the banana harvest yields dataset.

Previously, different conventional models such as autoregressive integrated moving average (ARIMA) [5], [6], [7], [8] are used in forecasting more specifically using TS data. However, a major drawback in its used in forecasting is its incapability to recognized nonlinearities [9], [10], [4]. These classical techniques have been replaced by DL algorithms [11]. DL approaches are capable of identifying non-linearity and complexity of data [12], [8] in TS forecasting. Hence, this advance approach is referring to as a future promising tool [13] in forecasting yields in the field of agriculture.

The more recent structures of DL are Deep Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks (RNN), and Q-learning [11]. Among these DL architecture, RNN model presents elevated performance in prediction, as it can capture the time features and the architectures demonstrate dynamic temporal behavior [14]. However, the training is difficult and the major weakness of RNN is carried out during the requirement of learning long-range time dependencies [15]. This limitation is addressed through the development of LSTM algorithm [16], [17], [18]. LSTM is developed to seize the “vanishing gradient problem” encountered in RNN. It is also capable of learning long term dependences [19], [10]. It is the state-of-the-art technique for sequence learning [20], [21], [16] and TS prediction such as in financial market [22], hydrology [19], petroleum production [23], energy [24], [25], [26], [27], neurocomputing [28], [24], [18], [17], expert systems [16], internet of things [10]. However, the weakness of this technique has perceived as it did not perform satisfactorily in dealing with TS forecasting. Its shallow architecture makes it incapable to exemplify the complex characteristics of TS data more specifically in handling extremely nonlinear and long interval TS datasets [15] such as in banana harvest yields data. Furthermore, this limitation compels LSTM to be unclear if it is the best design to work out real problem especially in using the harvest yield dataset and the optimization issues due to the size of the data and the model tuning strategy applied. Also, LSTM has not been used in forecasting harvest yields such as in bananas since its inception.

In this paper, the enhanced deep learning-based approach is used in forecasting harvest yields of banana production. The performance of the model is then evaluated in terms of accuracy measures. The result of this study will be a great contribution to the consistent management for the improvement of harvest yields and to the overall production. It would also provide a new technique to assist the agrarian reform beneficiary (ARB) cooperative of Dapco in its individual farming scheme, decision-making process and advance planning.

The following parts of the paper are structured as follows: Section II looks into the idea of Deep Learning, time series forecasting, application, and challenges. Section III features the fundamentals of LSTM and the enhanced model. Section IV highlights the experiments as well as the outcomes and finally, suggestions and conclusions are offered in Section V.

II. Related Literature

A. Deep Learning Approach

DL establishes a current, modern technique for data analysis, with likely results and significant capability [29]. As DL has been effectively applied in several domains, it has recently entered also in agriculture. Moreover, the performance of the model is generally high and its potentials can also be applied in a wide variety of agricultural related problems not only involving images but also in forecasting TS data. The efficiency of its testing time is relatively faster than any other outmoded methods and the likelihood to develop simulated datasets to train model [13] is another advantage to solve real-world problems.

DL algorithm consists of various components like RNN [11] and is the most popular method in forecasting task based on the solution intricacy, the desired accuracy in prediction and features of data [15]. More approaches are adopting RNN particularly LSTM exploiting the time dimension to perform higher performance prediction [13] and more sophisticated architecture in dealing with large datasets which could improve its performance [30], [31].

B. Time Series Forecasting Applications, Challenges and Methods

Investigating time series and dynamic demonstration is an interesting exploration. Analysis of TS data aims to research the observations trail and construct a model to depict the structure of data and forecast future values. Hence, it is vital to conceptualize an applicable model aiming at improving accuracy of the forecast. There are some different domains which already tested the capacity and adopted the used of LSTM in TS forecasting problems such as predicting emergency event occurrence [16] which solves classification and regression problem and exhibits better performance which is proven effective over conventional methods. In forecasting petroleum production, [15] which case study involves the production of two separate oil depot at a particular time period, proves the capability and eligibility of LSTM to be applied in the nonlinear forecasting problems and outperforms outmoded and traditional mathematical forecasting models. In predicting traffic flow [24], where it involves short term traffic flow at a time interval of 1 to 5 minutes, LSTM and Attention Mechanism shows excellent performance in dealing with 5- or 1-minute-long historical data. However, the performance of the model declines swiftly as the length of the sequence increases because of the collection of errors: the longer the sequence in LSTM, the greater the error. Thus, it considers time and space features combination. Otherwise, in predicting the remaining useful life of proton exchange membrane fuel cell [18], LSTM model quickly and accurately predicts the remaining service life and suitable for online residual life prediction but the robustness and generalizations performance need to be further strengthened and improved. It also outperforms the outmoded techniques in predicting hourly day-ahead solar irradiance [25] though error is encountered when using weather forecast but it shows less overfitting and better generalization. Thus, further evaluation and assessment is needed.

Moreover, it also shows simplicity and effectivity than ARIMA and back propagation neural network in the tourism flow prediction [10]. It suggests further that more hidden states are to be tested and superiority over the classic feed forward neural network and the double-LSTM models in predicting water table depth [19]. A dropout strategy is being implemented to avoid overfitting though it has a resilience to learn TS data but has insufficiency in its fitting ability. In financial market predictions, it is effective in extracting meaningful information from a noisy financial TS [22] compared to random forest, standard deep nets and logistic regression and it turns out to be an advancement of the domain with respect to prediction accuracy and daily returns after transaction cost. It also outperforms conventional techniques in short-term load forecasts [26] and exhibits consistency for the ambiguity with snowball in hours to forecast. LSTM proves appropriateness in TS modelling and forecasting with reduction in percentage errors and shows simplicity towards managing the information instead of working over complex equations.

III. The Enhanced Approach

This segment depicts the enhanced deep learning approach used in forecasting banana harvest yields from data preparation to evaluating the forecast. Also, important steps are discussed and described in this section.

A. *Data*

The banana harvest yield data set features the number of stems cut, the number of boxes produce and box-stem ratio. The author chose the “number of stems cut” as the harvest yields time series data to be used during the experiment. In this study, the term “stems” is referring to the whole bunch of the banana fruit cut from the plant. The harvest data came from the ARB cooperative of Dapco in Davao del Norte, Philippines. The cooperative is one of the key players from the small-medium scale banana growers in the region and has contributed to the overall production on the exportation of bananas to other countries like Japan. The dataset contains thirty-five thousand series of observations approximately from year 2014 to 2018 where each year is composed of thirteen (13) periods and it usually starts from second half of the first month of the year to the first half of the following year.

B. *Data Preprocessing*

Series of data transformations are done before fitting the model to the dataset and making a forecast. It includes converting the TS data into supervised learning to make it stationary. A lag differencing is used to strip off the increasing trend in the data. Transformation of data into a supervised learning problem and scaling to values to meet the hyperbolic tangent activation function of the model is also done. All these transforms will be inverted back on forecast to revert data into its original scale before evaluation and in determining the error score and splitting data into training and testing sets respectively. Aiming to obtain the best outcomes, iterative optimization is used which means attaining the outcomes several times and select the utmost optimal iteration that has less errors. Each important parameter, such as epoch, batch size and neurons, is given a varied value and the experiment for each parameter are run in several repetitions as desired. The best value for each parameter is identified through the summary of performance report using the RMSE scores from each population of results.

C. *The Multi-Dynamic Long Short Term Memory (mdLSTM) Model*

It is indispensable to briefly explain the fundamentals of LSTM memory block as it is the precedent of the enhanced model prior to its introduction.

1) Fundamentals of the Long Short Term Memory (LSTM) Model.

LSTM was first proposed in 1997 (Sepp Hochreiter, Jürgen Schmidhuber), driven by an analysis of error flow in prevailing RNNs (Hochreiter et al., 2001) [16]. It shows suitability for processing and forecasting using TS data. The LSTM block, shown in Fig. 1, depicts a cell state (C_t) which resembles a conveyor belt. It thoroughly takes turn in the chain and controlled by constitution of gates; an elective inlet means of information. Gates are comprised of a sigmoid neural net layer and a *pointwise multiplication* and *addition* operation. These gates and the memory cell allow an LSTM unit to respectively *forget*, *memorize* and *expose* [14] the memory content.

It has an input at time step *t* denoted by (x_t), and the hidden state from the previous time step (S_{t-1}) that is introduced to LSTM block, and then the hidden state (S_t) is computed through the forget gate (f_t), input gate (i_t) and output gate (o_t) where the *input and forget gates* are responsible of how much new content should be remembered (*memorized*) and how much old content should be disregarded (*forgotten*). Gates are computed using the following set of formulas arrange in steps:

Step 1: Determine what content is going to be disregarded from the *cell state* which will be decided by the *forget gate* (f_t):

$$f_t\text{=sigmoid}(W_{xf}x_t+W_{sf}h_{t-1}+b_f) \quad (1)$$

Step 2: Determine which new content is going to be kept in the *cell state* which will be decided in two phases: First, the *input gate*(i_t) layer decides which values to be updated. Second, a *tanh* layer that forms a vector of new candidate values ($\hat{C}_t$). These two phases can be illustrated as follows:

$$i_t\text{=sigmoid}(W_{xi}x_t+W_{si}h_{t-1}+b_i) \quad (2)$$

$$\hat{C}_t = \tanh\,(W_{ci}x_t+W_{ci}h_{t-1}+b_c) \quad (3)$$

Step 3: Update the previous *cell state* (C_{t-1}) into the new *cell state* ($\hat{C}_t$), which can be conveyed as:

$$C_t = f_t * C_{t-1}+i_t * \hat{C}_t \quad (4)$$

Step 4: Lastly, determine the desired output to be produced. The output will be a streamed form and will be based on the cell state. The output gate (o_t) in this step will decide what part of the cell state is going to be produced as output and then goes through the *tanh* layer, impelling values to be between -1 and 1, and multiplying it to the *output gate* as illustrated in the equation below:

$$o_t\text{=sigmoid}(W_{xo}x_t+W_{ho}h_{t-1}+b_o) \quad (5)$$

$$S_t=o_t*\tanh(C_t) \quad (6)$$

The LSTM is represented with the two sets of parameters from the preceding six equations. These are: W_{xf}, W_{sf}, W_{xi}, W_{si}, W_{ci}, W_{xo}, W_{ho} which are referred to as weights and b_f, b_i, b_c *and* b_o are biases, respectively.

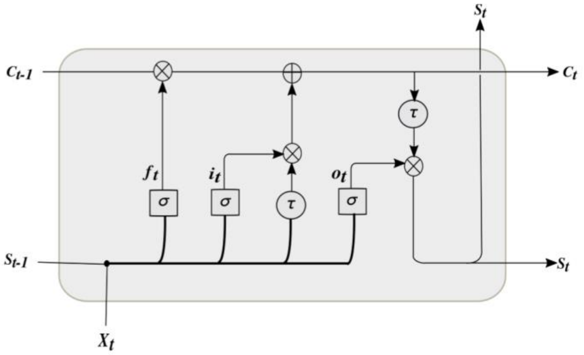


Fig. 1. The LSTM Memory Block, where ⊕ is the Pointwise Addition, ⊗ is the Pointwise Multiplication, (τ) Hyperbolic Tangent Activation Function and [σ] is the Sigmoid Activation.

2) The multi-dynamic Long Short-Term Memory (mdLSTM)

An effective means to a better overall performance of the neural network is to augment its deepness [11]. The development of expound LSTM recurrent neural network is encouraged by the remarkable learning capabilities of profound recurrent network design to be utilized in TS forecasting applications. The enhanced model, shown in Fig. 2, has several LSTM layers, heaped one after the other joined to blend the advantage of a sole LSTM layer in an expound recurrent network manner feed with multiple value of neurons. The aim of the enhanced model is to construct the characteristics in a hierarchical design where the lower layer separates the input data disparity factors and these demonstrations are merged at the upper layer. Such deep structure will simplify well owing to a trimmed representation than a shallow design in case of large or complex datasets [32], [33] such as in the banana harvest yields.

D. Training and Forecasting

To execute the model, Keras library along with Theano and Tensorflow backend are properly installed and configured, splitting of dataset into 80% training and 20% testing is done respectively to capture the thirteen (13) periods of the year 2018 as test dataset. The number of epochs, batch size and neurons are assigned a minimal value from the summary result of the performance done in the tuning step. The “mean squared error” loss function with “ADAM” optimization algorithm are used in compiling the model. RMSE [34], [26], [22], [8] is used to evaluate the performance of the model to forecast. It calculates the variance between actual value and the predicted value and used to evaluate different models for a certain data and not amid datasets. The following formula is used for computing the RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{obs} - y_i^{pred}\right)^2} \qquad (7)$$

This measure is calculated by associating the target values for the time series and its consequent time series predictions, where $\boldsymbol{n}$ is the total number of observations, $\boldsymbol{y}_i^{obs}$ is the actual value; whereas, $\boldsymbol{y}_i^{pred}$ is the predicted value.

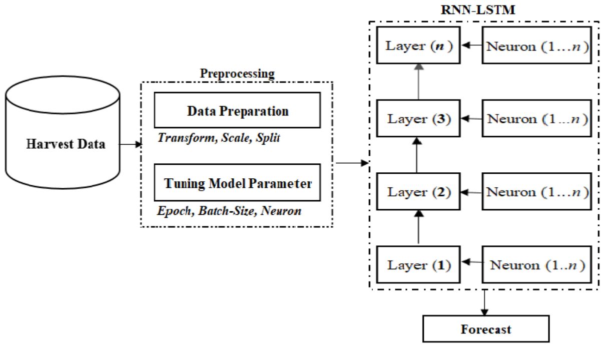


Fig. 2. The Flowchart of the Enhanced Model.

IV. Experimental Results and Discussion

The enhanced deep learning model was implemented using the banana harvest yields dataset. Given the nature of the dataset, long interval time and the missing (zero) observations are seen and considered. In view of the fact that missing values were observed, simply removing records containing zero value is the simplest strategy ever devised during the experiment with the goal of obtaining the optimal outcome of the enhanced model.

For consistency and fair assessment, splitting of data to 80% training and 20% for testing is done. Moreover, the model uses the optimal outcome from the lowest mean RMSE of the model parameter tuning for the values to be assigned for *epochs, batch size* and *neurons*. Noticeably, the number of epoch is very minimal because there is no proof that suggests the exact number of cycles (epochs) to train a model. Experiment wise, it is evident that setting the value of epoch and neurons minimally produces a sensible model outcome in terms of reduction of errors and forecasting accuracy most likely if using large datasets. The size of the diverse dataset is considerably important in setting the number of epochs because it shows different behavior to optimally train the network.

During the experiment, the model uses the optimal value of the parameters such as the number of epochs, batch size and neurons with respect to the result obtains from the model parameter tuning. Also, it has been noticed that executing the model more than once with the same parameter values does not guarantee better performance and at some instances, worsen the result. There are three sets of experiment done in forecasting using banana harvest yield dataset: *first*, using single LSTM layer, *second*, using multiple LSTM layers assigned with same value of neurons in each layer and *third*, using the enhanced deep learning model where multiple layers feed with multiple value of neurons. The latter used the precedent numbers of the optimal value of neuron obtained from the model parameter tuning. The results of experiments are exhibited in Table I, Table II and Table III, respectively. For uniformity, all experiments were done with and without a dropout rate. The second and third experiment used two up to four LSTM layers only.

TABLE I. Results Obtain using Single LSTM Layer

LSTM Layer	Parameters			RMSE	
	Epoch	Batch Size	Neurons	No Dropout	With Dropout
1	4	1	5	***61.602***	68.331

TABLE II. Results Obtain using Multiple LSTM Layers

LSTM Layers	Parameters			RMSE	
	Epoch	Batch Size	Neurons	No Dropout	With Dropout
2	4	1	5, 5	64.025	72.502
3	4	1	5, 5, 5	63.748	63.362
4	4	1	5, 5, 5, 5	***63.225***	68.702

TABLE. III. RESULTS OBTAIN USING THE ENHANCED APPROACH

LSTM Layers	Parameters			RMSE	
	Epoch	Batch Size	Neurons	No Dropout	With Dropout
2	4	1	1,2,3,4,5	***34.805***	46.867
2	4	1	1,2,3,4	36.670	37.020
2	4	1	1,2,3	44.606	47.688
2	4	1	1,2	35.030	36.909
3	4	1	1,2,3	37.947	36.670
3	4	1	1,2,3,4	47.498	49.837
3	4	1	1,2,3,4,5	36.371	45.562
4	4	1	1,2,3,4,5	50.104	50.085
4	4	1	1,2,3,4	39.290	36.771

Results showed that all the experiments convey better performances in the absence of dropout rate. The optimal result of the second experiment is composed of four LSTM layers using the same value of neurons with 63.225 in terms of RMSE. The third experiment performs best with 34.805 (in terms of RMSE) in the composition of two LSTM layers with the different minimal value of neurons. The enhanced deep learning approach outperforms the single LSTM layer and multiple LSTM layers with 43.5% and 44.95% reduction of error rates, respectively. Therefore, the deep structure of LSTM is proficient in forecasting highly nonlinear and long interval time data such as in banana harvest yields dataset. The result of the performance of the enhanced approach is considered significant since there is no standard (reduction of errors) available and no proof that it has been used in forecasting harvest yields of any agricultural crop such as bananas. Furthermore, the graphical representation of the best results of the performance obtained from each experiment is shown in the following figures: Fig. 3 represents the graphical plot of the single LSTM layer; Fig. 4 represents the multiple LSTM layers, while Fig. 5 is the graphical representation of the enhanced approach.

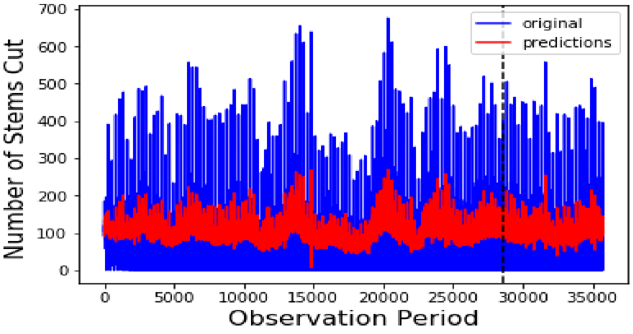


Fig. 3. Graphical Representation of the Optimal Result using the Single LSTM Layer.

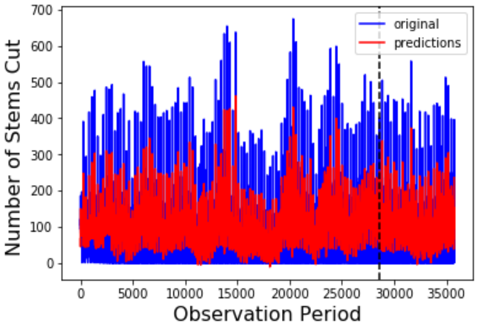


Fig. 4. Graphical Representation of the Optimal Result using Multiple LSTM Layer.

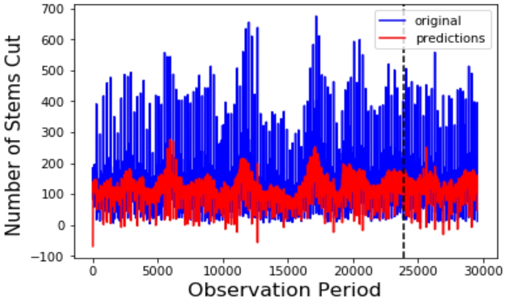


Fig. 5. Graphical Representation of the Optimal Result of the Enhance Approach.

V. CONCLUSION

This technical quest aspired to build deep multifaceted system proficient in forecasting banana harvest yields. In this paper, the enhanced model was trained from raw observations. The model achieved better performance with 43.5% and 44.95% reduction of error rates compared to single LSTM and multiple LSTM layers, respectively. It has been found out that applying dropout strategy and employing larger number of neurons does not guarantee optimal results on the datasets described in Section III-A. This outcome is a contribution to the unceasing growth of research in the agriculture domain since it has not been used for forecasting yields of crop such as bananas. It is also significant to the extensive planning for sustainability of ARB cooperatives subsidizing the over-all production. For future endeavors, it is suggested to investigate the performance of the model in diverse datasets specifically in handling multiple input features (multivariate) datasets. Furthermore, with the aim of providing help and support to the ARB cooperatives-IFS strategy and officers to the efficient and extensive planning for production sustainability, embedding the model to a web-based along with a handy application for recording and monitoring the interventions and factors influencing yields are well-thought-out for future work.

ACKNOWLEDGMENT

The author of this paper would like to thank the Agrarian Reform Beneficiaries (ARB) cooperative of Dapco in Davao del Norte, Philippines for the support and for providing the relevant data for this study.

REFERENCES

[1] "Kentaro Kuwata and Ryosuke Shibasaki The University of Tokyo IIS , The University of Tokyo , 4-6-1 Komaba , Meguro-ku , Tokyo 153-8505 , JAPAN," pp. 4–6, 2015.

[2] A. Koirala, K. B. Walsh, Z. Wang, and C. Mccarthy, "Deep learning – Method overview and review of use for fruit detection and yield estimation," Comput. Electron. Agric., vol. 162, no. January, pp. 219–234, 2019.

[3] S. Nagini, "Agriculture Yield Prediction Using Predictive Analytic Techniques," pp. 783–788, 2016.

[4] S. Rathod and K. N. Singh, "Hybrid Time Series Models for Forecasting Banana Production in Karnataka Hybrid Time Series Models for Forecasting Banana Production in Karnataka State , India," no. January 2018, 2017.

[5] B. Garg, S. Aggarwal, and J. Sokhal, "model R," Comput. Electr. Eng., vol. 0, pp. 1–21, 2017.

[6] D. Elavarasan, D. Raj, V. Sharma, and A. Y. Zomaya, "Forecasting yield by integrating agrarian factors and machine learning models : A survey," Comput. Electron. Agric., vol. 155, no. August, pp. 257–282, 2018.

[7] P. Surya and I. L. Aroquiaraj, "Crop Yield Prediction in Agriculture using Data Mining Predictive Analytic Techniques," vol. 5, no. 4, pp. 783–787, 2018.

[8] S. Siami-namini and N. Tavakoli, "A Comparison of ARIMA and LSTM in Forecasting Time Series," 2018 17th IEEE Int. Conf. Mach. Learn. Appl., pp. 1394–1401, 2018.

[9] S. Rathod and G. C. Mishra, "Statistical Models for Forecasting Mango and Banana Yield of," vol. 20, pp. 803–816, 2018.

[10] Y. Li and H. Cao, "ScienceDirect ScienceDirect ScienceDirect Prediction for Tourism Flow based on LSTM Neural Network Prediction for Tourism Flow based on LSTM Neural Network," Procedia Comput. Sci., vol. 129, pp. 277–283, 2018.

[11] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," 2015.

[12] R. Zuo, Y. Xiong, J. Wang, and E. J. M, "PT NU SC," Earth-Science Rev., 2019.

[13] A. Kamilaris and F. X. Prenafeta-boldú, "Deep learning in agriculture : A survey," vol. 147, no. July 2017, pp. 70–90, 2018.

[14] J. Chung, "Gated Feedback Recurrent Neural Networks," vol. 37, 2015.

[15] A. Sagheer and M. Kotb, "Neurocomputing Time series forecasting of petroleum production using deep LSTM recurrent networks," Neurocomputing, vol. 323, pp. 203–213, 2019.

[16] B. Cortez, B. Carrera, Y. Kim, and J. Jung, "PT US CR," Expert Syst. Appl., 2017.

[17] Y. Rizk and M. Awad, "Neurocomputing On extreme learning machines in sequential and time series prediction : A non-iterative and approximate training algorithm for recurrent neural networks," Neurocomputing, vol. 325, pp. 1–19, 2019.

[18] A. Elsheikh, S. Yacout, and M. Ouali, "PT US CR," Neurocomputing, 2018.

[19] J. Zhang, Y. Zhu, X. Zhang, M. Ye, and J. Yang, "Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas," J. Hydrol., vol. 561, no. April, pp. 918–929, 2018.

[20] K. Greff, R. K. Srivastava, J. Koutn, and B. R. Steunebrink, "LSTM : A Search Space Odyssey," pp. 1–12.

[21] K. Yao, T. Cohn, K. Vylomova, and C. Dyer, "Depth-Gated Recurrent Neural Networks," pp. 1–5.

[22] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," Eur. J. Oper. Res., vol. 270, no. 2, pp. 654–669, 2018.

[23] A. Sagheer and M. Kotb, "AC US CR," Neurocomputing, 2018.

[24] B. Yang, S. Sun, J. Li, X. Lin, and Y. Tian, "US CR," Neurocomputing, 2018.

[25] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," Energy, vol. 148, pp. 461–468, 2018.

[26] S. Muzaffar, A. Afshari, and S. Muzaffar, "ScienceDirect ScienceDirect ScienceDirect Short-Term Load Forecasts Using LSTM Networks on District and Cooling Short-Term Load Forecasts LSTM feasibility of using the heat demand-outdoor Masdar for," Energy Procedia, vol. 158, pp. 2922–2927, 2019.

[27] Y. Qin et al., "Hybrid forecasting model based on long short term memory network and deep learning neural network for wind signal," Appl. Energy, vol. 236, no. October 2018, pp. 262–272, 2019.

[28] J. Liu, Q. Li, W. Chen, Y. Yan, Y. Qiu, and T. Cao, "ScienceDirect Remaining useful life prediction of PEMFC based on long short-term memory recurrent neural networks," pp. 1–11, 2018.

[29] A. Singh, B. Ganapathysubramanian, A. K. Singh, and S. Sarkar, "Machine Learning for High-Throughput Stress Phenotyping in Plants," Trends Plant Sci., vol. 21, no. 2, pp. 110–124, 2016.

[30] R. Sensing, S. I. Sciences, and C. Vision, "Multi-Temporal Land Cover Classification with Long Short-Term Memory Neural Networks," vol. XLII, no. June, pp. 6–9, 2017.

[31] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks," pp. 1–5, 2017.

[32] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling q," vol. 42, pp. 11–24, 2014.

[33] M. Hermans and B. Schrauwen, "Training and Analyzing Deep Recurrent Neural Networks," pp. 1–9.

[34] W. Xu, Z. Zhang, D. Gong, and X. Guan, "Neural Network Model for the Risk Prediction in Cold Chain Logistics," vol. 9, no. 8, pp. 111–124, 2014.

Developing a Dengue Forecasting Model: A Case Study in Iligan City

Ian Lindley G. Olmoguez[1], Mia Amor C. Catindig[2], Minchie Fel Lou Amongos[3], Fatima G. Lazan[4]
Bachelor of Science in Information Technology[1, 3, 4]
College of Computer Studies, IT Department[2]
MSU–Iligan Institute Technology, Iligan City, Philippines[1, 2, 3, 4]

***Abstract*—Dengue is a viral mosquito-borne infection that is endemic and has become a major public health concern in the Philippines. Cases of dengue in the country have been recorded to be increasing, however, it is reported that the country lacks predictive system that could aid in the formulation of an effective approach to combat the rise of dengue cases. Various studies have reported that climatic factors can influence the transmission rate of dengue. Thus, this study aimed to predict the probability of dengue incidence in Iligan City per barangay based on the relationship of climatic factors and dengue cases using different predictive models with data from 2008 to 2017. Multiple Linear Regression, Poisson Regression, and Random Forest are integrated in a mini-system to automate the display of the prediction result. Results indicate that Random Forest works better with 73.0% accuracy result and 33.58% error percentage, with time period and mean temperature as predictive variables.**

Keywords—Dengue; predictive models; Pearson's correlation; multiple linear regression; Poisson regression; random forest

I. Introduction

Dengue has been an endemic infection in over 100 countries in the world, in tropical and subtropical regions. One of the four dengue viruses had been classified as dengue serotypes (DENV-1, DENV-2, DENV-3, and DENV-4), carried by the main vectors, the Aedes Aegypti and Aedes Albopictus [1]. Reports were gathered from the World Health Organization (2018) that the Dengue fever was the most critical and rapidly spreading mosquito-borne viral disease in the world for over the past 50 years with 390 million dengue infections per year in 3.9 billion people in 128 countries at risk of infection.

In the Philippines, Dengue fever has become one of the major health problems among the populace. The Department of Health (DOH) reported a total of 138,444 dengue cases nationwide from January 1 to October 6 2018 representing 21% increase on the number of cases in the same period in 2017 [2]. Just recently, there were 115, 986 dengue cases recorded in the Philippines including 491 deaths reported from January 1 to July 6 2019; 86 % higher than in 2018 [3]. Same report cited Region X as one of the regions with highest incidence rate having 9,354 cases [3]. Incidence of Dengue was caused by several factors, one of which was the climatic conditions referring to temperature, relative rainfall, and relative humidity were reported to be important influential Dengue transmitters [4]. Studies discovered that places with high temperatures and higher rainfall such as in the Philippines had high dengue transmission rates resulting to more steady water as potential breeding grounds for mosquitoes [3]. Humidity had been a consistent, significant weather factor that provided favorable conditions for Dengue vectors [5].

Despite the effort of the government to look for possible ways to control the increase of dengue cases in the Philippines, there is still no specific solution or response on how to control the dengue outbreaks up to this writing. This circumstance necessitates the implementation of primary safety measures to reduce and prevent dengue infections, to control mosquito populations, and limit the spread of dengue cases nationwide. Since climate conditions influence the dengue transmission cycle [6] [7] [8], a dengue risk-prediction system based on the relationship of dengue incidence and climatic conditions is investigated in this study. The development of a risk prediction system could forecast the locale of possible high incidence rate of dengue thus will have significant contribution in controlling the spread of dengue by reducing the transmission of mosquitoes [9].

Predictive analytical approach using a variety of machine learning, modeling, statistics, artificial intelligence, and data mining algorithms could be input data to predict unknown events in the future. Also, the use of statistical methods, correlations between dengue incidence and climatic variables were established to predict potential outbreaks in specific areas. Promprou [6] sampled a predictive model to predict the Dengue Haemorrhagic Fever (DHF) in Thailand using Multiple Linear Regression model to explain the relationship between the household's activities and DHF patients. Results of the study revealed a 26.9% of the variation of DHF patients using a number of water storage containers, Aedes Aegypti in drainage of refrigerators, pH and temperature of water in container. In the study by Ong et al. [7] predicting the Dengue incidence with the use of Random Forest approach, predicted the risk rank of dengue transmission in Singapore with dengue cases, population, entomological and environmental data. The evaluation using the latest dengue case data in the study showed a strong predictive ability for the model, compared to the study results of Tilwani, Dave, & Nadurbarkar [8], that adopted a regression approach with Poisson Regression and Negative Binomial to investigate the correlation between dengue incidence and climatic fluctuations including relative humidity, temperature, and pressure. Dengue cases with 70% accuracy showed the impact of climatic fluctuations in dengue transmission.

Although studies had been conducted by the foregoing authors, there remained a dearth of the study in Iligan City, Region X of the Mindanao Province, Philippines. It was on this premise that this study was conceptualized and developed to predict the number of dengue incidence based on the correlation between climatic conditions (temperature, relative rainfall, and relative humidity). Three predictive models were used by the study: Poisson Regression, Multiple Linear Regression and Random Forest. The data were limited only to climatic factors including the temperature, relative rainfall, and relative humidity, as provided for by the concerned office. The study will attempt to evaluate and compare the results of these predictive models that best fit in the case of Iligan City and discover and visualize the probability of dengue to arise in a particular area.

II. Related Studies

A study by Ong et al. (2018) [10] used a Random Forest approach that predicts the risk rank of dengue transmission in Singapore with dengue; population, entomological and environmental data. The predicted risk ranks are then categorized and mapped to color-coded risk groups which were evaluated with dengue cases and cluster data. According to its findings, the study demonstrates the potential of Random Forest and its strong predictive ability to stratify Singapore's spatial risk of dengue transmission. It suggests that population density, dengue burden and abundance of Ae. aegypti are significant risk factors for dengue transmission. Evaluation using the latest data of dengue cases showed a significant predictive ability for the model. Strong positive correlation between the observed and predicted ranks of risk and an almost perfect agreement between the predicted levels of risk and the density of the case were observed.

In the study of Carvajal, Viacrusis, Hernandez, Ho, Amalin, & Watanabe (2018) [11], various predictive models such as General Additive Modeling, Seasonal Autoregressive Integrated Moving Average with exogenous variables, Random Forest, and Gradient Boosting were used to predict the temporal pattern of dengue incidence in Metropolitan Manila, and to compare their predictive accuracy. Among the statistical modelling techniques, Random Forest showed a better predictive accuracy.

With the above mentioned studies, it has been supported that predictive analytics with the support of statistical methods are useful for predicting the possibilities of dengue incidence in various areas. Moreover, identifying the relationship of climatic factors and dengue transmission is considered to be one of the effective predictor. The presented study intends to predict the probability of dengue as well with dengue incidence rate and climatic factors such as temperature, rainfall, and humidity as the basis. Given the results from the related studies presented, it is perceived that predictive models are promising in predicting dengue incidences as they can provide significant prediction result.

III. Method

A. Study Area

The study area is in Iligan City in Lanao del Norte, Philippines. It is an urbanized city with 44 barangays, with estimated population of 400,000 according to 2015 census.

B. Research Design

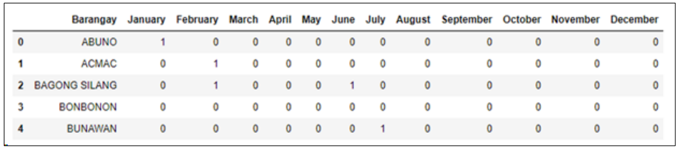
A quantitative study was utilized using mathematical models and statistics to analyze and provide more objective numerical results [12]. The presented research was quantitative in nature as it relied on the analysis of the secondary numerical data collected to describe and predict the probability of dengue incidence in Iligan City using predictive models.

C. Data Collection and Analysis

This study used secondary data obtained from the proper authority. A reported monthly dengue fever incidence data of 44 barangays in Iligan City over the period from 2008 to 2017 were provided by the City Health Office (CHO). The climatic data over the same period comprised monthly average temperature, maximum and minimum temperature, relative humidity and relative rainfall were also collected from the City Health Office (CHO). These data were used to analyze factors affecting the occurrence of dengue cases and to predict the next incidents of dengue in Iligan City.

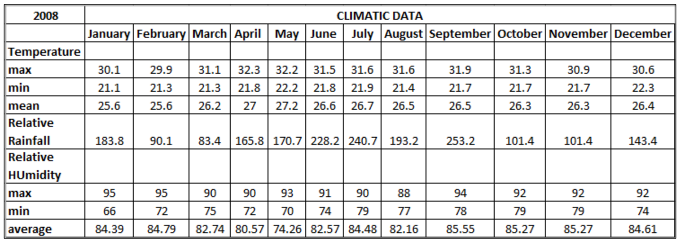
The collected data were imported to Python to evaluate inaccuracy or inconsistency in the data such as duplicate columns, not a number (NAN) values, and columns and rows that were not part of the explanatory factors. NAN values were replaced with 0, while gender and age were excluded as independent variables for these variables were presented daily while climatic data are presented by month. Fig. 1 shows a sample of the cleaned data while Fig. 2 is a sample of the climatic values.

A separate the data per year was kept by separating them to tabs and changing the variable names to max temperature, min temperature, mean temperature, relative rainfall, max relative humidity, min relative humidity, and average relative humidity.



	Barangay	January	February	March	April	May	June	July	August	September	October	November	December
0	ABUNO	1	0	0	0	0	0	0	0	0	0	0	0
1	ACMAC	0	1	0	0	0	0	0	0	0	0	0	0
2	BAGONG SILANG	0	1	0	0	0	1	0	0	0	0	0	0
3	BONBONON	0	0	0	0	0	0	0	0	0	0	0	0
4	BUNAWAN	0	0	0	0	0	0	1	0	0	0	0	0

Fig. 1. Cleaned Sample Data of Dengue Cases.



2008	CLIMATIC DATA											
	January	February	March	April	May	June	July	August	September	October	November	December
Temperature												
max	30.1	29.9	31.1	32.3	32.2	31.5	31.6	31.6	31.9	31.3	30.9	30.6
min	21.1	21.3	21.3	21.8	22.2	21.8	21.9	21.4	21.7	21.7	21.7	22.3
mean	25.6	25.6	26.2	27	27.2	26.6	26.7	26.5	26.5	26.3	26.3	26.4
Relative Rainfall	183.8	90.1	83.4	165.8	170.7	228.2	240.7	193.2	253.2	101.4	101.4	143.4
Relative HUmidity												
max	95	95	90	90	93	91	90	88	94	92	92	92
min	66	72	75	72	70	74	79	77	78	79	79	74
average	84.39	84.79	82.74	80.57	74.26	82.57	84.48	82.16	85.55	85.27	85.27	84.61

Fig. 2. Sample Climatic Data in 2008.

D. Development of the Predictive Model

Correlation analysis was used to determine the strength of the relationships between the monthly number of dengue cases, dependent variables and climatic factors (minimum temperature (tmin), maximum temperature (tmax), mean temperature (tmean), relative rainfall (rr), minimum relative humidity (rhmin), maximum relative humidity (rhmax), and average relative humidity (rhmean)) and time period (timeperiod) as independent variables, covering 120 months in total from 2008 to 2017.

To see how the data sets were correlated, Pearson's Correlation Coefficient was used. Given that the data are continuous, this method is suitable to perform as it is generally used when variables are continuous in nature such as in ratio or interval scale variables. Pearson's correlation coefficient is indicated by r and defined by:

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma y^2 - (\Sigma y)^2\}}} \qquad (1)$$

The value of r always ranges from -1 to +1. The relationship between the variables is said to be not related when the value of r comes down to 0. If the value of r lies to +1, then the variables are said to be positively correlated, while the variables are said to be negatively correlated if the value of r is -1.

The development of the forecasting model is based on Multiple Linear Regression (MLR), Poisson Regression, and Random Forest.

Multiple Linear Regression analysis is a statistical technique that used several explanatory (independent) variables to predict the outcome of a response (dependent) variable. The model is developed with the following equation.

$$y = b0 + b1x1 + b2x2 + b3x3 + ... + bnxn \qquad (2)$$

Where

bi = y-intercept

y = dengue cases

xi = climatic factors and time period

The value of R-squared and adjusted R-squared were calculated in order to test how well the data fit the regression model. Usually, the higher R-squared value between 0 and 100 means the better the regression model fitted the observations. To identify the significance of each of the independent variables, P-value was also computed.

The Poisson Regression on the other hand is designed to fit a model of regression in which counts were made with the dependent variable Y (dengue cases). The fitted model Y to one or more X predictor variables (climatic factors and time period), which were either quantitative or categorical. A poisson regression model defined as:

$$z = e + b0 + b1x1 + b2x2 + b3x3 + ... + bkxk \qquad (3)$$

where z = log(y) , transformed from y in generalized linear modelling called link function. This was done so that a linear regression modelling in z satisfied all required assumptions [11]. The corresponding coefficients (b1, b2 … bk), which were originally modelled for z, have to be transformed back properly to a model for y for interpretation. To estimate the b values, a process called maximum likelihood estimation (MLE) or weighted least squares may be used where it fitted a model.

Random forest is an ensemble of simple tree predictors used to determine the final outcome. It used a bootstrap sampling approach to generate k different training data subsets from an original dataset, and then k decision trees were constructed by training these subsets. The final value was calculated by taking the average of all the predicted values by all the trees in forest. Since there was no given regression coefficients in building the Random Forest model, variable selection was done through performing the Feature Importance method, which variables had the most effect on the model. Visualizing tree in Random Forest was to evaluate the predictions for each row using all the trees in the model, how each variable contributed to the final prediction. To calculate for the result, propents have used the RandomForestRegressor class of the sklearn.ensemble library in Python.

However, before performing the prediction process in MLR, selecting highly significant independent variables based on their p-value was done first in order to build a best model, called the Stepwise Regression.

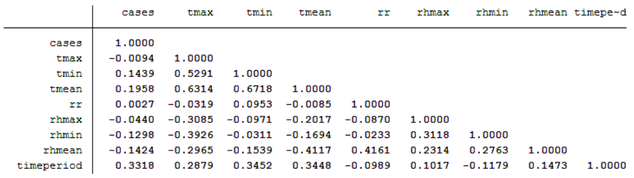
Dengue Prediction System: The proposed system was implemented using Python and PostgreSQL as the backend, while the front end was developed using CSS, Javascript, HTML, Leafleat, and Django. This system will automate the prediction of the model chosen, provided that data are presented in the correct format.

IV. Results and Discussion

As shown in Fig. 3, most of the values were not close to 1 which indicates that there is a weak correlation between independent variables. On the other hand, timeperiod associated with cases has an r of 0.3318 and has the greatest correlation coefficient among the other independent variables. Other independent variables such as tmin and tmean have positive correlation with dengue cases having r values equal to 0.1439 and 0.1958, respectively while there is negative correlation with rhmin and rhmean with cases havingr equal to -0.1298 and -0.1424 consecutively.

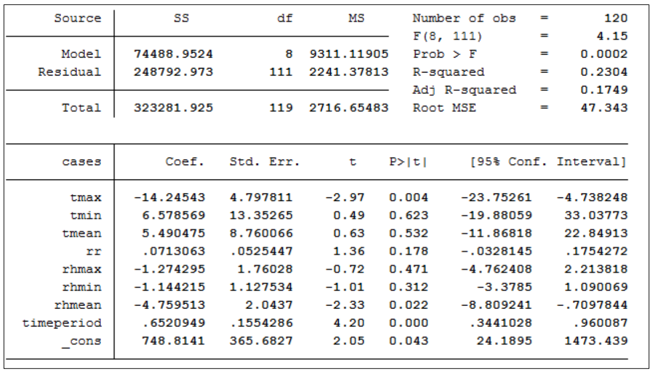
Result of the Multiple Linear Regression displays in Fig. 4, shows that r2 is equal to 0.2304. This indicates that all eight of the independent variables explain 23.04% of the variability on the number of dengue cases. Moreover, it was shown that some independent variables had a p-value greater than 0.05 lowering their significance in predicting dengue incidence. From this model, only the maximum temperature, average relative humidity, and time period have p-values less than 0.05; thus, have significant impact on dengue cases.

Stepwise regression, as seen in Fig. 5, was conducted to eliminate non-significant variables to build a reliable model out of the remaining independent variables. The variables tmin, rhmax, tmean, and rhmin were removed from the model. These variables were omitted in building the Multiple Linear Regression since their respective p-values were higher than 0.05.



```
             |    cases     tmax     tmin    tmean       rr    rhmax    rhmin   rhmean timepe~d
-------------+-----------------------------------------------------------------------------------
       cases |   1.0000
        tmax |  -0.0094   1.0000
        tmin |   0.1439   0.5291   1.0000
       tmean |   0.1958   0.6314   0.6718   1.0000
          rr |   0.0027  -0.0319   0.0953  -0.0085   1.0000
       rhmax |  -0.0440  -0.3085  -0.0971  -0.2017  -0.0870   1.0000
       rhmin |  -0.1298  -0.3926  -0.0311  -0.1694  -0.0233   0.3118   1.0000
      rhmean |  -0.1424  -0.2965  -0.1539  -0.4117   0.4161   0.2314   0.2763   1.0000
  timeperiod |   0.3318   0.2879   0.3452   0.3448  -0.0989   0.1017  -0.1179   0.1473   1.0000
```

Fig. 3. Correlation Matrix between the Dengue Incidence Rate and Weather Variables with Time Period of 0 to 120 Months from 2008 to 2017 in Iligan City.



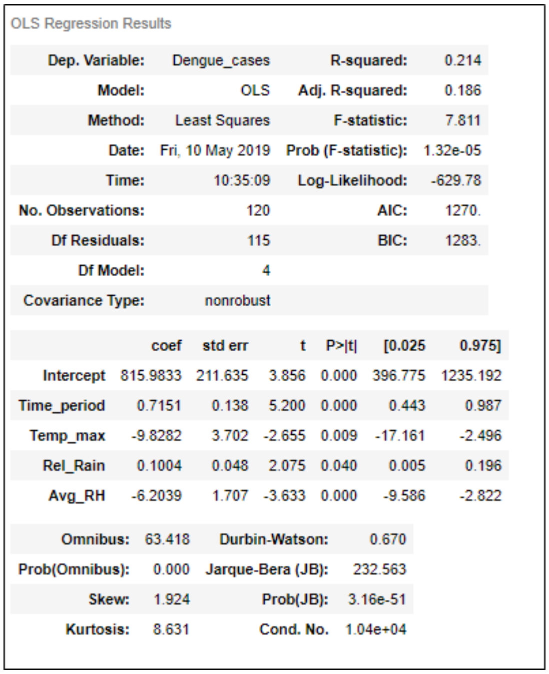
```
      Source |       SS           df       MS      Number of obs   =       120
-------------+----------------------------------   F(8, 111)       =      4.15
       Model |  74488.9524         8  9311.11905   Prob > F        =    0.0002
    Residual |  248792.973       111  2241.37813   R-squared       =    0.2304
-------------+----------------------------------   Adj R-squared   =    0.1749
       Total |  323281.925       119  2716.65483   Root MSE        =    47.343

------------------------------------------------------------------------------
       cases |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        tmax |  -14.24543   4.797811    -2.97   0.004    -23.75261   -4.738248
        tmin |   6.578569   13.35265     0.49   0.623    -19.88059    33.03773
       tmean |   5.490475   8.760066     0.63   0.532    -11.86818    22.84913
          rr |   .0713063   .0525447     1.36   0.178    -.0328145    .1754272
       rhmax |  -1.274295    1.76028    -0.72   0.471    -4.762408    2.213818
       rhmin |  -1.144215   1.127534    -1.01   0.312      -3.3785    1.090069
      rhmean |  -4.759513     2.0437    -2.33   0.022    -8.809241   -.7097844
  timeperiod |   .6520949   .1554286     4.20   0.000     .3441028     .960087
       _cons |   748.8141   365.6827     2.05   0.043      24.1895    1473.439
------------------------------------------------------------------------------
```

Fig. 4. Performance Analysis of Multiple Linear.

```
p = 0.6232 >= 0.0500 removing tmin
p   =   0.4801   >=   0.0500   removing
rhmax
p   =   0.3408   >=   0.0500   removing
```

Fig. 5. Variable Selection in Multiple Linear Regression using Stepwise Regression.

Fig. 6 shows the progress of the result of the Multiple Linear Regression model after eliminating insignificant independent variables.



```
OLS Regression Results
      Dep. Variable:      Dengue_cases          R-squared:     0.214
             Model:                OLS     Adj. R-squared:     0.186
            Method:      Least Squares        F-statistic:     7.811
              Date:   Fri, 10 May 2019  Prob (F-statistic):  1.32e-05
              Time:           10:35:09     Log-Likelihood:   -629.78
  No. Observations:                120                AIC:     1270.
      Df Residuals:                115                BIC:     1283.
          Df Model:                  4
   Covariance Type:          nonrobust

                 coef   std err       t   P>|t|    [0.025     0.975]
   Intercept  815.9833  211.635   3.856   0.000   396.775   1235.192
 Time_period    0.7151    0.138   5.200   0.000     0.443      0.987
    Temp_max   -9.8282    3.702  -2.655   0.009   -17.161     -2.496
    Rel_Rain    0.1004    0.048   2.075   0.040     0.005      0.196
      Avg_RH   -6.2039    1.707  -3.633   0.000    -9.586     -2.822

        Omnibus:  63.418     Durbin-Watson:     0.670
 Prob(Omnibus):   0.000   Jarque-Bera (JB):   232.563
           Skew:  1.924           Prob(JB):  3.16e-51
       Kurtosis:  8.631          Cond. No.  1.04e+04
```

Fig. 6. Performance Analysis of Multiple Linear Regression after Removing Non-Significant Independent Variables.

Based on the result, time period, maximum temperature, relative rainfall, and average humidity, all indicates significant impact on predicting the number of dengue cases with p-value of less than 0.05. Comparing its result from the first model, the value of adjusted R-squared increases, adjusted R2 = 0.186 or 18.6% which conveyed that omitting non-significant independent variables could improve the accuracy result of the model.

Using the coefficient values of the identified significant factors would generate the predictive model in Eq. (4).

cases = 0.7151(Time_period) - 9.8282(Temp_max) + 0.1004(Rel_Rain) - 6.2039(Avg_RH) + 815.9833 (4)

Fig. 7 shows the relation between the number of cases and time period for multiple linear regressions. Time period represents the total number of months involved in the study from January 2008 to December 2017. It is evident in the results that there are a number of gaps between the actual and predicted values.

A. Poisson Regression

One of the major assumptions in Modeling Poisson regression model is the equality of the mean and variance. As shown on Table I, the data that was used for Poisson Regression model did not meet the assumption wherein the mean and variance should be equal. The result showed that the variable Time_period had 60.5 mean and 1210 variance which is an overdispersion whereas Temp_max has 32.09333 mean and 1.753401 variance which results to underdispersion.

B. Random Forest

By performing the *Feature Importance* method in Python, the significant independent variables with the highest significance out of the eight are time period with a 0.48 and mean temperature with a value of 0.12. Results are shown in Fig. 8. These variables were used to train the random forest algorithm. Fig. 9 illustrates the adjusted R-Squared of using Random Forest having a value of *73.0%.* This means that timeperiod and tmean in this model explain *73.0%* of the variability of the dependent variable. It is evident in Fig. 10 that the actual and predicted values of random forest from the time period of January 2008 to December 2017 are closer from the actual number of dengue cases. The Mean Absolute Percentage Error (MAPE) of the model using Random Forest is *33.58%.*

C. Evaluation of the Models

Based on the summary of results of the three models shown in Table II, it indicates that Random Forest has a greater accuracy with 73.0% compared with Multiple Linear Regression with an adjusted R-Squared value of *18%*. Further, Random Forest has lower MAPE score result with *33.58%* while Multiple Linear Regression has *67.14% MAPE* value. Comparing the results of the two models, Random Forest having two significant independent variables specifically the time period and the mean temperature clearly performs relatively better having higher accuracy score and lowest MAPE score than Multiple Linear Regression. This output explains the flexibility and the potential of Random Forest in predicting the number of dengue incidence based on climatic factors.

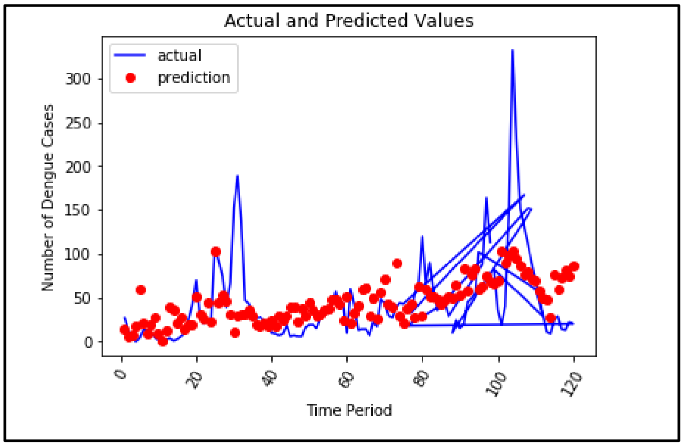
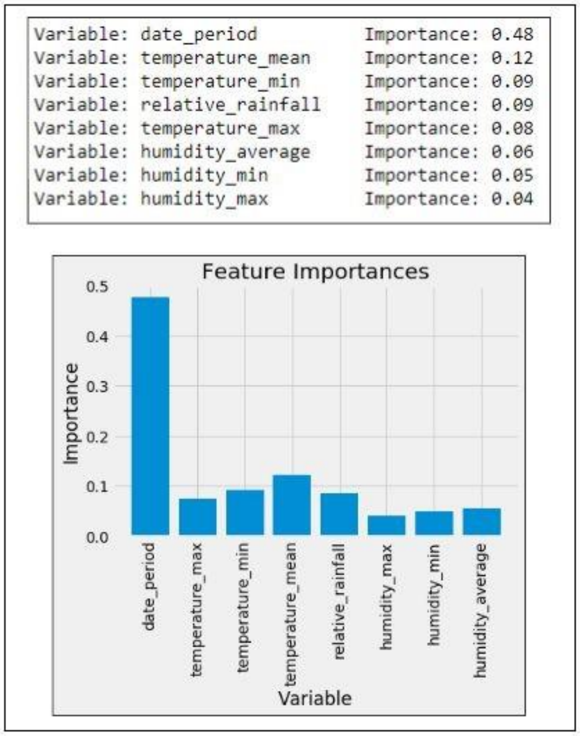


Fig. 7. The Actual Values and Predicted Number of Dengue Incidence from Time Period of January 2008 to December 2017 for Multiple Linear Regression.

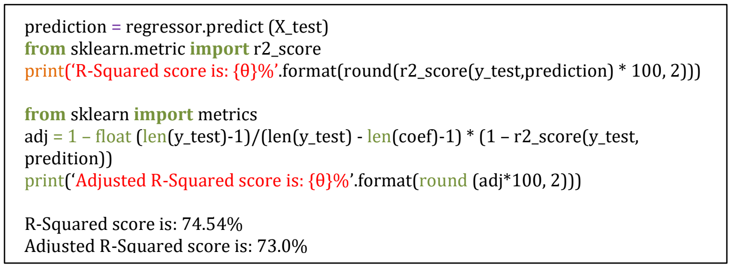
TABLE. I. Descriptive Analysis of the Data for Poisson Regression

Variable	**Mean**	**Variance**
Time_period	60.5	1210
Temp_max	32.09333	1.753401
Temp_min	22.52833	.2198627
Avg_Temp	27.32917	.708806
Rel_Rain	159.6975	10250.92
RH_max	91.88333	7.566106
RH_min	73.00833	21.40329
Avg_RH	82.913	9.554999



```
Variable: date_period          Importance: 0.48
Variable: temperature_mean     Importance: 0.12
Variable: temperature_min      Importance: 0.09
Variable: relative_rainfall    Importance: 0.09
Variable: temperature_max      Importance: 0.08
Variable: humidity_average     Importance: 0.06
Variable: humidity_min         Importance: 0.05
Variable: humidity_max         Importance: 0.04
```

Fig. 8. Feature Importance Performed for Random Forest.



```
prediction = regressor.predict (X_test)
from sklearn.metric import r2_score
print('R-Squared score is: {0}%'.format(round(r2_score(y_test,prediction) * 100, 2)))

from sklearn import metrics
adj = 1 – float (len(y_test)-1)/(len(y_test) - len(coef)-1) * (1 – r2_score(y_test,
predition))
print('Adjusted R-Squared score is: {0}%'.format(round (adj*100, 2)))

R-Squared score is: 74.54%
Adjusted R-Squared score is: 73.0%
```

Fig. 9. The Adjusted R-Squared using Random Forest.

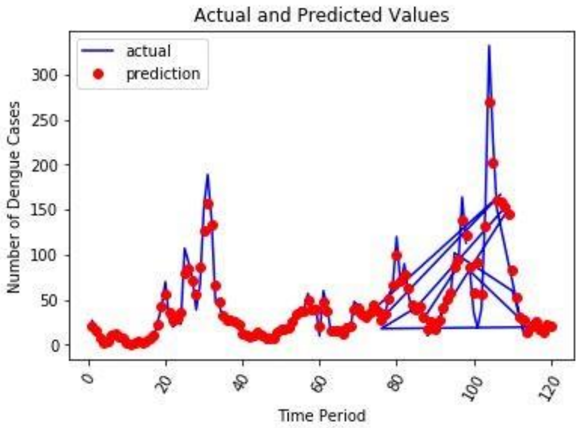
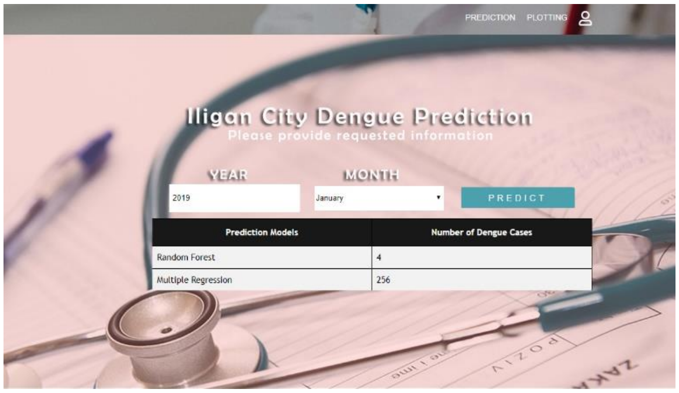


Fig. 10. The Actual Values and Predicted Number of Dengue Incidence of Random Forest from the Time Period of January 2008 to December 2017 using Random Forest.

TABLE. II. Comparison of the different Predictive Models

Predictive Models	Accuracy Result (Adjusted R-Squared)	Mean Absolute Percentage Error (MAPE)
Random Forest	73.0%	33.58%
Multiple Linear Regression	18%	67.14%

The proposed system generates prediction of dengue based on historical data and reported dengue cases in each barangay that can assist the society and local health offices in Iligan City. Fig. 11 is a screenshot of the system as output from the results of this study.



Iligan City Dengue Prediction

Please provide requested information

YEAR: 2019

MONTH: January

Prediction Models	Number of Dengue Cases
Random Forest	4
Multiple Regression	256

Fig. 11. A Screenshot of the Developed Dengue Prediction System.

V. Conclusion

The correlation analysis between each climatic factor has different level of significance in relation with dengue cases. This means that some of the independent variables can significantly affect the transmission of dengue while some factors are not. Time period is one of the highly correlated variable and even with most of the climatic factors. This only shows that time period has valuable significance as it helps to determine dengue in a given period of time.

Predictive models, specifically Multiple Linear Regression and Random Forest, can potentially predict dengue cases in Iligan City based on the results compared to Poisson Regression which was excluded due to the violation of the assumptions of the model. In building the Multiple Linear Regression model, coefficient of the significant independent variables were considered and formulated as:

dengue cases = 0.1003793 (rr) - 9.828208 (tmax) - 6.203866 (rhmean) + 0.7150807 (time period) + 815.9834

On the other hand, the result of the evaluation with Random Forest performed better in terms of its goodness-of-fit and error difference between its actual and predicted values, having an accuracy percentage of 73% and 33.58% error result, than Multiple Linear Regression with only 18% accuracy percentage and error result of 67.14%.

Considering the results, Random Forest has the highest accuracy output and smallest error measure compared with multiple linear regressions. This indicate that this model has a positive impact in providing a prediction model for dengue cases in Iligan City. However, the accuracy percentage of the predictive model is lower than the expected percentage of the researchers. One factor is possibly due to the limited data of climatic factors used in the study. Considering other factors in dengue prediction may increase accuracy results of the predictive model.

VI. Recommendation

It is recommended that additional possible independent variables alongside climatic data must be considered such as the topographic profile of the area, ecological, biological and sociological aspects. Gathering of climatic data is recommended if possible to be daily or weekly rather than monthly and must be specific to every barangay to further investigate the impact of climatic factors in predicting dengue cases specific for every barangay. Further, temperature, rainfall, and humidity alone can contribute in predicting number of dengue cases. But considering other climatic factors including wind speed, precipitation, air pressure, and other weather data can provide better prediction result. Analysis and the use of other predictive models are highly recommended to test further the accuracy of forecasting dengue cases in Iligan City.

Acknowledgment

The study would like to acknowledge the Center of Health in Iligan by providing the data concerned in the development of the project.

References

[1] T. W. Chuang, K. C. Ng, T. Nguyen and L. Chaves, "Epidemiological characteristics and space-time analysis of the 2015 dengue outbreak in the metropolitan region of Tainan City," Taiwan. International journal of environmental research and public health, 2018.

[2] A. De Vera, "Dengue Cases Up," Manila Bulletin, 24 October 2018. [Online]. Available: https://news.mb.com.ph/2018/10/24/dengue-cases-up. [Accessed 22 December 2018].

[3] Philippines Department of Health Surveillance Reports, "Dengue Outbreak," WHO, Philippines, 2019.

[4] V. G. Ramachandran and e. al, "Empirical model for estimating dengue incidence using temperature, rainfall, and relative humidity: a 19-year retrospective analysis in East Delhi," Epidemiology and health, vol. 38 e2016052, 2016.

[5] University Corporation for Atmospheric Research, "Climate Change and Vector-Borne Disease," [Online]. Available: https://scied.ucar.edu/longcontent/climate-change-and-vector-borne-disease. [Accessed 20 February 2019].

[6] S. Promprou, "Multiple Linear Regression Model to Predict Dengue Haemorrhagic Fever (DHF) Patients in Kreang Sub-District, Cha-Uat District, Nakhon Si Thammarat, Thailand," JOURNAL OF APPLIED SCIENCES RESEARCH, pp. 6193-6197, 2014.

[7] J. Ong, X. Liu, J. Rajarethinam, S. Y. Kok, S. Liang, C. S. Tang and G. Yap, "Mapping dengue risk in Singapore using Random Forest," PLoS Neglected Tropical Diseases, vol. 12:e0006587, 2018.

[8] T. K, D. G and a. N. V, "Impact of Climatic Fluctuation on Dengue Virus Etiology," Journal of Molecular and Genetic Medicine, vol. 12(1): 331, 2018.

[9] C. Hettiarachchige, S. von Cavallar, T. Lynar, R. I. Hickson and M. Gambhir, "Risk prediction system for dengue transmission based on high resolution weather data," Plos One, vol. 13(12)e0208203, no. e0208203, 2018.

[10] J. Ong, X. Liu, J. Rajarethinam, S. Yheng Kok, S. Liang, C. S. Tang, A. Cook, L. C. Ng and G. Yap, "Mapping dengue risk in Singapore using," PLoS Negl Trop Dis 12(6):e0006587, 2018.

[11] T. M. Carvajal, K. M. Viacrusis, L. F. Hernandez, H. T. Ho, D. M. Amalin and K. Watanabe, "Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in Metropolitan Manila, Philippines," BMC Infectious Disease, Vols. vol. 18, no. 1, 2018, 2018.

[12] F. Moore, "Qualitative vs. Quantitative Design," DOI: 10.13140/RG.2.2.34861.49128, 2016.

Performance Evaluation of Network Gateway Design for NoC based System on FPGA Platform

Guruprasad S.P[1]
Research Scholar, Dept. of ECE
Jain University, Bangalore, India

Dr.Chandrasekar B.S[2]
Director, CDEVL
Jain University, Bangalore, India

***Abstract*—Network on Chip (NoC) is an emerging interconnect solution with reliable and scalable features over the System on Chip (SoC) and helps to overcome the drawbacks of bus-based interconnection in SoC. The multiple cores or other networks have a boundary which is limited to communicate with devices, which are directly connected to it. To communicate with these multiple cores outside the boundary, the NOC requires the gateway functionality. In this manuscript, a cost-effective Network Gateway (NG) model is designed, and also the interconnection of a network gateway with multiple cores are connected to the NoC based system is prototyped on Artix-7 FPGA. The NG mainly consists of Serializer and deserializer for transmitting and receiving the data packets with proper synchronization, temporary register to hold the network data, electronic crossbar switch is connected with multiple cores which are controlled by switch controller. The NG with the Router and different sizes of NoC based system is designed using congestion-free adaptive-XY routing. The implementation results and performance evaluation are analyzed for NG based NoC in terms of average Latency and maximum Throughput for different Packet Injection Ratio (PIR). The proposed Network gateway achieves low latency and high throughput in NoC based systems for different PIR.**

***Keywords*—Network gateway; network on chip; FPGA; routing; network interface; crossbar switch**

I. Introduction

The NoC will play an emerging role in future high-performance Chip Multi-Processor (CMP) to address the problems of interconnections. In recent years, most of the research focused on a packet-switched NoC design, which improves the system performance by using optimization techniques in the network for better Latency and bandwidth and also supports on-chip and off-chip communications. The NoC based photonic communications support a mechanism for large data transmission with higher bandwidth and less power consumption. The photonic based NoC supports Multiple cores interface using gateway switch [1] [3]. Multiple cores residing in a single chip (MPSoC) exist towards mixed-criticality system includes dependability, security, and different block access with shared resources. The outside network real-time messages are communicating to MPSoC using a gateway [2]. In general, the network gateway is a node which connects two different networks with different transmission protocols and simplifies the internet connectivity into one electronic device. The gateway node acts as a firewall and proxy server for business use. Gateways are a protocol which provides the compatibility between two different protocols and will be operating in any of the Open system interconnection (OSI) layers. The multifunctional intercommunication supported by the Gateway on a single-core chip. The different protocol standards like Bluetooth, Modbus, serial bus, Process Filed bus, and Controller area Network (CAN) provides intercommunication using gateway [4]. The intelligent Gateway has interoperation and achieves better communication among different bus networks with reconfigurability and also supports fast conversion speed, flexibility, intellectual control ability, reliability, and higher-level interface. The protocol converting Gateway works on Most of the OSI layers [5-6]. The high-performance computation needs high-speed interconnection like Ethernet and Infiniband. The data transmission between two heterogeneous networks needs an efficient network gateway to improve system performance in terms of bandwidth and Latency [7] [12].

The gateway terminology is used commonly for most of the applications for protocol conversion and data packets transfers. The network gateways are used in most of the real-time embedded and Internet of Things (IoT) applications. The home gateway requires a standard ARM chip with SoC chip which integrates the Customer Electronics Bus (CEbus) with home appliances like TV, microwave oven, refrigerator, and washing machine. The user sends a command to the internet; the network control module receives the command, issues request signal to Chip to control the home appliances [8]. The heterogeneous Gateway provides different interfaces to internet, GSM, CDMS, PSTN, and so on, to support different application scenarios [9]. The embedded Gateway is a backbone for smart grid home networks [10], wireless applications [11] [14], indoor high precision positioning systems [13], and IoT applications [15] for communicating with other networks.

In this manuscript, a cost-effective Network Gateway model is designed along with Gateway based NoC system using Adaptive XY Routing. The Network gateway results are hardware resource-efficient, works at low Latency, and High Throughput for input traffics which are evaluated for NoC based system. Section II explains about related work on Gateway mechanism used for different applications and also explains about research findings. Section III elaborates the Network Gateway architecture using electronic crossbar switch with an explanation. Section IV explains the Network Gateway based NoC based system with router architecture. The results and performance evaluation are analyzed with tables and graphs in Section V. Finally concludes the overall proposed work with Future scope in Section VI.

II. Related Works

In this section, the general Gateway related work and applications of Gateway are reviewed. Shi et al. [16] presented an embedded dual home architecture with secured Gateway both on hardware and software platform. The Gateway improves the transmitting information risk by the user and network isolation module using FPGA is incorporated to improve the security features using data signature and key management. The secured embedded Virtual private network (VPN) gateway is presented by Han et al. [17] to improve the data transmission security with protection capability in application terminals. This VPN gateway is worked under L3, L4, and L7 layers with firewall protection, VPN Functioning, and network isolation modules. Ajami et al. [18] presented an FPGA Based embedded network firewall which supports highly customized data packet filtering on a network gateway. These firewall customized in real-time by changing the TCP/UDP port id, Source MAC address, and source-destination IP address. Abuteir et al. [19] introduced a gateway design to establish the hierarchical platform for multi-core chips interaction either on on-chip or off-chip networks. The software-based Gateway supports message classification, message – scheduling, traffic shaping services, downsampling, service, protocol conversion, egress-queuing, ingress-Queuing, Virtual-Link queuing, and also supports serialization services.

Obermaisser et al. [20] described the mixed-criticality systems for end to end real-time communication, which involves gateways between multiple off-chip networks, Gateway between off-chip and on-chip networks. The gateway node resolving the contention between source controlled and autonomous networks and also supports end-to-end addressing and routing. The cloud storage gateway was presented by Dumitru et al. [21] on FPGA platform. The secured data encryption and transparency are resolved by using FPGA between host and outside interface in cloud infrastructure. Lee et al. [22] presented a high-performance hardware-software based gateway design for In-Vehicle Network (IVN) for CAN/ FlexRay controllers. The data conversion between CAN to FlexRay and vice-versa is achieved using Routing table converter block with AXI interface on Zed board. Shreejith et al. [23] described the vehicular Ethernet Gateway connected with multiple network protocols like FlexRay, CAN, and Ethernet with embedded computing Units. The Ethernet gateway is designed using Switch fabric between FlexRay and Ethernet controller. The switch fabric is designed using Crossbar switch.

The embedded Gateway for Fourth Generation (4G) mobile network and process Fieldbus (PB) with decentralized Periphery (DP) is described by Zhou et al. [24] on FPGA platform. The AES algorithm is used for secured data transaction in Gateway. The Gateway is used to connect two different protocol 4G and PB conversion in terms of data. The Korona et al. [25] introduced an Internet Protocol security (IPsec) gateway for multi-gigabit networks which includes security association database to store secure information, Internet key exchange to set secure channels, and responsible for all security operation with packet encapsulation. The programmable-SoC (PSoC) based cyber-physical production system (CPPS) gateway is described by Urbina et al. [26] to meet the industry 4.0 standards. The industrial network architecture includes CPPS Gateways, which are interconnected with multiple peripherals, electronic and electrical devices using different network protocols like Profinet, Profibus, and High availability Seamless Redundancy (HSR). Kwak et al. [27] present the trust domain gateway system to solve the untrusted internet structural problems.

Gaps in the research: Most of the work carried on traditional software-based gateway designs lacks with latency and throughput issues. Hardware-based Network gateway designs use bus-based interconnections for embedded real-time applications and lack of scalability and reliability problems. The existing research work is done on protocol conversion using gateways, but not on NoC based system. In order to resolve these problems, a cost-effective Network gateway with NoC based system is designed.

III. Network Gateway Design

The Gateway provides the network and access information to the four gateway cores, and the hardware architecture of the network gateway is represented in Fig. 1. The network gateway mainly consists of deserializer and Serializer for receiving and transmitting the data information's with proper synchronization, Temporary register, Electronic crossbar switch, Switch controller, priority encoder, and four gateway cores. The gateway cores are processors, buffers, caches, peripheral devices, etc. The FIFO buffers are considered in the design.

The data information is received from the network either from the interface or from the Router to deserializer, which receives the data signals serially, works based on Serial In Parallel Out (SIPO) manner. The received 8-bit data converts to 32-bit data to parallel using shifting operation along with issuing the synchronization signal to Serializer. The synchronization is achieved between Serializer and deserializer using counter method and proper clocking mechanism. The temporary register receives the deserialized data, holds for access to the electronic crossbar. This temporary register is only used to store the received deserialized data signals and that are scheduled towards for the gateway cores through switch controller. The electronic crossbar switch receives the temporary data along with gateway core (buffer) inputs and works based on switch controller, and its hardware architecture is represented in Fig. 2.

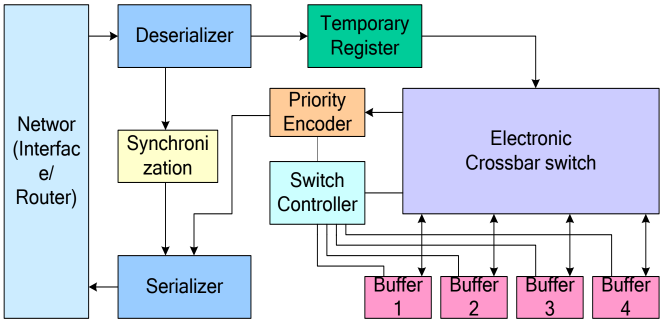


Fig. 1. Hardware Architecture of Network Gateway.

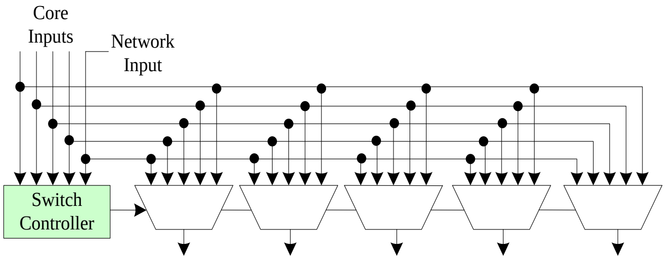


Fig. 2. Electronic Crossbar Switch Diagram.

The electronic crossbar switch have mainly five multiplexors, and each multiplexer is five inputs, four from the gateway cores and one from the temporary register (which is coming from the network). The switch controller issues the select line based time slot and priority to the crossbar switch and generates the prioritize output. The switch controller receives the five request inputs from the gateway cores and temporary register. The switch controller works based on the arbitration and time slot. The controlling mechanism is incorporated in the switch controller using Finite State Machine (FSM), which receives the input requests, gives priority to the corresponding input and other requests in waiting for the state.

The electronic crossbar issues the data signals based switch controller to priority encoder as an input. The same select signal issue the prior encoded data signal to Serializer. The Serializer is ready to transmit the data signals to the network form the crossbar switch. The Serializer converts the 32-bit parallel data information to 8-bit serially in PISO manner with synchronization and sends to network. The received and transmitted data of the network is same in the Gateway, which proves that the designed Gateway is working effectively.

IV. Network Gateway for NoC System

The network gateway is interconnected to the NoC based systems which offer on-chip and off-chip data flow control and arbitration between many gateway cores interconnections of the NoC based Multiprocessing SoC (MPSoC). The MPSoC chips are considered as an FPGA or ASIC devices for prototyping the network gateway with NoC. The network Gateway interconnected to NoC, and it is represented in Fig. 3. This is an example of 4x4 Mesh topology-based Network gateways with NoC Connection. It mainly contains 16 routers, 16 network gateways with 64-processing cores and all are interconnected with linked wires. This architecture is flexible to support any of the64- processing core information's that can transmit to any of the 16 routers via network gateways using Adaptive routing algorithm.

The network gateway with cores is connected to routers via a network interface (NI). In design, Mesh topology is selected to design 2x2, 3x3, and 4x4 NoC architectures. In Fig. 3, the 4x4 NoC has 16 routers (R1 to R16), and all the routers are interconnected using linked wires. All the network Gateway with cores inputs are received to the corresponding routers via the network interface and perform the data transaction based on the destination address of the corresponding routers.

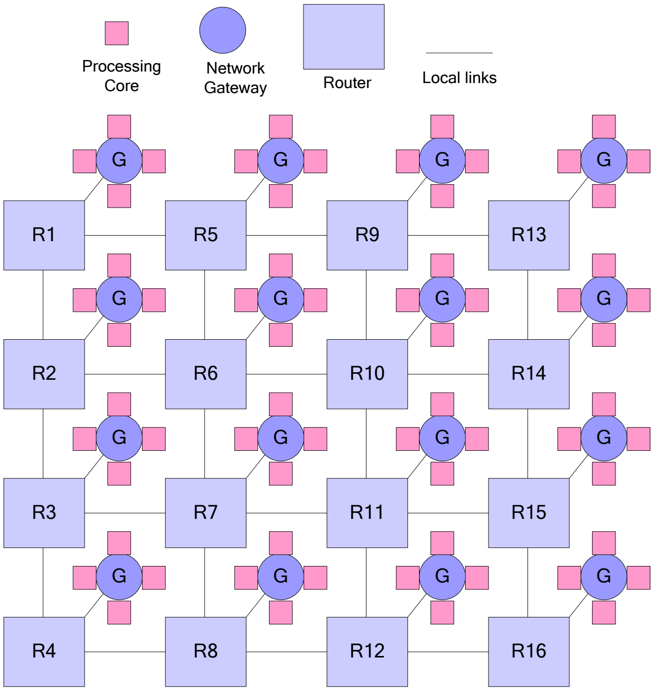


Fig. 3. Network Gateway based NoC Design Architecture.

The router architecture of Network Gateway is represented in Fig. 4. The designed Router is congestion-free Router which finds the shortest route to reach the destination. Each Router has five–port input registers followed by packet formation with priority-based arbitration and adaptive XY routing algorithm. The five-port input register receives gateway data information and stores it in local input port (Li), and For NoC, supported service inputs are East (Ei), West (Wi), South (Si) and North (Ni) are presented to route to corresponding destination locations.

The 8-bit local gateway data are used for packet formation along with user address and request input. The Network Gateway based Router packet formation is represented in Fig. 5. The packet is framed based on a request, destination address provided by the user, and gateway input. So The NoC is having a 13-bit packet which includes 1-bit request, 2-bit destination X address, 2-bit destination Y address, and 8-bit Gateway data.

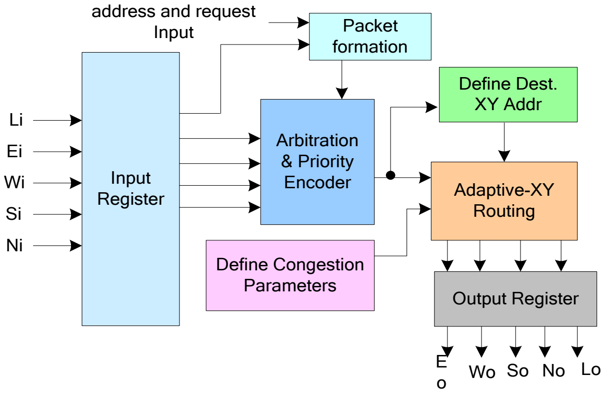
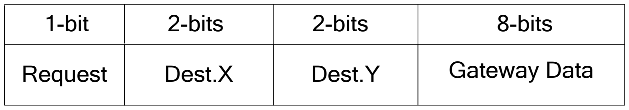


Fig. 4. Hardware Architecture of Network Gateway based Router.



1-bit	2-bits	2-bits	8-bits
Request	Dest.X	Dest.Y	Gateway Data

Fig. 5. Packet formation for Network Gateway based Router.

The framed packet, along with four more from input register is input to priority encoder. The priority encoder works based on the arbitration. The Arbiter receives the MSB bits from all the five ports and considered as requests and generates the 5-bit grants based on the priority. These grants are acts as a select line to priority encoder. The encoded data is a prioritize packet data, and it sends as an input to the adaptive routing-XY algorithm. Each Router, R1 to R16, has fixed 4-bit current XY address and which is easy to identify the Router. For example, in design, R4 is set to “0011," and R14 is “1101”.

To perform the routing computation, first, define the congestion parameters along with Destination-XY address from the encoded packet. The adaptive–XY routing is congestion-free routing and adaptive form of normal –XY routing [28]. The X or Y direction with less number of routing path is defined and the routing packet id assisted to the destination with less congestion. Based on congestion parameters, which finds the shortest routing path to reach the destination with less traffic. The Network Gateway based single router, 2X2, 3X3, and 4X4 NoC’s are designed and prototyped on FPGA, which are explained detail in the next section.

V. Results and Performance Analysis

The results and performance evaluation are analyzed in this section for Network Gateway (NG) Module, and NG Based NoC using Mesh topology. The NG and NG-Based NoC are designed using Verilog-HDL on Xilinx platform and implemented on Artix-7 FPGA.

A. Implementation Results

The Network gateway implementation results after a place and route process on Artix-7 FPGA are tabulated in Table I. The resources in terms of Area-Slices, LUT's, Design operating frequency, and total power utilized are represented. The NG utilizes 450 slice registers, 893 slice LUTs and operating at 319.642 MHz frequency. The NG utilizes 0.104W total power, which includes 0.022W dynamic power using X-power analyzer.

The NG Module is designed for NoC based Multiprocessing SoC applications. The NG Based Router is designed using Adaptive –XY routing algorithm. The different network sizes like 2X2, 3X3, and, 4X4 are designed using mesh topology. The Chip area utilization for NG Based NoC designs are represented in Table II. The graphical visualization of the NG based NoC designs for area utilization is represented in Fig. 6.

The total power (W) analysis of NG based NoC design with respect to Different clock frequencies are represented in Fig. 7.

The Power analysis results are generated using Xilinx X-Power analyzer and the ambient temperature, and the initial source voltage is set to 25^{o}C and 1Volt, respectively. The NG router and NG-4X4 NoC utilizes 1.032W and 1.10W respectively for 5000MHz clock frequency. The network gateway based NoC designs are implemented effectively on FPGA with better chip area, speed, and power tradeoffs have been achieved.

TABLE. I. Network Gateway Resource Implementation Results

Resources	**Utilized on Artix-7 FPGA**
Slice Registers	450
Slice LUTs	893
LUT-Flipflops	252
Max. Frequency (MHz.)	319.642
Total power (W)	0.104

TABLE. II. Resource Utilization–for NG-NoC Design

Area Utilization	**Network Gateway –NoC designs**			
	NG Router	***NG-2X2 NoC***	***NG-3X3 NoC***	***NG-4X4 NoC***
Slice Registers	470	603	823	1193
Slice LUTs	914	1011	1216	1551
LUT-FF pairs	273	365	561	891

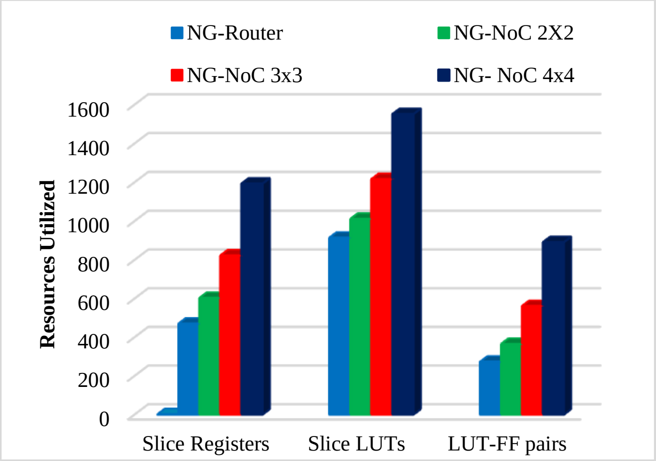


Fig. 6. Network Gateway-NoC Designs Area utilization on Artix-7 FPGA.

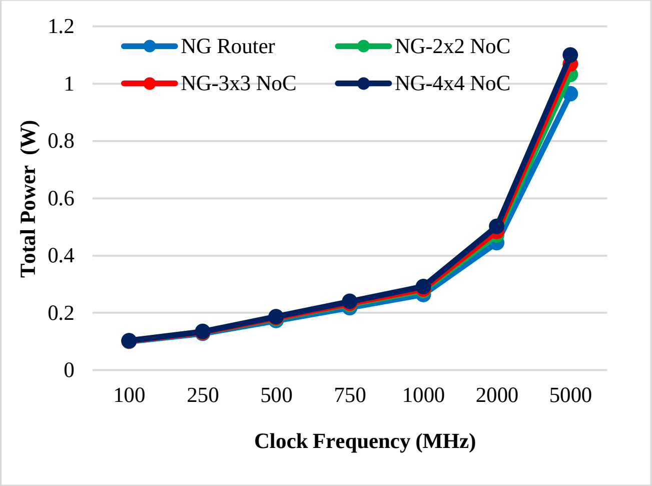


Fig. 7. Total Power v/s different Frequencies for NG based NoC Design.

B. Performance Evaluation

The performance analysis of this work is evaluated using average Latency and maximum Throughput with respect to input traffic. The wormhole switching method and uniform traffic patterns are considered for analysis purpose. The Packet Injection Rate (PIR) is defined as the total number of data packets that can be sent on a single clock cycle. The average latency for network gateway is calculated using below equation (1).

$$(\text{Avg. Latency})_{NG} = \text{Min.NG latency} + \text{No. of Flits} \qquad (1)$$

The minimum Network gateway (NG) latency in terms of clock cycles is 18.5. The number of flits used in the design is 8. So average latency for NG is 26.5 clock cycles. For the NG Based NoC design, the Average Latency for NG based NoC design is expressed in the below equation (2).

$$(\text{Avg. Latency})_{NoC} =$$

$$(\text{No. of PE's} * (\text{Avg. Latency})_{NG}) + (\text{No. of PE's} * 2) \qquad (2)$$

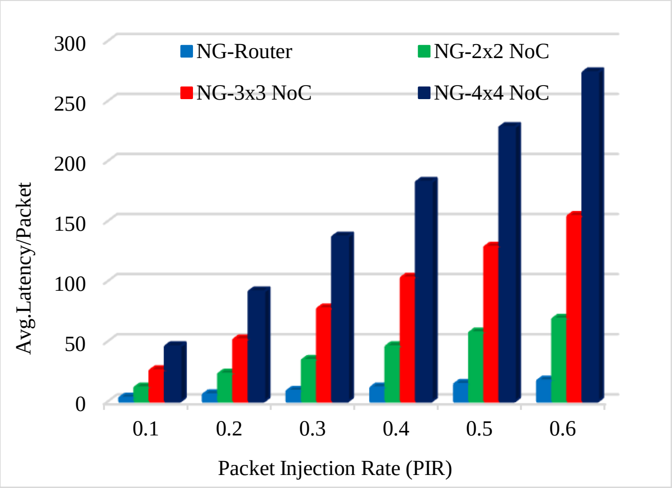


Fig. 8. Average Latency v/s Input Traffic for Network Gateway-NoC Designs using Mesh Topology.

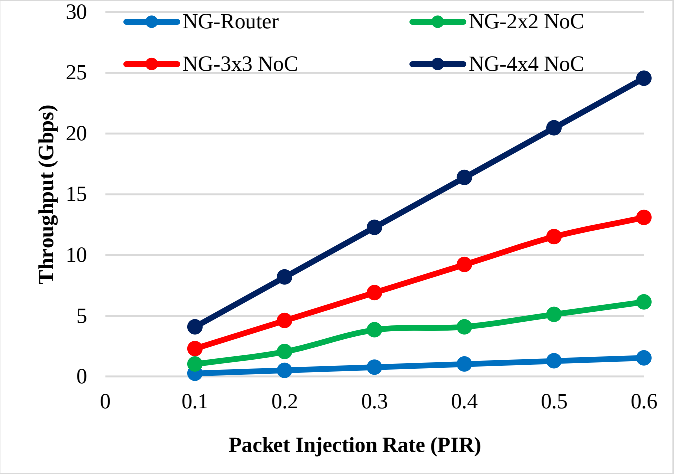


Fig. 9. Max.Throughput v/s Input Traffic for Network Gateway-NoC Designs.

The number of Processing elements (PE's) is defined based on Mesh Topology used in NoC. The 2 clock cycles are considered additionally, which is time taken to forward the packet from source to destination in NoC Network. The input traffic interms of PIR are evaluated in each NG –NoC designs are represented in Fig. 8 for average latency calculation. The average Latency for NG based NoC is represented in terms of Clock cycles (ns).

The maximum Throughput for NG based NoC is defined based on a number of PE's followed by Gateway data width, PIR, and Maximum operating frequency (MHz). And it is represented by using the below equation (3).

$$(\text{Throughput })_{NoC} = \text{No.of PE's} * \text{Datawidth} * \text{PIR} * F_{max} \qquad (3)$$

The PE's are connected NoC boundary via a network interface. The throughput calculation depends upon the data width used in the Gateway. For example, the 4X4 NoC with 8-bit data packet are connected 16 PE's and operated on maximum frequency (F_{max}) of gateway design used in artix-7 FPGA. The maximum Throughput with respect to Input traffic is represented in Fig. 9. The maximum Throughput of NG-Router and NG-4X4 NoC operated at 1.5342 Gbps and 24.548 Gbps respectively. The maximum Throughput varies based on data width selection. In design 8-bit data width is selected.

VI. Conclusion and Future Work

This manuscript presents an efficient and cost-effective Network Gateway (NG) model using Electronic crossbar switch along with Network gateway in NoC based system. The NG design is flexible to support multiple cores and easy to prototype on on-chip devices. The NG with a single Router and different sizes of NoC using mesh topology is designed using Adaptive XY routing. The NG implementation results on Artix-7 FPGA utilizes <1% hardware resources and NG based 4X4 NoC utilizes >2% resources. The NG operates at 319.6 MHz and consumes less total power around 0.104W on FPGA. The Performance analysis of NG based NoC is evaluated using Average Latency and Maximum Throughput with respect to different Input traffic. The average Latency for NG and NG Based 4X4 NoC design utilizes 15.9 and 273.6 at 0.6 PIR, respectively. The maximum Throughput for NG and NG Based 4X4 NoC design works at 1.53 Gbps and 24.54 Gbps at 0.6 PIR respectively for 8-bit data width. This architecture can be incorporated in futuristic researches with the security features to Network Gateway and NoC based systems to strengthen the data packets from attacks.

References

[1] Petracca, Michele, Keren Bergman, and Luca P. Carloni. "Photonic networks-on-chip: Opportunities and challenges." In 2008 IEEE International Symposium on Circuits and Systems, pp. 2789-2792. IEEE, 2008.

[2] Petrakis, Polydoros, Mohammed Abuteir, Miltos D. Grammatikakis, Kyprianos Papadimitriou, Roman Obermaisser, Zaher Owda, Antonis Papagrigoriou, Michael Soulie, and Marcello Coppola. "On-chip networks for mixed-criticality systems." In 2016 IEEE 27th International Conference on Application-specific Systems, Architectures and Processors (ASAP), pp. 164-169. IEEE, 2016.

[3] Hendry, Gilbert, Johnnie Chan, Shoaib Kamil, Lenny Oliker, John Shalf, Luca P. Carloni, and Keren Bergman. "Silicon nanophotonic network-on-chip using TDM arbitration." In 2010 18th IEEE Symposium on High-Performance Interconnects, pp. 88-95. IEEE, 2010.

[4] Hu, Yonghong, and Lu Ding. "Design and Realization of Multi-functional Gateway Based on Single Chip." In 2009 2nd International Congress on Image and Signal Processing, pp. 1-4. IEEE, 2009.

[5] Guo, Tiantian, Ming' a Zhou, and Qing Shen. "A Reconfigurable Intelligent Gateway for Heterogeneous Networks." In Proceedings of 2013 Chinese Intelligent Automation Conference, pp. 485-493. Springer, Berlin, Heidelberg, 2013.

[6] Guo, Xiaodong, and Haijun Ren. "Multimode communication gateway design in heterogeneous network environments, intelligent distribution, and utilization." In 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 326-329. IEEE, 2015.

[7] Shi, Wei, Gaofeng Lv, Zhigang Sun, and Zhenghu Gong. "HiTrans: An FPGA-Based Gateway Design and Implementation in HPC Environments." In International Conference on Algorithms and Architectures for Parallel Processing, pp. 561-571. Springer, Cham, 2015.

[8] Ni, Binbin, Mingguang Wu, and Yanpeng Liu. "Design of Embedded Home Network Gateway for CE Bus Based on ARM." In 2006 4th IEEE International Conference on Industrial Informatics, pp. 1380-1384. IEEE, 2006.

[9] Luo, Hong, Cheng Chang, and Yan Sun. "Advanced sensor gateway based on FPGA for wireless multimedia sensor networks." In 2011 International Conference on Electric Information and Control Engineering, pp. 1141-1146. IEEE, 2011.

[10] Nguyen, Minh-Triet, Lap-Luat Nguyen, and Tuan-Duc Nguyen. "On the design of gateway node for smart gird home network." In 2015 International Conference on Communications, Management and Telecommunications (ComManTel), pp. 57-61. IEEE, 2015.

[11] Shaofeng, Lin, Tao Bo, Pan Jin, Wan Juan, and Du Jia. "Design and Implementation of Embedded Wireless Gateway." In 2015 International Conference on Intelligent Transportation, Big Data and Smart City, pp. 270-273. IEEE, 2015.

[12] Zheng, Qi. "The design and the implementation of communication gateway between CAN bus and Ethernet." In 2015 IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), pp. 862-866. IEEE, 2015.

[13] Wang, Wenhua, Weiwei Xia, Rui Zhang, and Lianfeng Shen. "Design and implementation of gateway and server in an indoor high-precision positioning system." In 2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE), pp. 540-541. IEEE, 2014.

[14] Baoxia, Sun, Wang Weixing, Tie Fenglian, and Weng Jiangpeng. "Design and implementation of gateway for hybrid antenna clustering routing algorithm in paddy monitoring." In 2016 6th International Conference on Electronics Information and Emergency Communication (ICEIEC), pp. 310-313. IEEE, 2016.

[15] Zhong, Chang-Le, Zhen Zhu, and Ren-Gen Huang. "Study on the IOT architecture and gateway technology." In 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), pp. 196-199. IEEE, 2015.

[16] Shi, Yonghong, Jianzhong Shen, Lin Zhang, Qian Zhang, and Shaofeng Lin. "Design of Security Gateway Based On Dual-Homed Architecture." In 2016 International Conference on Robots & Intelligent System (ICRIS), pp. 159-163. IEEE, 2016.

[17] Kun Han, Junjie Liu, Demin Yang and Quan Yuan, "The design of secure embedded VPN gateway," 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA), Ottawa, ON, 2014, pp. 350-353.

[18] Ajami, Raouf, and Anh Dinh. "Embedded network firewall on FPGA." In 2011 Eighth International Conference on Information Technology: New Generations, pp. 1041-1043. IEEE, 2011.

[19] Abuteir, Mohammed, Romn Obermaisser, Zaher Owda, and Thierry Moudouthe. "Off-chip/on-chip gateway architecture for mixed-criticality systems based on networked multi-core chips." In 2015 IEEE 18th International Conference on Computational Science and Engineering, pp. 120-128. IEEE, 2015.

[20] Obermaisser, Roman, Zaher Owda, Mohammed Abuteir, Hamidreza Ahmadian, and Donatus Weber. "End-to-end real-time communication in mixed-criticality systems based on networked multi-core chips." In 2014 17th Euromicro Conference on Digital System Design, pp. 293-302. IEEE, 2014.

[21] Dumitru, Laurențiu A., Sergiu Eftimie, and Dan Fostea. "An FPGA-Based cloud storage gateway." In 2nd International Conference SEA-CONF, Academia Navală Mircea Cel Bătrân, Constanța. 2016.

[22] Lee, Trong-Yen, Chia-Wei Kuo, and I-An Lin. "High performance CAN/FlexRay gateway design for in-vehicle network." In 2017 IEEE Conference on Dependable and Secure Computing, pp. 240-242. IEEE, 2017.

[23] Shreejith, Shanker, Philipp Mundhenk, Andreas Ettner, Suhaib A. Fahmy, Sebastian Steinhorst, Martin Lukasiewycz, and Samarjit Chakraborty. "VEGa: A high performance vehicular Ethernet gateway on hybrid FPGA." IEEE Transactions on Computers 66, no. 10 (2017): 1790-1803.

[24] Zhou, Yuan, Wenping Xiao, Mingshan Liu, and Xiaokun Li. "Design of the embedded gateway for 4G and PROFIBUS-DP based on FPGA." In 2017 3rd IEEE International Conference on Computer and Communications (ICCC), pp. 748-752. IEEE, 2017.

[25] Korona, Mateusz, Krzysztof Skowron, Mateusz Trzepiński, and Mariusz Rawski. "FPGA implementation of IPsec protocol suite for multigigabit networks." In 2017 International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 1-5. IEEE, 2017.

[26] Urbina, Marcelo, Armando Astarloa, Jesús Lázaro, Unai Bidarte, Igor Villalta, and Mikel Rodriguez. "Cyber-physical production system gateway based on a programmable SoC platform." IEEE Access 5: 20408-20417, 2017.

[27] Kwak, Byeong Ok, and Tae Soo Chung. "Design and Implementation of the Trust Domain Gateway System." In 2018 International Conference on Information and Communication Technology Convergence (ICTC), pp. 925-927. IEEE, 2018.

[28] Guruprasad, S. P., and B. S. Chandrasekar, "Design and Performance analysis of Adaptive-XY over N-XY and DO Routing on FPGA Platform," IJRECE, pp. 2166-2171, Vol. 6 Issue 3, July-September, 2018.

Wireless Multimedia Sensor Networks based Quality of Service Sentient Routing Protocols: A Survey

Ronald Chiwariro[1]
Research Scholar, Computer Science Engineering
JAIN (Deemed-to-be University)
Bangalore, India

Dr. Thangadurai. N[2]
Professor and Research Coordinator
Department of Electronics and Communication Engineering
JAIN (Deemed-to-be University), Bangalore, India

Abstract—**Improvements in nanotechnology have introduced contemporary sensory devices that are capable of gathering multimedia data in form of images, audio and video. Wireless multimedia sensor networks are designed to handle such type of heterogeneous traffic. The ability to handle scalar and non-scalar data has led to the development of various real-time applications such as security surveillance, traffic monitoring and health systems. Since, these networks are an emergent of wireless sensor networks; they inherit constraints that exist in these traditional networks. Particularly, these networks suffer from quality of service and energy efficiency due to the nature of traffic. This paper presents the characteristics and requirements of wireless multimedia sensor networks and approaches to mitigate existing challenges. Furthermore, a review of recent research on multipath routing protocols and multi-channel media access protocols that have quality of service assurances and energy efficiency in handling multimedia data have included.**

Keywords—Quality of service; multipath routing; multi-channel media access control; energy efficiency

I. Introduction

Wireless Multimedia Sensor Networks (WMSNs) have enhanced the data gathering capability of the traditional Wireless Sensor Networks (WSNs) which were restricted only to gathering scalar data. WMSNs have sensor nodes equipped with cameras and microphones that enable these networks to gather multimedia data in various forms like live data streams, videos, audio, images and so on[1]. Recent advances in feature engineering, image-processing techniques, machine learning and communication technologies have given birth to various research to applications of WMSNs. Applications include health care industry, military and general surveillance systems, real time intelligent transportation systems and environmental monitoring [2], [3].

The WMSNs are, descendent of WSNs hence the same benefits such as self-organization, flexibility, disposition simplicity and scalability are also characteristic. However, the added features and capabilities in WMSNs present a number of challenges that are inherent with these constrained networks such as limited energy, storage, communication bandwidth as well as processing capacity. The large volumes of data generated by these multimedia networks to be reliably transmitted over the wireless medium in real-time further exacerbate these challenges. Research, on this domain, aims at the development of computation algorithms and protocols that are highly energy-efficient and Quality of Service (QoS) aware. Due to these variations, the solutions developed for WSNs cannot be directly applied to WMSNs. Therefore there is need to modify these techniques before they can be applied to WMSNs. Furthermore, new techniques at all layers from physical layer to application layer suitable for these networks are required. Surveys on such research ranging from hardware to the network model layers and other cross-layer designs are[1], [4]–[6]. Some extensive studies on various hardware and software architecture test beds are in [7]. Transport protocols designed to be reliable are in [8]. A comparison of energy efficient and QoS aware routing protocols is done in [9]–[11]. Accordingly, a review of QoS cognizant and multi-channel Media Access Control (MAC) protocols are in [12], [13]. AlSkaif et al. present a comparative study on WSNs MAC protocols investigating their suitability on WMSNs through the analysis of some network parameters on node energy drain [14]. Authors in [15], [16] identifies cross-layer optimization solutions to problems inherent in WMSNs packet delivery, energy preservation and error recovery. Discussions of security requirements in WMSNs and classification of the security threats as well as some protection mechanisms are in [22], [23]. Finally,[19] discusses energy-efficiency issues with regard to all sensor application designs as well as extension of network lifetime while [20] [21] proposes a classification of energy-efficient target tracking schemes according to sensing and communication subsystems on a particular node.

This survey will thus concentrate on the important aspects required to deliver QoS-aware routing protocols in WMSNs, thus energy-efficiency, real-time multimedia streaming and data volumes. The paper will also highlight challenges and proffered solutions to guide related research. Network designers and architects will also immensely benefit from the clarity on characteristics and requirements of WMSNs as well as existing solutions. Furthermore, presented is a survey of communication MAC and routing protocols with emphasis on energy-efficiency, scalability, QoS guarantee, prioritization schemes, multipath routing and service differentiation [17][18]. The conclusion will also give future directions on discussed issues.

The remaining paper is as follows: Section 2 highlights the characteristics and design requirements of WMSNs with design challenges and existing remedies. A study of multipath routing protocols is done in Section 3 followed by WMSNs proposed MAC protocols. Lastly, Section 5 presents conclusions to the survey.

II. Wireless Multimedia Sensor Networks

WMSNs are an emergent technology out of the traditional WSNs. As such, they inherit many constrains that exist in these networks as well as new challenges and requirements that come because of the requirement for real-time multimedia services and handling of increased volumes of data. The gathered data traffic handled by these networks requires delivery in real-time due to the nature of applications that require the data. Examples of such applications include security surveillance, health systems and traffic management systems. The multimedia data collected by the camera sensors is voluminous for a particular event; hence, bandwidth requirements for the transmissions are increased. As summarized in Table I, WMSNs have opened many doors to research due to their characteristics and capabilities. This section discusses the characteristics, design requirements of WMSNs as well as proposed approaches [24].

A. *Power Constraints*

The camera sensor nodes in WMSNs are generally battery-powered. The batteries are expected to power the sensor nodes for protracted periods without replacement. Therefore, the functionality of such nodes should take into cognizance these power constraints and limit energy consumption in its computations and communication. In traditional WSNs, energy drain due to computations can be insignificant compared to WMSNs where computations tend to consume extremely high energy. To, capture and processing of a simple frame in a vehicle tracking system can constitute up to 12% of total energy consumption of the overall event. It is therefore recommended to adopt energy-efficient algorithms in image processing [21] and likewise in video compression. Due to the large volumes of multimedia data to be transmitted, it is prudent that the communication protocols at every layer be energy-efficient. For example, the transport layer protocols reduce the number of control messages according to desired levels of reliability, with routing protocols employing load balancing and energy estimation techniques across the network and at the MAC layer protocols can avoid idle listening by inactive nodes. Dynamic power management is another important technique to be used as it ensures that idle components of a sensor node are selectively shutdown or hibernated to prevent unnecessary power consumption.

B. *Real-Time Multimedia Data*

In most applications involving multimedia data, QoS is difficult to achieve. Transmission of data to the sink without any packet loss or delays above a threshold is very crucial in WMSNs. Therefore there is a need to impose severe QoS demands on the networks. Delays cannot be tolerated in applications that involve multimedia data for example in security surveillance or traffic management systems. This implies that prioritization and service differentiation will play a pivotal role in these real-time systems. MAC protocols should give access or assign greater quality channels to higher priority data. Routing protocols need to select paths that will have the least delay to meet the required QoS as illustrated in. Reliability is also crucial in ensuring QoS to WMSNs. Retransmissions are done at transport layer for example in TCP while redundancy is at bit-level or at packet-level as presented in [8]. However, these methods must be used with consideration that they increase traffic hence consume more networks resources. The heterogeneous traffic in WMSNs that include multimedia and scalar data intended for different applications with varying QoS demands will require variable levels of priority even within the same traffic type.

TABLE I. Characteristics and Design Requirements of Wireless Multimedia Sensor Networks

Characteristics	Requirements	Design Approaches
Power Constraints	Energy efficiency	Energy-efficient computations • Image compression algorithms • Video compression algorithms Dynamic power management
		Energy-efficient computations • Transport Protocols • Routing protocols • MAC protocols
Real-time multimedia data	Quality of Service	Delay • Routing protocols • MAC protocols Reliability • Routing protocols [8] • MAC protocols
		Prioritization and service differentiation • Routing protocols • MAC protocols Local processing
Volumes of multimedia data	Reduction of data redundancy	Multimedia in-network processing • Multimedia data fusion • Multi-view video summarization Distributed source coding
		In-network data storage and query processing
	Higher Bandwidth Requirement	Multipath routing Multi-channel MAC protocols Ultra Wideband technique

C. Volumes of Multimedia Data

Typically, WMSNs have limited bandwidth hence transmission of large volumes of sensory data presents a major challenge to QoS guarantee. Techniques for data compression and redundancy reduction are vital to decrease data volumes prior to transmission. One such technique is local processing where on-board analysis of the captured images is used to extract only important events. The downside of local processing is the requirement for added hardware resources. Another technique is In-network processing of multimedia data that encompass data fusion where the sink node collects heterogeneous data from various nodes and create a summarized version of events to reduce data redundancy and enhance inferences. To deal with the resource limitation problems associated with centrally coding data from multiple sensor cameras, WMSNs use distributed source coding (DSC) where encoding of data is done independently at each sensor before transmission to the sink for decoding. This reduces the power consumption as well as required hardware resources. Typically, WSNs transmit all collected data to the sink for subsequent processing and querying. Due to technological advancements, it is now possible to equip sensors with processors and flash memory that enable them to process and store data. After processing, only analyzed data is transmitted to the sink. In terms of queries, only the result is sent to the network after querying historical data. However, proper data ageing schemes needs to be incorporated into the local databases as they fill up in order to maintain data integrity. It is also important to note that the sensors will form distributed databases which require efficient query engines to retrieve the data efficiently. Mitigating the bandwidth constraint that is extreme in WMSNs due to the large volumes and nature of traffic is also an important factor in achieving QoS communications. At the MAC layer, sensor nodes can communicate simultaneously using different channels. Data traffic can be routed through multiple paths. However, radio equipment that has considerable bandwidth such as ultra-wideband (UWB) can be utilized in WMSNs.

III. Quality of Service Aware Multipath Routing Protocols for WMSNs

Routing techniques for WSNs have been extensively studied over the years to improve communications. However, these techniques cannot be directly implemented in WMSNs due to variations with traditional WSNs. Routing in WSNs aims at finding the shortest path for transmission scalar data. Applying the same routing concepts to large volumes of multimedia data will result in network congestions and increased power drain on nodes. Therefore, the robust approach will be to send data in parallel through multiple paths. Routing in WSNs is particularly concerned with energy-efficiency whilst WMSNs also consider the QoS due to real-time traffic and reliability concerns.

This section presents some multipath routing protocols in WMSNs with QoS assurances. This survey looks at different protocols than those recently surveyed in. Furthermore, the chosen multipath routing protocols have single path routing support. For further comparison of the surveyed multipath routing protocols with QoS assurances, particularly to WMSNs refer to Table II.

TABLE. II. Comparison of Multipath Routing Protocols under Review

Protocol	**DGR**	**AntSensNet**	**Z-MHTR**	**GEAM**	**LCMR**
Routing Method	geographic routing	ant colony based routing	ZigBee cluster tree routing	geographic routing	ad-hoc on-demand distance vector routing
Routing Metric	geographic distance and deviation angle	pheromone value of residual energy, delay, packet loss rate and available memory	network address	geographic distance	end-to-end delay
Routing States	one hop neighbour table	one hop neighbour table, routing pheromone table	one hop neighbour table, tree branches used in routing and/ or interfering node table	one hop neighbour table, district information	routing table
Disjoint Paths	yes	yes	yes	yes	no
QoS Metrics	reliability and throughput	reliability, delay, throughput	throughput	throughput	delay
Path Recovery	yes	yes	no	yes	no
Scalability	good	good	good	good	poor
Congestion Control	no	yes	no	no	no
Prioritization	no	yes	no	no	no
Service Differentiation	no	yes	no	no	no
Energy Efficiency	medium	medium	good	good	poor
Clustered	no	yes	yes	no	no
Interference Aware	no	no	yes	yes	no
Year	2007	2010	2014	2013	2017

A multipath routing protocol based on ant colony optimization called AntSensNet with QoS assurances is presented in. It has three phases of operation: Formation of the cluster, route discovery phase, data transmission and route maintenance. The cluster formation is initiated by the sink that releases some cluster ants (CANTs). Those within close proximity to the sink are selected as cluster heads (CH) and will receive the CAs first. Upon receiving the CANTs, they will be responsible for the reduction of the time-to-live (TTL). The cluster head will then advertise the CANTs to non-cluster heads within its communication radius so that those who are willing to join the cluster can join. Once clusters are formed, the CH begins route discovery. Each CH manages a pheromone table and shares with its neighbors according to traffic classes following four parameters i.e. Energy, packet drop, memory and delay. Traffic specific paths to the sink is created by broadcasting a forward ant (FANT) which will collect traversed node identities and the four parameters (queue delay, ratio of packet, residual energy and available memory) as it propagates. When a node receives a FANT, an update is done to its information before sending it to the next hope that satisfies the QoS requirements and a corresponding backward ant (BANT) is transmitted in the reverse path for path reservation. On receipt of the BANT, nodes update their pheromone tables. For establishment of multiple paths for video transmission, a video forward ant (VFANT) is broadcasted in the same manner as the FANT and the sink responds by sending multiple VBANTs. The VBANTs will be used to choose paths for sending video data. Once routes are ready data, delivery starts. A maintenance ant (MANT) is used for route maintenance. This protocol gives differentiated service to ensure QoS delivery by offering each traffic separate routes. The use of cluster heads is a drawback on scalability. However, the multipath routing technique is viable for video data only.

Z. Bidai et al. proposed the ZigBee Multipath Hierarchical Tree Routing (Z-MHTR) protocol. It allows source to use non-parent neighbors to search for other paths. The source node maintains a record of all branches used for tree routing (TR). The source node will construct disjoint paths using three basic principles. If a selected next hop node branch has not been utilized for TR path by the source node then a node disjoint can be established from that node to the sink using TR. If the branch has already been utilized for TR path by the source then the next hop will depend upon the depth of a node common to the TR path used by the source and the node that has used the node branch for TR. If all neighbor’s branches have been utilized in TR then it selects the neighbor node that is not in any TR path. The rules are applied to any subsequent nodes until the sink. The number of disjoint paths corresponds to the number of branches forming the topology. Furthermore, the author proposed for reduction of interference in which nodes lists interfering neighbors except the ones on the same paths. This is done by checking whether they can hear data packets that are not destined to them. The disjoint paths that reduce inter-path interloping are preferred. Based on the ZigBee tree topology and address assignment, multipath routing is achieved through neighbor table and a record of routing tree usage on a particular branch. The further work mitigates multiple paths interferences caused by route coupling. However, the restriction is only to ZigBee tree topology hence the paths are proportionate to available branches.

M. Chen et al. recommended the directional geographical routing (DGR) protocol for real-time video communications.The nodes in this protocol implement the global coordinate system to create virtual coordinates upon receipt of a broadcast probe. The virtual coordinates are obtained by mapping the source and sink position along the x-axis to the destination or intermediate node. A node is selected to be a forwarding candidate only if it falls within the transmission range, the optimal mapping location and the threshold of the source. Next hop will be a candidate that has the smallest distance to the optimal mapping hence; it will have a smaller timer than other competing nodes. If a timer expires, the node sends a reply message REP to the source. On receipt of an REP, the source confirms with SEL message. Nodes that hear the REP or SEL cancels their timers. The winner node will not establish any other path to the same source in order to guarantee path disjointedness. In turn, the connected node will send its own probing messages following the same procedure with an adjusted deviation angle to create a path towards the sink. For establishment of multiple paths, the source will send a number of probe messages with variations in the initial deviation angle. For video routing, the source broadcasts the complete frame initially to all single hop neighbors. Those neighbors within the chosen paths will retransmit the video using respective paths only those packets specified by the source. The packet delivery in this protocol is fast and reliable through multipath and the forwarding equivalence class. It also scales well due to the stateless geographic based routing paradigm. However, if a node fails, the path recovery takes longer as well as the new route discovery. In addition, it considers only a single active source for video transmissions that might not be practical in some scenarios.

A. Bhattacharya and K. Sinha following the principles of ad-hoc on-demand distance vector routing (AODV) developed the least common multiple routing (LCMR) protocol. As opposed to calculating the shortest path by number of hops, it uses the routing time taken or end-to-end delay to choose multiple paths. During route discovery, the route reply message RREP has to arrive before the deadline otherwise it will not be accepted. The source node uses the RREP message to check the routing time taken by the corresponding route request message RREQ before reaching the destination. From the accepted x paths that have routing time $\{T_1, T_{2, \ldots,} T_x\}$, it calculates the least common multiple L of $\{T_1 ,T_{2, \ldots ,} T_x\}$. The packets sent over path i are decided such that $= \sum_{i=1}^{x} L/Ti$ packets, L/T$_i$ packets will be routed along that path i. The total time it takes to deliver k packets gives the maximum routing time T$_{max}$of $\{T_1 ,T_{2, \ldots ,} T_x\}$. This protocol ensures avoidance of congested routes through the end-to-end calculation of routing time during its route discovery process. In order to reduce the transmission time, the number of packets allotted to a particular route is reduced according to time L and the routing time T$_i$ of the path. However, this may lead to early node death if most traffic is continuously routed through a node with least end-to-end routing time. Adaption to congestion and route breakage needs improvement.

Unlike DGR, that uses the deviation angle for controlling the directions of multiple paths, Li et al proposed the division of the topology into different districts for specific paths using the geographic energy-aware non-interfering multipath routing (GEAM) protocol. After division into virtual coordinates just like in DGR, the source and sink areas are restricted within the transmission radius. Each packet is piggybacked with boundary information of the selected district by the source before transmission. The subsequent nodes will then use greedy perimeter stateless routing (GPSR) to forward the packet to the respective district. For load balancing and even distribution of energy, GEAM the data transmissions are organized in runs of same lengths. To further avoid interference within multiple routing paths, it applies division of runs into three rounds, where a district D_x belongs to round k if $D_x\%3 = k$. During the first run, loads are distributed evenly to all districts. After each run, the sink collects residual energy from all nodes within a district and sends back to the source. Based on these statistics the source adjusts the rate of utilization for every district and those with higher energy levels get more loads in the next run. GEAM achieves balanced traffic loads and energy consumption as well as avoids interference by the division to the topology into various districts. Scalability is also guaranteed using GPSR. However, piggybacking every packet with border information and making it collect network statistics increases the overhead. It also does not consider some QoS metrics such as delay and reliability that are of paramount importance to delivery of multimedia data.

IV. Quality of Service Aware Media Access Control Protocols for WMSNs

MAC protocols present a challenge during their design and implementation when aiming for energy efficiency and coordinating transmission of large volumes of multimedia sensory data and meeting QoS in MWSNs. The dynamic and burst traffic predominant in WMSNs it requires application of duty cycling techniques in saving energy deeper analysis. Reduction of collisions is also an important factor in MAC protocol design especially when it involves real-time multimedia data. Controlling media access through prioritization and differentiation of services is also an important factor when handling heterogeneous traffic. This section will elaborate some of the energy-efficient MAC protocols that have QoS assurances. A summary of the same is in Table III.

M. Arifuzzaman et al. proposed the intelligent hybrid MAC (IH-MAC) protocol. This protocol combines CSMA/CA and TDMA techniques as a single mechanism that implements local synchronization. The protocol prioritizes the node holding data with high QoS such as real-time data. If nodes have same priority and mapped to same slot, then they contend for that slot. For energy preservation, it adjusts its transmission output during the contentions. The protocol scales well and reduces collisions as well as improves on channel utilization and access delays that are challenges in CSMA/CA by fusion of CSMA/CA and TDMA.

TABLE. III. Comparison of Media Access Control Protocols

Protocol	**EQ-MAC**	**Saxena**	**Diff-MAC**	**MQ-MAC**	**IH-MAC**	**AMPH**	**PA-MAC**
MAC Mechanism	hybrid of CSMA/ CA and TDMA	CSMA/ CA	CSMA/ CA	IEEE 802.15.4	hybrid of CSMA/ CA and TDMA	hybrid of CSMA/ CA and TDMA	IEEE 802.15.4
Synchronization	global, precise	not required	not required	local, precise	local, precise	global, precise	global, precise
QoS Guarantee	delay	throughput, delay	reliability, delay	reliability, delay	delay	reliability, delay	throughput, delay
Prioritization Scheme	traffic types	traffic types	traffic types, traversed hop count of packets	traffic types, packet lifetime	traffic types	traffic types, dynamic	traffic types
Service Differentiation Scheme	dynamic slot allocation	adaptive contention window, dynamic duty cycle	adaptive contention window, dynamic duty cycle, weighted fair queueing	dynamic channel allocation, dynamic slot allocation, adaptive contention window	adaptive contention window, dynamic slot allocation	adaptive contention window, dynamic slot allocation	dynamic channel access control
Scalability	poor	good	good	medium	medium	poor	poor
Adaptation to Dynamic Traffic	good	medium	medium	poor	good	good	poor
Collision Rate	low	medium	medium	low	low	low	high
Energy Efficiency	good	medium	medium	good	medium	poor	good
Message Passing	no	no	yes	no	no	yes	no
Clustered	yes	no	no	yes	yes	no	no
Year	2008	2008	2011	2015	2013	2014	2016

An energy-efficient hybrid MAC scheme (EQ-MAC), was proposed by Yahya and Ben-Othman. It uses the cluster mechanism in which the cluster head schedules slots using TDMA. It uses frames for communication. The cluster head sends the initial broadcast frame for synchronization. Once synchronization is completed, the cluster members start transmission of data through the cluster head. The cluster head issues TDMA slots upon request from the cluster members with consideration of traffic priorities. The cluster head then broadcasts allocated TDMA slots to cluster members for transmissions to begin. Sleep mechanism will also apply to those cluster members without data to transmit. Real-time data is placed in a queue that is served instantaneously. The sleep mechanism saves energy and channel utilization. The protocol assures delivery of real-time data especially multimedia due to prioritization of traffic. However, this may starve low priority traffic.

An efficient QoS provisioning protocol by M. Souil (AMPH), is a hybrid channel access method. The notable difference between AMPH and IH-MAC is that the latter is CSMA/CA centered and AMPH is TDMA centered. AMPH divides transmissions into slots and two-hop radius for each node. Prioritization for medium access is done by separation of real-time and best effort traffic and based on slot ownership. Contending nodes are separated into four groups according to traffic priority: real-time by owner, real-time by non-owner, best effort by owner and best effort by non-owner. To avoid starvation, the protocol allows best effort traffic ahead of real-time traffic in limited slots per cycle. To conserve energy, it allows nodes to switch of their radios in the waiting state. The use of any slot coupled with traffic prioritization achieves optimum channel utilization and QoS guarantees to heterogeneous traffic. However, there is need for a robust differentiation of traffic that caters for more traffic types that exist in WMSNs.

Bhandari et al. proposed a multi-channel priority based adaptive MAC protocol (PA-MAC) that is based on the IEEE 802.15.4 standard. The protocol traffic classification is grouped into four categories according to priority: emergency (medical), on-demand, normal, non-medical. It uses the contention access periods (CAP) for the four classifications of traffic. Traffic with higher priority is allowed access to slots for lower priority traffic and the lower priority traffic transmits during the contention free period (CFP). The nodes enter into sleep until next transmission. Collisions are mitigated by traffic differentiation and transmission of lower priority data (e.g. multimedia data in medical scenario) at CFP. However, the protocol gives less priority to multimedia data hence it cannot be applied directly to WMSNs.

Related CSMA/ CA based protocols with QoS assurances were proposed by Saxena et al. and Diff-MAC. The protocols use adaptive contention window (CW) and dynamic duty cycling mechanisms. The CW sizes for real-time traffic are set to be less than low priority traffic. The protocols differ in that, Saxena et al. aims for fairness by making sensors adjust their CW size after checking with neighboring sensors if chances of a collision remain after last CW size changes whereas sensors in Diff-MAC continue to change their CW sizes towards the threshold CW size. Diff-MAC also employs the hybrid weighted fair queuing (WFQ) technique to allow channel access to real-time traffic while Saxena et al. uses a FIFO mechanism. Diff-MAC avoids starvation to same traffic type by prioritization of packets belonging to the same queue prioritizing them based on traversed hops. It further segments video frames and transmit the in bursts to lower retransmission cost. Both protocols use the dynamic duty cycle technique. The protocols offer good QoS, fairness and energy-efficiency in WMSNs. However, constantly monitoring of various states in a network leads to idle listening and as for Diff-MAC, the constant intra-queue prioritizations may not scale well with high traffic.

MQ-MAC is a cluster based slotted CSMA/ CA MAC protocol. The cluster head is responsible for key responsibilities that include channel sensing, time slot allotments and channel allocation. It divides its super frame into active and sleep periods, with the active being sub-divided into three phases namely; sensing, channel selection and data transmission requests. Once the cluster head receives results of channel sensing and transmission requests from the cluster members, it will allocate slots and transmission channels. QoS is guaranteed through slot allocation. The requests once received from cluster members are classified according to arrival time and traffic type as well as consideration of the packet lifetime. Early slots are allocated to requests with higher priority. The slot allocations are allows data traffic from cluster members to the cluster head to be collision free. After the transmission phase, the sensor nodes will sleep and wake up when another super frame starts. QoS is guaranteed through allocation of slots and channels for different traffic types according to priority. However, the presence of many control messages during sensing and switching are not desirable due to overheads.

V. Conclusion

WMSNs are becoming more popular in various IoT applications due to their ability to handle heterogeneous data from various sensory devices. Considerable research has been done to enhance these networks. However, some challenges are still prevalent due to the distinctive characteristics of the WMSN and resource constraints. This paper covered the unique characteristics and requirements for WMSNs as well as some design approaches to the constraints highlighted. Furthermore, the survey includes multipath routing protocols and MAC protocols, which are two important communication parameters to improve QoS provision in any network. Multipath routing is significant to the provision of QoS and delivery of multimedia data in WMSNs. These protocols are able to distribute the voluminous multimedia data across the network, thus balancing the load as well as energy consumption. It is important for the protocols to counter interference in multiple parallel paths to avoid route coupling issues. However, most multipath routing protocols consider load balancing and energy management without due diligence for other QoS metrics such as prioritization, and differentiation of service. Traffic in these networks is heterogeneous in nature therefore prioritization and service differentiation should not only be fixed to a particular type of traffic as in the case with most protocols that dedicate only to video traffic. Route recovery and congestion control must also be given great significance to improve QoS in WMSNs. Finally, Efficient MAC protocols intended for WMSNs must be able to handle heterogeneous traffic and vast volumes of multimedia data. In literature, there exist CSMA/ CA based MAC protocols that are scalable and adapt to different variable traffic situations although suffer bottlenecks in QoS provision and energy efficiency. Hybrid protocols combining CSMA/ CA and TDMA are promising to be an important part of WMSNs since CSMA/ CA and TDMA are used to handle low data rates and high data rates respectively thereby improving throughput and the reduction of collisions. It is important to note that QoS is of great importance to WMSNs hence the future research focus should be on handling multimedia data collected by camera sensors.

References

[1] T. Almalkawi, M. G. Zapata, J. N. al-Karaki, and J. Morillo-Pozo, “Wireless multimedia sensor networks: Current trends and future directions,” Sensors, vol. 10, no. 7, pp. 6662–6717, 2010.

[2] T. Semertzidis, K. Dimitropoulos, A. Koutsia, and N. Grammalidis, “Video sensor network for real-time traffic monitoring and surveillance,” IET Intell. Transp. Syst., vol. 4, no. 2, p. 103, 2010.

[3] N. B. Bo et al., “Human mobility monitoring in very low resolution visual sensor network,” Sensors (Switzerland), vol. 14, no. 11, pp. 20800–20824, 2014.

[4] S. Soro and W. Heinzelman, “A Survey of Visual Sensor Networks,” Adv. Multimed., vol. 2009, pp. 1–21, 2009.

[5] A. Sharif, V. Potdar, and E. Chang, “Wireless multimedia sensor network technology: A survey,” IEEE Int. Conf. Ind. Informatics, no. May 2014, pp. 606–613, 2009.

[6] I. F. Akyildiz, T. Melodia, and K. R. Chowdury, “Wireless multimedia sensor networks: A survey,” IEEE Wirel. Commun., vol. 14, no. 6, pp. 32–39, 2007.

[7] I. F. Akyildiz, T. Melodia, and K. R. Chowdhury, “Wireless multimedia sensor networks: applications and testbeds,” Proc. IEEE, vol. 96, no. 10, pp. 1588–1605, 2008.

[8] M. A. Mahmood, W. K. G. Seah, and I. Welch, “Reliability in wireless sensor networks: A survey and challenges ahead,” Comput. Networks, vol. 79, pp. 166–187, 2015.

[9] M. Radi, B. Dezfouli, K. A. Bakar, and M. Lee, “Multipath routing in wireless sensor networks: Survey and research challenges,” Sensors, vol. 12, no. 1, pp. 650–685, 2012.

[10] A. M. Zungeru, L.-M. Ang, and K. P. Seng, “Classical and swarm intelligence based routing protocols for wireless sensor networks: A survey and comparison,” J. Netw. Comput. Appl., vol. 35, no. 5, pp. 1508–1536, Sep. 2012.

[11] S. Ehsan and B. Hamdaoui, “A Survey on Energy-Efficient Routing Techniques with QoS Assurances for Wireless Multimedia Sensor Networks,” IEEE Commun. Surv. Tutorials, vol. 14, no. 2, pp. 265–278, 2012.

[12] O. D. Incel, “A survey on multi-channel communication in wireless sensor networks,” Comput. Networks, vol. 55, no. 13, pp. 3081–3099, Sep. 2011.

[13] M. A. Yigitel, O. D. Incel, and C. Ersoy, “QoS-aware MAC protocols for wireless sensor networks: A survey,” Comput. Networks, vol. 55, no. 8, pp. 1982–2004, 2011.

[14] T. AlSkaif, B. Bellalta, M. G. Zapata, and J. M. BarceloOrdinas, “Energy efficiency of MAC protocols in low data rate wireless multimedia sensor networks: A comparative study,” Ad Hoc Networks, vol. 56, pp. 141–157, Mar. 2017.

[15] D. G. Costa and L. A. Guedes, “A survey on multimedia-based cross-layer optimization in visual sensor networks,” Sensors, vol. 11, no. 5, pp. 5439–5468, 2011.

[16] N.Thangadurai, Dr.R.Dhanasekaran and R.D.Karthika, "Dynamic Energy Efficient Topology for Wireless Ad hoc Sensor Networks", WSEAS Transactions on Communications, Vol. 12, Iss. 12, pp. 651-660, 2013.

[17] L. D. P. Mendes and J. J. P. C. Rodrigues, “A survey on cross-layer solutions for wireless sensor networks,” J. Netw. Comput. Appl., vol. 34, no. 2, pp. 523–534, 2011.

[18] N. Thangadurai and Dr.R.Dhanasekaran, “Energy Efficient Cluster based Routing Protocol for Wireless Sensor Networks”, International Journal of Computer Applications, Vol. 71, No. 7, pp. 43-48, 2013.

[19] N.Thangadurai, Dr.R.Dhanasekaran and R.D.Karthika, "Dynamic Traffic Energy Efficient Topology based Routing Protocol for Wireless Ad hoc Sensor Networks", International Review on Computers and Software, Vol. 8, No. 5, pp. 1141-1148, 2013.

[20] T. Winkler and B. Rinner, “Security and Privacy Protection in Visual Sensor Networks,” ACM Comput. Surv., vol. 47, no. 1, pp. 1–42, 2014.

[21] M. Guerrero-Zapata, R. Zilan, J. M. Barceló-Ordinas, K. Bicakci, and B. Tavli, “The future of security in Wireless Multimedia Sensor Networks : A position paper,” Telecommun. Syst., vol. 45, no. 1, pp. 77–91, 2010.

[22] T. Rault, A. Bouabdallah, and Y. Challal, “Energy efficiency in wireless sensor networks: A top-down survey,” Comput. Networks, vol. 67, no. March, pp. 104–122, 2014.

[23] O. Demigha, W. K. Hidouci, and T. Ahmed, “On Energy efficiency in collaborative target tracking in wireless sensor network: A review,” IEEE Commun. Surv. Tutorials, vol. 15, no. 3, pp. 1210–1222, 2013.

[24] S. Bhandari and S. Moh, “A priority-based adaptive MAC protocol for wireless body area networks,” Sensors (Switzerland), vol. 16, no. 3, 2016.

Embedded System Interfacing with GNSS user Receiver for Transport Applications

Mohmad Umair Bagali[1]
Research Scholar
Department of Electronics and Communication Engineering
JAIN (Deemed-to-be University), Bangalore, India

Dr. Thangadurai. N[2]
Professor and Research Coordinator
Department of Electronics and Communication Engineering
JAIN (Deemed-to-be University), Bangalore, India

***Abstract*—The real time vehicle movement traces using waypoint display on the base-map with IRNSS/NavIC and GPS dataset in the GUI simultaneously. In this paper, a portable electronic device with application software has been designed and developed, which would be used to capture the real time positional information of a rover using IRNSS-UR. It stores the positional information into database and displays the real time vehicle positional information like date, time, latitude, longitude and altitude using both GPS and IRNSS/NavIC receiver simultaneously. The designed hardware device with an application software developed helps in mapping the real time vehicle / rover movement at the same time which also helps in identifying the region with data loss, varying positional information, comparing the distance travelled by rover and also aid in retrieving the past surveys and mapping the traces of both IRNSS and GPS simultaneously. The vehicle movement using both IRNSS/NavIC and GPS are tracked on the base map to find the similarity and differences between two. During this research work it can be conclude that that the rover position using GPS and IRNSS were accurate and continuous in our survey duration except in few places. In that few places the data loss is observed because of the satellite visibility variations. For Indian region the IRNSS/NavIC can be a better replacement for GPS.**

***Keywords*—GNSS; GPS; IRNSS; embedded systems**

I. Introduction

NavIC/IRNSS is a free provincial regional satellite constellation being developed by India. IRNSS will give two kinds of administration services, to be specific, Standard Positioning Service (SPS) which is given to every one of the customers and Restricted Service (RS), which is an encoded administration given distinctly to the approved clients only. It is intended to give precise position data administration to customers in India just as the locale extending out up to 1500 km from its limit, which is its essential administration region. An all-inclusive administration territory lies between essential administration region from Latitude 30^0 S to 50^0 N and Longitude 30 degree East to 130 degree East [2]. The IRNSS system is relied upon to give a position precision of superior to 10 m in the essential administration region [5].

The NavIC system comprises a ground segment which is supported by a space segment.

Space segment: 8 satellites constellation, where 4 satellites are in Geosynchronous Orbit (GSO) crossing the equator at 55° East and two at 111.75° East and three satellites are located approximately 36,000 km (22,000 mi) above earth surface in Geostationary Orbit (GEO) at 83° East, 32.5° East, and 131.5° East longitude. The movement of four GSO satellites will be in the form of figure of “8” as shown in Fig. 1.

Ground segment: The maintenance and operation of the IRNSS constellation is held in ground segment (Fig. 2) are:

1) IRNSS Spacecraft Control Facility (IRSCF)
2) ISRO Navigation Centre (INC)
3) IRNSS Range and Integrity Monitoring Stations (IRIMS)
4) IRNSS Network Timing Centre (IRNWT)
5) IRNSS CDMA Ranging Stations (IRCDR)
6) Laser Ranging Stations
7) IRNSS Data Communication Network (IRDCN)

CDMA extending is being completed by the four IRCDR stations all the time for all the NavIC satellites. The INC built up at Byalalu performs remote tasks and information accumulation with all the ground stations. The IRNWT has been set up and is furnishing IRNSS framework time with an exactness of 2 ns (2.0×10−9 s) (2 sigma) with respect to UTC. 14 IRIMS are right now operational and are supporting NavIC operations. Laser extending is being completed with the help of ILRS stations the world over. Route Software is operational at INC since 1 August 2013. The IRDCN has built up earthly and VSAT connects between the ground stations. Seven 7.2 m FCA and two 11 m FMA of IRSCF is right now operational for LEOP and on-circle periods of IRNSS satellites [3]. All the route parameters viz. satellite ephemeris, clock revisions, trustworthiness parameters and auxiliary parameters viz. iono-postpone remedies, time counterbalances with respect to UTC and different GNSS, chronological registry, instant message and earth direction parameters are created and uplinked to the shuttle naturally [1].

Signal: Standard Positioning Service are modulated by a 1 MHz BPSK signal and a Precision Service will use BOC (5,2) are the two NavIC signals consists of S1 band (2492.028 MHz) and L5 (1176.45 MHz). An informing or data interface is implanted in the NavIC framework. The navigation signals themselves would be transmit in the S band recurrence (2–4 GHz) and communicate through a staged exhibit reception apparatus i.e., antenna to keep up required inclusion and signal quality. This element permits to send admonitions to a particular geographic region. For instance, fishermen utilizing the framework can be cautioned about a cyclone [3].

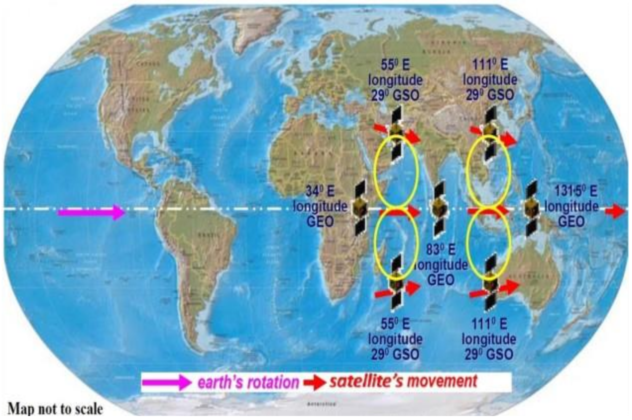


Fig. 1. NavIC/ IRNSS Constellation.

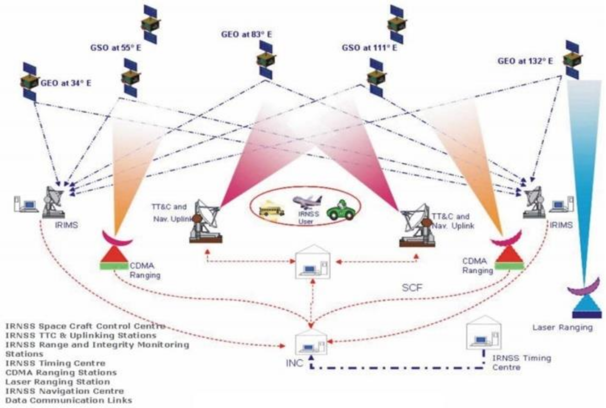


Fig. 2. IRNSS Operation Model [ISRO].

Accuracy: The framework is expected to give a complete position exactness of superior to anything 10 meters all through Indian landmass and superior to anything 20 meters in the Indian Ocean just as an area broadening roughly 1,500 km (930 mi) around India. NavIC has dual frequencies i.e., S1 and L5 bands whereas GPS which is dependent on L band only. At the point when low-recurrence sign goes through the climate, its speed changes because of air unsettling influences. For India's situation, the real delay is evaluated by estimating the distinct differences in two frequency i.e., L and S band. Thusly, NavIC isn't subject to any model to discover the recurrence mistake and is more precise or near to the GPS [6][8][10].

Major applications of NavIC are: Disaster Management, Terrestrial, Precise Timing, Aerial and Marine Navigation, Mapping, Vehicle tracking, Integration with mobile phones, Voice and visual navigation for drivers and fleet management, Terrestrial navigation aid for hikers and travelers, and Geodetic data capture.

There are various challenges involved in replacing the existing GPS navigation system with autonomous regional navigation system. In this paper aims at plotting or mapping the rover position in real time using both GPS and NavIC receivers. The purpose of real time mapping of rover position is to identify the geographical region with similarity and differences between GPS and NavIC navigational systems [7] [9].

This paper covers these contents in the following chapters: That is about the field survey carried out by the research team, Application software and Hardware development, Observation and Analysis, Conclusion and Future work.

II. Field Surveys

A total of five field surveys had been conducted. The survey had been conducted on a rover with IRNSS-UR, antenna and Laptop with IRNSS-UR application installed inside the vehicle shown in Fig. 3. IRNSS/ NavIC system at the user end identifying the following parameters are (a) the variations in experimental setup when the rover is on move at different speeds due various factors like vibrations and soon; (b) no vehicle movement trace / marker on GUI provided by ACCORD during its movement like Google navigation map.

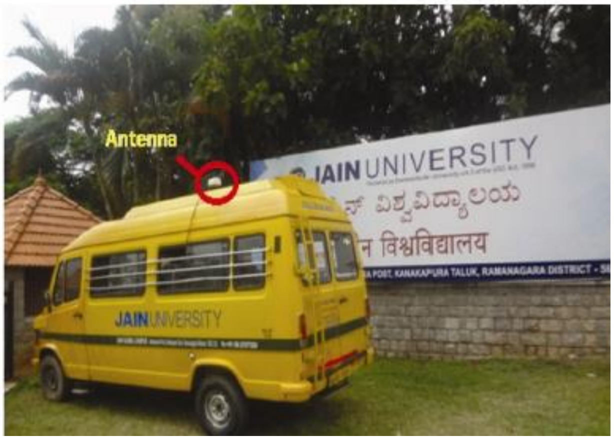


Fig. 3. Antenna Setup in Mobile Vehicle.

The hardware device which basically consists of Raspberry pi board and other peripherals have been developed an application in order to read NMEA (National Marine Engineering Association) data from the IRNSS-UR at 1 Hz.

The application is programmed to read the NMEA from the IRNSS-UR receiver, parsed the NMEA data, storage and display colored marker on GUI to indicate the current vehicle position. The hardware and application has been tested during this survey and noted the bugs / issue / short comings of both hardware and software.

Few of the bugs encountered during this survey are:

1) The device was able to read data but the delay for data reading was incremental and that leads to incremental time differences between current time and time read by the device.

2) When there is switch between poor signals strength to good signal strength (vice-versa) the parsing of NMEA data was incorrect.

Two separate applications developed to fix the problem of -“NMEA data reading at incremental delay which causing loss of data”.

1) One of the applications is meant for reading NMEA data, saving data in separate text file with auto count. The rate of creating the text file is fixed at the rate of every 30 seconds (new text file / 30 sec).

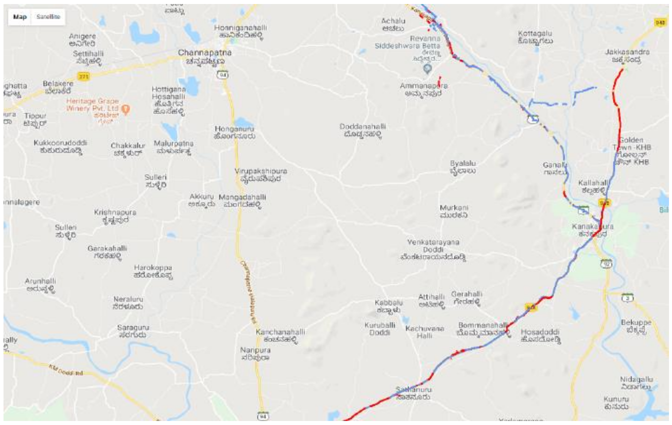


Fig. 4. Survey Data Plotted from NavIC and GPS.

2) The second application reads the NMEA data from these files and displays the details like date; time; lat; long; alt; positions of both IRNSS/ NavIC (blue color) and GPS (red color) separately on display shown in Fig. 4. The IRNSS and GPS traces are also displayed on the base maps to watch / trace the movement of vehicle in real time.

III. Application Software Development

The NavIC receiver interface and to retrieve the real-time positional data (latitude, longitude, altitude) in the form of waypoints, satellite information like Number of Satellites, Relative accuracy, number of SVs in view, time zone.etc. The data retrieved is a set of values such that for every IRNSS value there is a corresponding GPS value. The data retrieved is then used to display the information on GUI like distance travelled, delta value of waypoints (difference in value w.r.t previous waypoint), the positional information, the satellite Information and also the waypoints for GPS and IRNSS are displayed on the basemap with regional boundaries. Then all of retrieved information is logged in files which allow the application to revisit the journey step by step as in real-time survey, this allows us to retest the software after any alteration without the need for an additional field survey. The positional data is stored in database for various reasons like the ability to derive important information from the field survey, to plot the positional data on the map, create reports and perform data analysis (Fig. 5).

READER1: The reader1.py file takes user input for date (ddmmyy) and creates a folder (if it does not exist already) for log files for that date (it is up to the user to use new directory or use existing directory) and starts writing the data received from the receiver into the log files, if the data is unavailable the receiver will still be continuously flushing empty NMEA format data and writing it to the log file. The data is written in a new log file by incrementing the logfile count every 30s and if the receiver is disconnected it keeps waiting for the connection and continues from the previous point and if the application is restarted it can be made to continue from next log count.

MAIN1: The main1.py file again takes user input for the date (ddmmyy) and log count, checks if a folder exists and then starts reading the log files for that date from the given count (generally enter 1), this allows us to run the application for any survey data past or real time, if run in a real time session it reads all log files from the given log count and starts doing the following steps:

1) Parses NMEA data for GPGGA, IRGGA, GPRMC,GPGSA, GPGSV strings

2) Extracts Date, Time, Lat, Long, Alt for GPS and IRNSS

3) Inserts the extracted data into database

4) Displays the new information through the GUI

When running the application for a real-time session, reader1, py is started first and then main1, py, where as if user just wants to display previous journey data, directly main1.py is run by giving the date for the previous journey.

Database: This requires MySQldb middle to interface with MySQL database, we are all the coordinates from GPS and IRNSS separately and also a table to bookmark the locations. Each entry includes index, latitude, longitude and time.

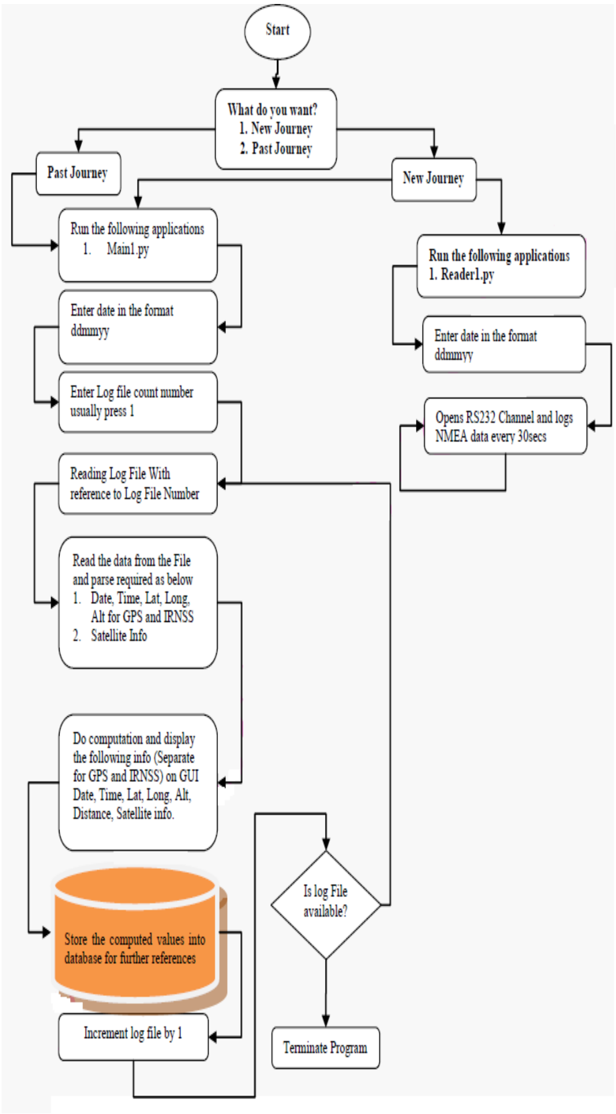

Start
What do you want?
1. New Journey
2. Past Journey
Past Journey
New Journey
Run the following applications
1. Main1.py
Run the following applications
1. Reader1.py
Enter date in the format ddmmyy
Enter date in the format ddmmyy
Enter Log file count number usually press 1
Opens RS232 Channel and logs NMEA data every 30secs
Reading Log File With reference to Log File Number
Read the data from the File and parse required as below
1. Date, Time, Lat, Long, Alt for GPS and IRNSS
2. Satellite Info
Do computation and display the following info (Separate for GPS and IRNSS) on GUI Date, Time, Lat, Long, Alt, Distance, Satellite info.
Is log File available?
Store the computed values into database for further references
Increment log file by 1
Terminate Program

Fig. 5. Flow of the Process.

SQL DATA BASE Host: Local ;UserName: faraz1 ; Pwd:ppp ; DB: temps db = MySQLdb.connect("localhost", "faraz1", "ppp", "temps") con=db.cursor() # Inserting parsed IRNSS values into Database (Similar to GPS values) sql = "INSERT INTO II1 (ILON,ILAT,IALT,ITIME,DATE) VALUES('"+str(ilon[it])+"','"+str(ilat[it])+"','"+str(ialt[it])+"',' "+str(itime[it])+"',"+d1+");" con.execute(sql) db.commit() # Inserting bookmarked GPS values into Database sql = "INSERT INTO gps_bookmark (NAME,LON,LAT,ALT) VALUES('"+str(name)+"','"+str(glon[0])+"','"+str(glat[0])+"',' "+str(galt[0])+"');" con.execute (sql) #Finalizing and closing connection db.commit() db.close()

Graphical User interface: This includes loading the main Application interface, Map components and tool bar to interact with the Map. The GUI updates itself every few times in a second and allows reflecting any triggered changes. Base-map allows us to plot the Map with some features like country boundaries, State boundaries; terrain etc. city layout features can also be added.

Inbuilt modules: Time, sys and OS modules are inbuilt module they are used as follows. OS module is used for performing directory and file related operations like making new directories and checking if directory or file exists etc.

Text extraction and pattern matching: re-module is used for text extraction using regular expressions, helping us to extract strings matching certain pattern.

Haversine: This module is used to calculate distance between 2 points on earth taking into consideration the curvature of the earth since directly using the coordinates will not give cutest results.

IV. Hardware Development

Fig. 6 shows the block diagram representation which consists of the portable electronic device includes raspberry pi model B, 7 inch touch screen and USB-TTL-RS232 converter. The IRNSS-UR is interfaced with the designed portable electronic device through USB-TTL-RS232 converter. The IRNSS-UR receives the L1, L5 and S1 band signals from IRNSS and GPS satellites and provides NMEA data format at RS232 channel. The portable electronic device reads the received NMEA data and stores the data in the text file format at a frequency of every 30 seconds. The algorithms were developed to read the real time NMEA information from the text files and are parsed to get date, time, latitude, longitude, altitude, signal strength and satellite information. These parsed data are further processed to display the information on the display screen and also to trace the corresponding waypoints of GPS and IRNSS on base-map using two different colors. The algorithms are also responsible to save the parsed data into SQL database for further offline analysis. This portable electronic device integrated with developed algorithms helps us to trace the real time waypoints received from GPS receiver and IRNSS receiver. This helps us to map both GPS and IRNSS rover positional information and to indentify the coverage area or rover movement simultaneously in real time.

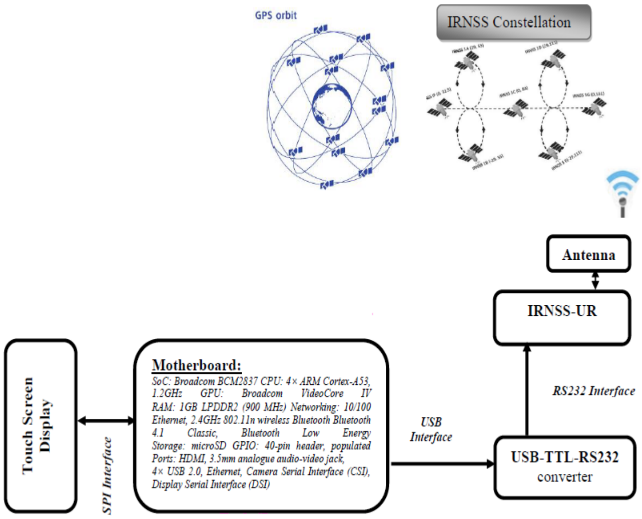

GPS orbit
IRNSS Constellation
Antenna
IRNSS-UR
RS232 Interface
Touch Screen Display
SPI Interface
Motherboard:
SoC: Broadcom BCM2837 CPU: 4× ARM Cortex-A53, 1.2GHz GPU: Broadcom VideoCore IV RAM: 1GB LPDDR2 (900 MHz) Networking: 10/100 Ethernet, 2.4GHz 802.11n wireless Bluetooth Bluetooth 4.1 Classic, Bluetooth Low Energy Storage: microSD GPIO: 40-pin header, populated Ports: HDMI, 3.5mm analogue audio-video jack, 4× USB 2.0, Ethernet, Camera Serial Interface (CSI), Display Serial Interface (DSI)
USB Interface
USB-TTL-RS232 converter

Fig. 6. Block Diagram Representation.

A. Raspberry Pi

The Raspberry Pi is a series of small single board computers developed in the UK by the Raspberry Pi Foundation. All models feature a Broadcom System on Chip (SoC) with an integrated ARM compatible CPU and On-chip Graphical Processing Unit (GPU). Processor speed ranges from 700 MHz to 1.2 GHz for the Pi 3 and on-board memory range from 256 MB to 1 GB RAM. Secure Digital cards are used to store the operating system and program memory in either SDHC or Micro SDHC sizes. Depending on the model; the boards have either a single USB port or up to four USB ports. For video output, HDMI and composite video are supported, with a standard 3.5mm phono jack for audio output. Lower level output is provided by a number of GPIO pins which support common protocols. The B-models have an 8P86 Ethernet port and the Pi 3 and Pi Zero W have on-board Wi-Fi 802.11n and Bluetooth. The default firmware is closed source, while an unofficial open source is available. The major components used in the raspberry Pi shown in Fig. 7.

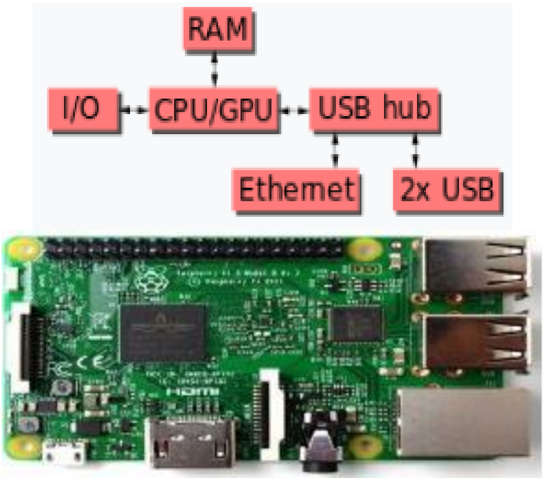


Fig. 7. Components of Raspberry Pi and Model.

The Ethernet adapter is internally connected to an additional USB port. In Model A, A+, and the Pi Zero, the USB port is connected directly to the System on Chip (SoC). On the Pi 1 Model B+ and later models the USB/Ethernet chip contains a five-point USB hub, of which four ports are available, while the Pi 1 Model B only provides two. On the Pi Zero, the USB port is also connected directly to the SoC, but it uses a micro USB (OTG) port.

A typical block diagram of the IRNSS-UR is shown in Fig. 8. IRNSS-UR is expected to receive, down convert and demodulate the transmitted satellite signals both at L5 (1176.45 MHz) and S1 (2492.028 MHz) band frequencies. Most importantly, IRNSS-UR generates measurements precisely with respect to the external/internal trigger such as 1 Pulse per Second (PPS). In addition, IRNSS-UR shall also include capability to process Global Positioning System (GPS) L1 Coarse/Acquisition (C/A) signals centered at 1575.42 MHz and generate measurements with respect to the external/internal trigger. The receiver shall output the user position computed using L5 only, S1 only; L1 only, combined L5 and GPS and combined S1 and GPS and combined IRNSS and GPS. Control inputs to the system include 10 MHz external reference clock, 1-PPS signal and commands/data through Ethernet and/or RS232 interface. The IRNSS-UR shall also have an USB port for IF sampled data collection into the PC and an external storage device.

B. Antenna Setup

Antenna module contains a mounting interface and an RF connector on bottom plate. Mounting interface is a standard M16 nut and can be mounted on standard M16 pipe/threaded pole (Fig. 9) [4]. Ensure there are no metal parts nearby the antenna for optimistic performance during installation. RF connector is a TNC Female type, through which the RF signal received by the antenna is fed to the receiver and the DC supply from the receiver is fed to the antenna LNA.

1) The antenna mounting site should provide full 360-degree visibility of the horizon. Any physical obstruction having an apex that makes an angle more than 5 degrees with the antenna phase centre, degrades the unit performance by blocking the satellite signals.

2) Ensure that there is no metal objects/plates touching or very close to the antenna. This is because the metal object alters the gain pattern of the antenna.

C. Receiver Setup

The back panel of the IRNSS-UR is as shown in the Fig. 10(a). The details and the functionalities of the ports in the back panel of the IRNSS-UR are as mentioned in Table I and the front panel of the IRNSS-UR is as shown in Fig. 10(b). The details and the functionalities of the ports in the front panel of the IRNSS-UR are as mentioned in Table II.

The steps to be followed depending on the Users' data requirement in a particular signal/data output are (Fig. 11):

a) To view the NMEA data in the GUI: Connect RS 232 cable from Laptop to the NMEA/LCD port of receiver through the USB-to RS-232 converter (Table III).

b) To collect IF samples into PC: Connect USB cables from Laptop to the IF SAMPLES TO PC port and SD Card Data TO PC port of receiver.

c) To collect IF samples into Hard-disk: Connect USB cable from External Hard disk to the IF SAMPLES TO HARD DISK port of receiver.

d) To run the Receiver with External clock source: For External clock source, connect SMA Cable from external clock source to the 10 MHz IN port of receiver.

e) To Latch the Receiver measurements with external PPS: Connect SMA Cable from external PPS source to the EXT PPS IN port of receiver.

f) To log the Navigation data of IRNSS into SD card: Insert SD card into the SD CARD slot of receiver.

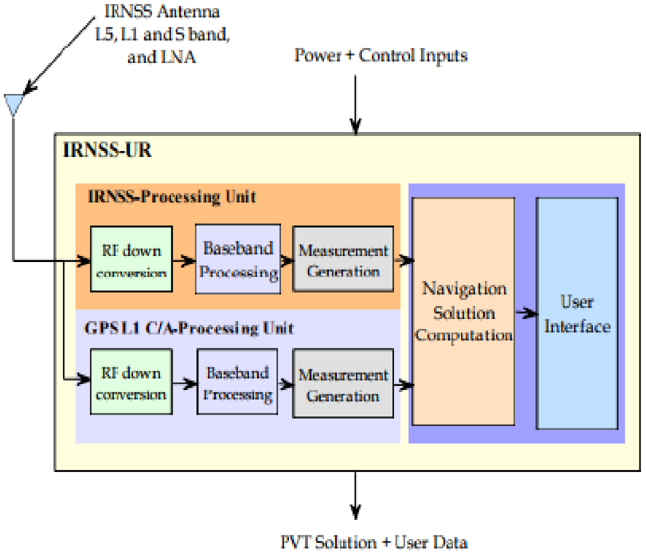


Fig. 8. High Level Block Diagram of IRNSS-UR.

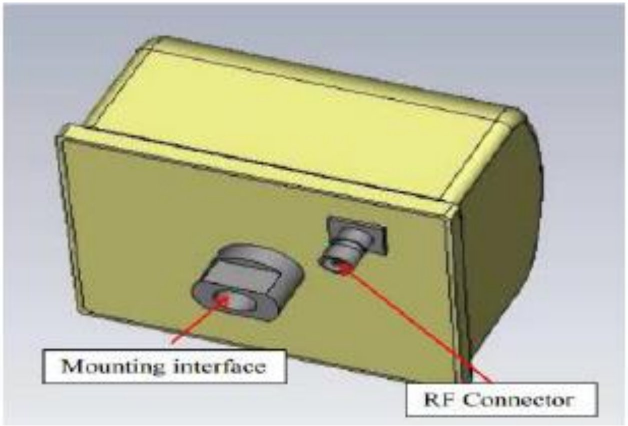
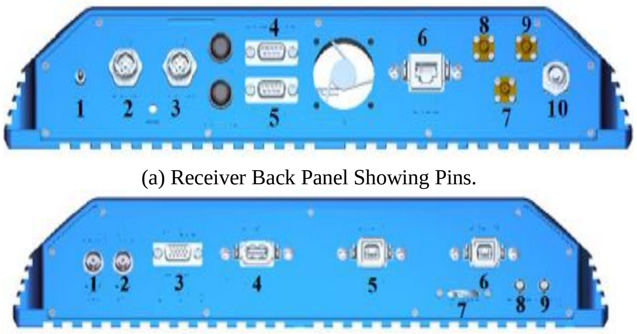


Fig. 9. IRNSS Antenna.



(a) Receiver Back Panel Showing Pins.

Fig. 10. (b) Receiver Front Panel Pins.

TABLE. I. RECEIVER BACK PANEL PINS AND THEIR FUNCTIONALITIES

Sl. No.	Ports	Functionality
1	ON/OFF	Gives control to power ON/OFF the IRNSS-UR
2	12 V DC IN	12 V DC supply is fed from battery
3	230 V AC IN	230V AC supply is fed from mains
4	NMEA/LCD	Used for NMEA data output, Ethernet IP configuration, and also for data communication with the detachable LCD (in the field environment)
5	SERVICE PORT	Factory Maintenance Port
6	ETHERNET PORT	To display status on Graphical User Interface running on host computer and also to receive commands
7	PPS IN	External PPS is fed through this port for measurement latching
8	10 MHz IN	External 10 MHz clock reference is fed through this port
9	10 MHz OUT	Receiver gives out 20 MHz clock reference through this port.
10	ANT	IRNSS L5, S and CPS L1 RF signals are fed through this port

TABLE. II. RECEIVER FRONT PANEL PINS AND THEIR FUNCTIONALITIES

Sl. No.	Ports	Functionality
1	IRNSS PPS OUT	IRNSS 1 PPS reference output
2	GPS PPS OUT	GPS 1 PPS reference output
3	RS 232 Monitoring	PRN code, PRN code epoch, bit synch, data and data clock are given outside through this port
4	IF samples to Hard disk	Intermediate frequency samples are transmitted to the external hard disk through this USB port
5	SD Card Data to PC	Stored 25 bps navigation data is transmitted to the host computer through this USB port
6	IF samples to PC	Intermediate frequency samples are transmitted to the host computer through this USB port
7	SD Card	25 bps navigation data is stored in the SD card
8	BATT	Indicates 12 V DC is connected to the IRNSS – UR
9	MAINS	Indicates 230 V AC is connected to the IRNSS-UR

TABLE. III. LIST OF CABLES

Cable Name	Quantity
230 V AC cable	1
12 V DC Battery cable	1
RS 232 cable (1m) for NMEA port	1
LCD cable (1 feet)	1
Ethernet cable	1
LMR 400 DB antenna cable	1
USB type B to type A cable	2
Battery charger adapter and cable	1-set
USB to RS-232 converter	1

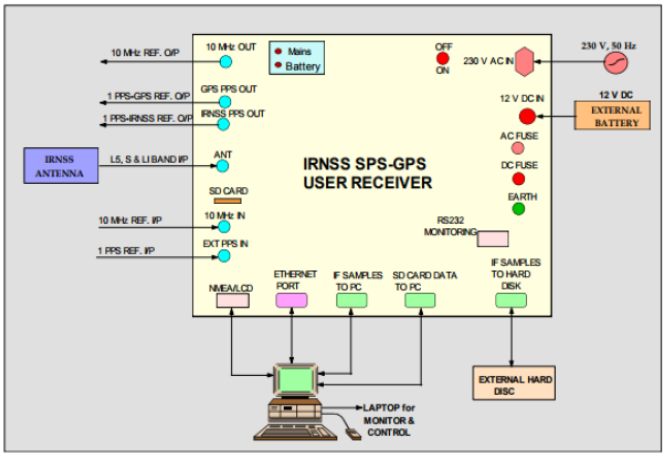


Fig. 11. Interconnection Diagram of IRNSS –UR.

D. Operating Systems

The Raspberry Pi Foundation recommends the use of Raspbian, a Debian-based Linux operating system which includes Ubuntu MATE, Snappy Ubuntu Core, Windows 10 IoT Core, and RISC OS.

To set up a blank SD card with NOOBS:

1) Format an SD card which is 8GB or larger as FAT.

2) Download and extract the files from the NOOBS zip file.

3) Copy the extracted files onto the SD card that you just formatted, so that this file is at the root directory of the SD card. Please note that in some cases it may extract the files into a folder; if this is the case, and then copy across the files from inside the folder rather than the folder itself.

4) On first boot, the "RECOVERY" FAT partition will be automatically resized to a minimum, and a list of OSes that are available to install will be displayed.

E. Hardware Packaging

3D Packaging: The below pictures represents the 3D design of Back, Front and Battery panels. The complete device is enclosed in these three panels. The complete product packaging is designed such that the device is portable and battery operated. Fig. 12 shows the display model along with 3D printing enclosure.

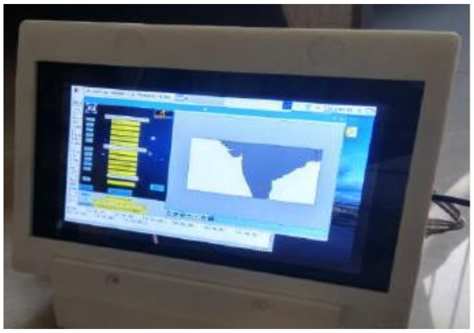


Fig. 12. Packed Display with Results.

V. Observation and Analysis

The application software running on the hardware during the field survey test is

Step 1: The user need to run the Main1.py application and application waits for the user to enter the date in the ddmmyyyy format (Fig. 13).

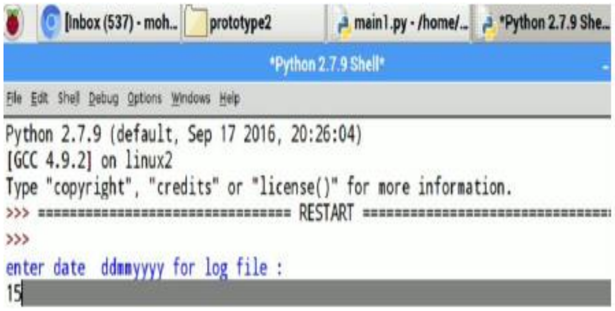
Step 2: After entering the date in the ddmmyyyy format (Fig. 14). The Graphical User Interface that gives the GPS and IRNSS receiver Positional information with date and time separately. The GUI is also integrated with basemap on which the waypoints are traced with red and green marker and finally click on start journey button (Fig. 14).

Step 3: After clicking on Start Journey button. The user needs to enter the count for the log file in the terminal window. After every 30 seconds the log files are created containing the NMEA data of both GPS and IRNSS receiver (Fig. 15). The waypoints are also displayed on the base maps.

The positional information of a rover with respective to GPS and IRNSS is displayed as shown in the Fig. 16. The parameter Date, Time, Latitude, Longitude and Altitude are displayed.

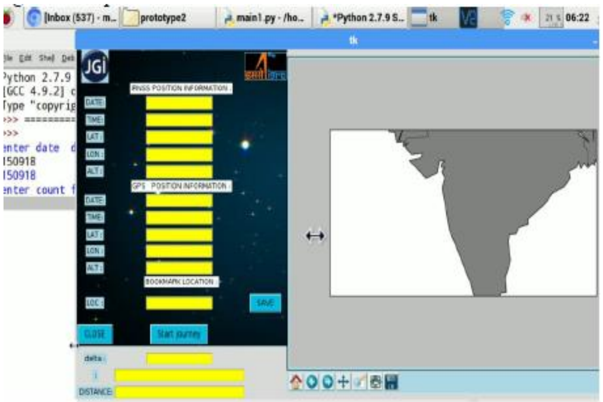
The Base map displaying waypoints in Red (IRNSS) and Green (GPS) color are traced to show the rover movement (Fig. 17). The distance travelled by the rover is also calculated using haversine expression and is displayed at the left bottom corner of the GUI. The base map can be zoom in and out, and can be moved.

The Rover distance travelled with respective to the IRNSS and GPS are determined separately and displayed on the GUI (Fig. 18). The satellite information like the number of satellite visible and the delta distance are displayed.



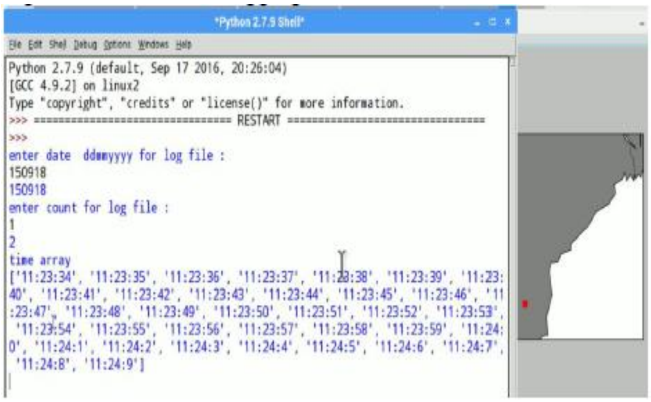
```
Python 2.7.9 (default, Sep 17 2016, 20:26:04)
[GCC 4.9.2] on linux2
Type "copyright", "credits" or "license()" for more information.
>>> ================================ RESTART ================================
>>>
enter date  ddmmyyyy for log file :
15
```

Fig. 13. Main1.py Application.



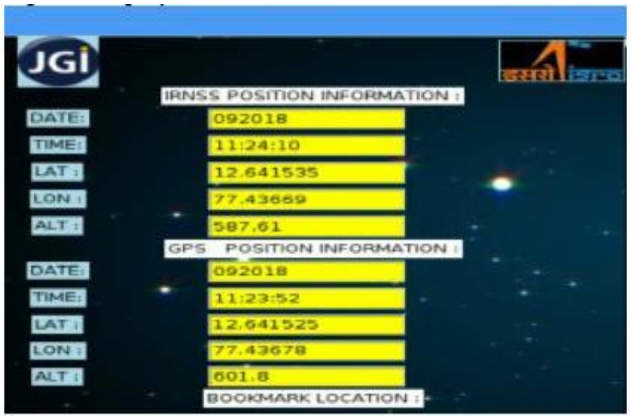
```
Python 2.7.9
[GCC 4.9.2] c
Type "copyrig
>>> ========
>>>
enter date  d
150918
150918
enter count f
```

Fig. 14. Graphical user Interface.



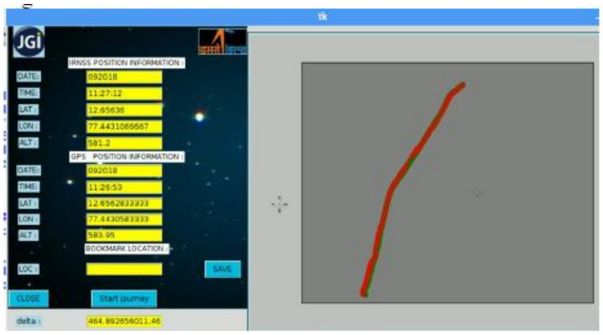
```
Python 2.7.9 (default, Sep 17 2016, 20:26:04)
[GCC 4.9.2] on linux2
Type "copyright", "credits" or "license()" for more information.
>>> ================================ RESTART ================================
>>>
enter date  ddmmyyyy for log file :
150918
150918
enter count for log file :
1
2
time array
['11:23:34', '11:23:35', '11:23:36', '11:23:37', '11:23:38', '11:23:39', '11:23:
40', '11:23:41', '11:23:42', '11:23:43', '11:23:44', '11:23:45', '11:23:46', '11
:23:47', '11:23:48', '11:23:49', '11:23:50', '11:23:51', '11:23:52', '11:23:53',
 '11:23:54', '11:23:55', '11:23:56', '11:23:57', '11:23:58', '11:23:59', '11:24:
0', '11:24:1', '11:24:2', '11:24:3', '11:24:4', '11:24:5', '11:24:6', '11:24:7',
 '11:24:8', '11:24:9']
```

Fig. 15. NMEA Data Logging.



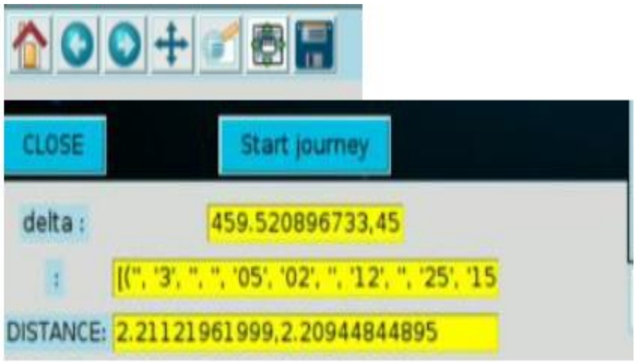
IRNSS POSITION INFORMATION :
DATE: 092018
TIME: 11:24:10
LAT : 12.641535
LON : 77.43669
ALT : 587.61
GPS POSITION INFORMATION :
DATE: 092018
TIME: 11:23:52
LAT : 12.641525
LON : 77.43678
ALT : 601.8
BOOKMARK LOCATION :

Fig. 16. Display Rover Positional Information.



IRNSS POSITION INFORMATION :
DATE: 092018
TIME: 11:27:12
LAT : 12.65636
LON : 77.4431066667
ALT : 581.2
GPS POSITION INFORMATION :
DATE: 092018
TIME: 11:26:53
LAT : 12.6562833333
LON : 77.4430583333
ALT : 583.95
BOOKMARK LOCATION :
LOC :
delta : 464.892656011.46

Fig. 17. Rover Movement.



delta : 459.520896733,45

: [('', '3', '', '', '05', '02', '', '12', '', '25', '15

DISTANCE: 2.21121961999,2.20944844895

Fig. 18. Rover Distance Travelled.

The waypoints i.e., latitude and longitude of both GPS and IRNSS are loaded into Google maps to trace the rover movement using device logged data and accord logged data shown in Fig. 18.

Fig. 19 shows the GPS waypoints from the NavIC user receiver depicted by blue color markers, the journey shows a more consistent waypoints dataset than the previous datasets although there is small discontinuity only due to the obstructions like flyovers and other tall buildings or due to signal loss, the GPS data received by the portable device is compared with the above image to highlight irregularities and bottlenecks associated with the portable device or the code. The journey was made from Jain University located near Jakkasandra and the stopping point was near JP Nagar. The data stayed consistent despite the speed changes and the waypoints correctly coincided with the roads marked on the map, we can infer from this that there was no error due to shift of values as observed in the earlier surveys from the portable device readings. It can also be observed in the image that there were no outliers in data that were observed in IRNSS readings from the NavIC user receiver.

Fig. 20 shows the IRNSS waypoints from the NavIC user receiver depicted by red color markers, the journey shows a more consistent waypoints dataset than the previous datasets which is also observed in the GPS data above, although there is small discontinuity due to the obstructions like flyovers and other tall buildings or due to signal loss and also we can observe few outliers that are clearly result of a minor fault since the vehicle wouldn't have travelled to the points depicted by the outliers, the IRNSS data received by the portable device is compared with the above image to highlight irregularities and bottlenecks associated with the portable device or the code. This image shows the journey data made from Jain University located near Jakkasandra to the stopping point that was near JP Nagar. As with the GPS data, the IRNSS data also stayed consistent despite the speed changes and the waypoints correctly coincided with the roads marked on the map and also with GPS data from the NavIC receiver, we can infer from this that there was no error due to shift of values as observed in the earlier surveys from the portable device readings.

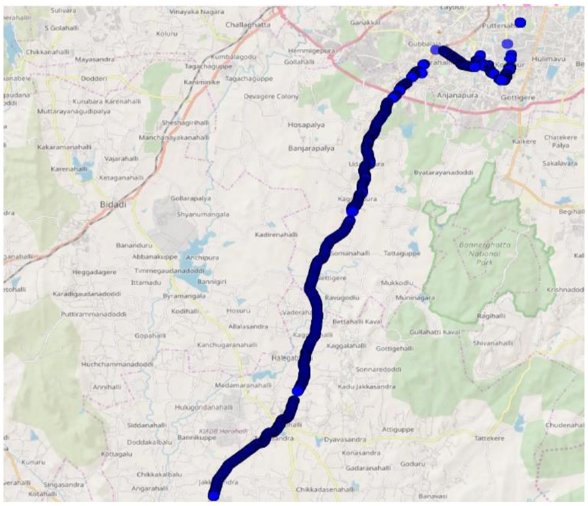


Fig. 19. Google Map using GPS Waypoints from Accord Logged Data.

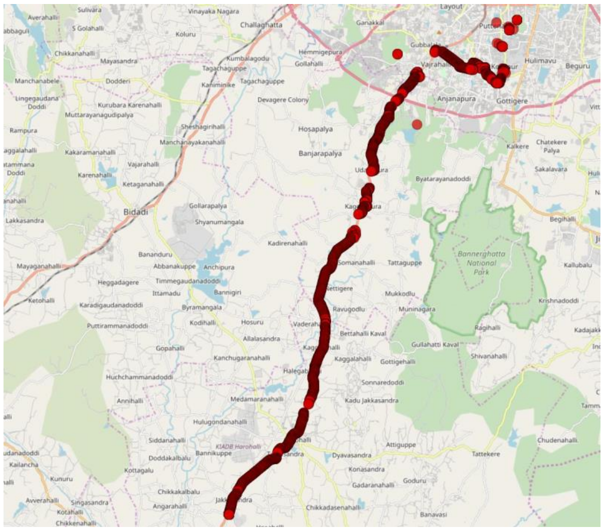


Fig. 20. Google Map using IRNSS Waypoints from Accord Logged Data.

VI. Conclusion and Future work

As with the IRNSS data from the NavIC user receiver, the portable device IRNSS data also stayed consistent despite the speed changes and the waypoints correctly coincided with the roads marked on the map and also with GPS data from the NavIC receiver, we can infer from this that there was no error due to shift of values as observed in the earlier surveys from the same device and the device data quality was comparable to that of the Receiver data. One important observation is that the data is a slightly sparser than the original data, this is due to the fact that code reduces the frequency of taking the way points to avoid redundancy of data points. This device can also be used in mapping and surveying project. By utilizing these devices capabilities in surveying can potentially save companies cost and time. This will allow surveying positions in the shortest time possible. Mapping can be done in case of highways, farmlands, rivers, power lines, etc.

This device can be further developed to guide and monitor the all transportation vehicles including aircrafts and keep them informed about the routes of the neighboring aircrafts in the airspace.

ACKNOWLEDGMENT

This research work was carried out by the grant sanctioned by ISRO File No. NGP22. The authors like to acknowledge the support of Space Applications Centre - Indian Space Research Organization (SAC-ISRO) by providing IRNSS receiver and encouraging us to work on this navigation studies. Authors thank Dr. G. Raju & JAIN (Deemed-to-be University), Bangalore for the support to complete this work successfully. Scholar also acknowledges the Directorate of Minorities, Government of Karnataka for their Ph.D. Fellowship.

REFERENCES

[1] Chandrasekar, M. V. et al., Modernized IRNSS broadcast ephemeris parameters. J. Control Theory Inf., vol. 5, Iss.2, 2015.

[2] ISRO, Indian Regional Navigation Satellite System Signal in Space ICD for Standard Positioning Service (Version 1.0, ISRO-IRNSSICD- SPS-1.0), Indian Space Research Organization, 2014.

[3] Mohmad Umair Bagali, Dr. Thangadurai N, “Embedded Board Development Interfaced with GPS/IRNSS/NavIC Receiver for Disaster Applications”, Proc. Of SSRN, International Conference on Sustainable Computing in Science, Technology & Management, pp. 416 – 426, Jaipur, 2019.

[4] Mohmad Umair Bagali, Naveen Kumar Reddy, Ryan Dias, Dr. Thangadurai N. The Positioning and Navigation System on Latitude and Longitude Map using IRNSS User Receiver. IEEE International Conference on Advanced Communication Control and Computing Technologies, Ramanathapuram, pp.122-127, 2016.

[5] Rao, V. G., Lachapelle, G. and Vijay Kumar, S. B., Analysis of IRNSS over Indian Subcontinent. J. Inst. Navigation, San Diego, 2011.

[6] Kaplan, “Understanding GPS: principles and applications”, 2nd Ed, Artech House, 2006.

[7] Ganeshan, A.S., Rathnakara S.C., Gupta R., and Jain, A.K., Indian Regional Navigation Satellite System (IRNSS) Concept, ISRO Satellite Center Journal of Spacecraft Technology, 2005, 15(2), pp. 19–23.

[8] Mohmad Umair Bagali, Thangadurai N, “Application Specific Embedded Board Development Interfaced with GPS/IRNSS Receiver for Environmental Monitoring”, International Journal of Innovative Technology and Exploring Engineering, Vol.8, Iss.8, pp. 2628–2637,2019.

[9] Grewal, M.S., Weill, L.R., and Andrews, A.P., Global Positioning Systems, Inertial Navigation and Integration, Wiley Publications, 2007, 2nd edition.

[10] Van Diggelen, F., GNSS Accuracy: Lies, Damn Lies, and Statistics, GPS world, 2007, pp. 26–32.

Empirical Performance Analysis of Decision Tree and Support Vector Machine based Classifiers on Biological Databases

Muhammad Amjad[1*], Abid Rafiq[3]
Department of Computer Science and Information Technology
University of SargodhaSargodha, Pakistan

Zulfiqar Ali[2]
Department of Computer Science and Information Technology
University of Central Pujab (UCP), Lahore, Pakistan

Nadeem Akhtar[4], Ali Abbas[6]
Department of Computer Science and Information Technology
The University of Lahore (UOL), Lahore, Pakistan

Israr-Ur-Rehman[5]
Department of Computer Science
Islamia College University, Peshawar, Pakistan

***Abstract*—The classification and prediction of medical diseases is a cutting edge research problem in the medical field. The experts of machine learning are continuously proposing new classification methods for the prediction of diseases. The discovery of classification rules from medical databases for classification and prediction of diseases is a challenging and non-trivial task. It is very significant to investigate the more promising and efficient classification approaches for the discovery of classification rules from the medical databases. This paper focuses on the problem of selection of more efficient, promising and suitable classifier for the prediction of specific diseases by performing empirical studies on bunch mark medical databases. The research work under the focus concentrates on the benchmark medical data sets i.e. arrhythmia, breast-cancer, diabetes, hepatitis, mammography, lymph, liver-disorders, sick, cardiotocography, heart-statlog, breast-w, and lung-cancer. The medical data sets are obtained from the open-source UCI machine learning repository. The research work will be investigating the performance of Decision Tree (i.e. AdaBoost.NC, C45-C, CART, and ID3-C) and Support Vector Machines. For experimentation, Knowledge Extraction based on Evolutionary Learning (KEEL), a data mining tool will be used. This research work provides the empirical performance analysis of decision tree-based classifiers and SVM on a specific dataset. Moreover, this article provides a comparative performance analysis of classification approaches in terms of statistics.**

Keywords—Classification; rules discovery; support vector machine; decision tree

I. Introduction

The Knowledge Discovery is processing of finding the non-trivial, useful and hidden patterns from a very large database. Knowledge discovery and data mining are a new trend in information technology. Traditionally a large part of the process was done by manually that is time-consuming task. With time new technologies invented and task shifted from manually to computerized form. Business knowledge is necessary in advance to compete in the world. Data storage is now a day reached to amount of terabyte size [1]. But it is necessary to extract useful knowledge from it for use. So knowledge discovery is the name of the discovery of hidden knowledge from large databases. Knowledge discovery contains the steps of data preparation, data preprocessing, and hypothesis generation, the formation of the pattern, evaluation, knowledge representation, knowledge refinement, and knowledge management [2]. It also includes many stages for databases updating.

Machine Learning methods and biological databases play a significant role in disease diagnosis. It helps in future for diagnosing of medicine. The biological database includes information about gene structure, function, and similarities of structure and sequences of biological data. Classification of the biological database can be done in two forms as a specialized and comprehensive database. The comprehensive database includes different species database, for example, GenBank [3] and specialized databases consist of a special organism or species databases, for example, WormBase [4].

Machine learning becomes a necessary part of solving the problem in every branch of science. In biomedicine to predict genetic sequence and protein structure machine learning has been used [5]. Machine learning is used to extract hidden knowledge for the different data set. It includes neural network, boosting, support vector machine and decision trees [6]. In machine learning, two ways are performed for data mining. It is supervised learning we make a dataset to extract new data from a large amount of data. New data and training data set match for validation of result. But in unsupervised learning, some pattern is used to classifying the data without explicit instruction [7]. Reinforcement learning focus on the reward and output achieve in the form of reward and punishment. An agent is required to gain the maximum reward to gain the result. Agent focuses on the positive situation to gain maximum reward. Negative situation decreases the reward. This type of learning is used in control theory, statistics, information theory, etc.

This research article investigates the performance of Decision Tree approach and Support Vector Machine Algorithm for the discovery of classification rules. The

*Corresponding Authors.

interesting and useful discovered rules are used for the building of classifiers. The classifiers are applied for the diagnoses of the various harmful diseases. In this paper, we use KEEL [8] data mining tool for the data processing and classification of the biological databases.

Section II provides the related work published in contemporary literature. Section III gives information about the decision tree-based classification and provides the empirical performance analysis of selected classifiers on medical databases. Section IV provides a basic understanding of SVMs and comparative empirical study on medical data sets. Sections V and VI provide the experimental setup and discussion on the results produced during the under focused research study and the last section concludes the findings of the research work.

II. Related Work

This section provides the literature review of the various research carried by the different researchers in this field. The following section gives information about the use of different classification for the discovery of rules and the classification of different biological diseases.

There are many techniques are used to find a pattern inpatient health data. The best system is one that is the efficient, adoptive, generic and affordable system. Many factors affect the result of analysis like an error in online databases, sensor's settlement. This study shows that ASP logic approach is the best use for incomplete biological data. Artificial Neural Network is best used for single purpose system. ANN generates best better result than ASP and another approach used in the health care system. If the hardware is costly then it difficult to use this system [9].

There are many data mining algorithm available but this study provides a comparative study of three algorithms Naïve Bayes, Decision Tree and Multi-Layer Perceptron Neural Network. In this study, window operating system 8.1 is used with WEKA data mining tool. Ebola Disease data set contain the range of 250-10000 instances that are stored in MySQL. According to this study, the Naïve Bayes algorithm shows a negative correlation, with the increase in the dataset it performance lead to a decrease. WEKA shows a positive correlation. Naïve Bayes is the best and popular machine learning algorithm is fast in training [10].

Mohammed H. Tafish and Dr. Alaa M. El-Halees proposed a model as Breast Cancer Severity Degree Predication Using Data Mining Techniques in the Gaza Strip described that in Gaza Area cancer disease and diabetes growth are top disease during the last decades. They used a data mining method to diagnose cancer and diabetes disease. They proposed a model using data mining techniques like SVM, KNN, and ANN. Breast cancer data taken from Gaza hospital used, after evaluation and test by applying the above techniques they obtain 77% accuracy for the prediction of the severity of breast cancer [11].

Manickam Ramasamy at el. proposed a model for predicting hepatitis in which they provide an empirical analysis of the decision tree algorithm by using Hepatitis data set taken from the UCI machine learning repository. They used different classification algorithm and accuracies of classification are performed by 10 cross validation techniques. By using different classifier they concluded that Random forest takes less running time with the highest accuracy of 87.50%. This accuracy gives help in ailment prediction and classification in the field of medical science [12].

In this study extended deep learning method is used for classifying multimedia data set. Convolution Neural Network is a deep learning method is costly but this paper feed low level features in this approach. To find the best result CNN is used with the bootstrapping method. TRECVID data set is used in this approach which is high-level imbalanced data set. This approach works effectively on the use of low-level features that reduced the training time in deep learning [13].

Anuj et al. describe Parkinson's disease. It is the connection between speech impairment and Parkinson's disease. In this paper classification based deep learning (Deep Neural Network, Dimensionality reduction techniques) and machine learning algorithms (Logistic regression, Naïve Bayes, K-Nearest Neighbor, Decision tree, Random forest) are employed with the use of Dimensionality reduction. The data set Parkinson's Speech is used in this approach that is obtained from the UCI machine learning repository. The result is extracted with the base of accuracy. KNN produced 95% highest accuracy with 10 features [14].

Sara Belarouci et al. propose meta-heuristics optimization methods for improvement of medical classifier performance. They are used many algorithms like Genetic Algorithm PSO, Simulated Anneeling to compare with Least Square Support Vector Machine to improve the classification with aspect to False Positive and Negative. Meta-heuristics Optimization is best for solving the problem of unbalance dataset. Five different datasets related to various diseases like Liver Disorder, Appendicitis, and Diabetes. This approach will help doctors to diagnose many diseases effectively [15].

Tharaha S and Rashika K proposed this research using Hybrid Artificial Neural Network and Decision Tree algorithm for disease recognition. They used Artificial Neural Network for training data and decision tree for classification of data because the Decision Tree algorithm is a good classifier. Datasets are taken from the human blood detecting and sensor counting, stored with different attributes. Time taken for test split in ANN is 0.09s and where decision tree took time is 0.14s. The result is shown by apply WEKA 3.8.1 version. The combination of these two algorithms gives the best result than separate used and provide the best help for disease diagnosing [16].

Dania Abed aljawad et al. proposed an empirical study of Bayesian Network and Support Vector Machines for Breast Cancer surgery Survivability Prediction. They used Haberman's survival dataset and evaluate the performance of the Bayesian network and Support Vector Machine using WEKA tool. Empirical research shows that Support Vector Machine best performs with an accuracy of 74.44% than Bayesian network with an accuracy of 67.56%, Imbalance data is converted into balance. This study helps the doctors to the prediction of the patient stage of cancer using old data as a sample to new data [17].

P. Hamsagayathri and P. Sampath proposed a Priority Based decision Tree Classifier for Breast cancer. Women mostly from 40-70 age affected with breast cancer. So they proposed a model for prediction of breast cancer. Classification provides a vital role in the detection of breast cancer and helps the researcher to analyze and classify data. SEER breast cancer data set is used in this paper. Two decision tree algorithm J48 and priority-based decision tree algorithm are used. The priority-based algorithm provides the best result with less time consuming to build the model. J48 used repetitive but priority base algorithm not used repletion step and 98.51 accuracies [18].

With the reference of above literature review, the specific medical data sets i.e. arrhythmia, breast-cancer, diabetes, hepatitis, mammography, lymph, liver-disorders, sick, cardiotocography, heart-statlog, breast-w and lung-cancer are not used to investigate the performance of Decision Tree (i.e. AdaBoost.NC, C45-C, CART, ID3-C) and Support Vector Machines. In this research study will Decision Tree based classifiers and SVM Machines for the discovery of classification rules. The problem statement and objectives of this research are given in the next sections.

III. Decision Tree based Classification

After Decision Tree is most popular supervised machine learning algorithm applied for the various classification problems. It is used for classification and regression problems. Decision tree provides the result which is easily understandable by humankind. A decision Tree provide output in a tree-like graph in which each node represents to attribute, each branch provide a rule and each leaf node provide a target class. Target class may be in discrete or in continuous form. Decision Rule may be in IF-then-Else rule. Big decision tree means the more complex rule.

Decision Tree is used as a top-down approach for making a decision tree. It begins from the root node to the leaf node. The decision is made on each internal node where attributes are split into further node if it contains information that can be divided further. More information leads to further classification. If a node cannot have information more then it considered as leaf node that refers to the target value.

Different methods are used to construct a decision tree. Every method used different information for the construction of a decision tree. Large decision tree not considered an accurate and efficient decision tree. Different research shows that the best decision tree is as small as possible. It based on the proper selection of attributes. Attributes selection measures are used to split attributes into further sub attribute. It is a recursive approach. Attributes selection measure checks the impurity of the attribute. Impurity measurement method includes Gain Ratio, distance measures, Gini-index and information gain. ID3, C4.5 focused information gain and CART use Gini-index for attributes selection.

A decision tree process can be divided into two steps: one constructs a decision tree and other to pruning a decision tree. Data mining works on real world data. Data may have some missing value, wrong value, containing noise or even less essential data, so this problem may lead to over-fitting and will destruct the predictive performance. There are two basic strategies for pruning the decision tree i.e. first forward pruning means pruning before completion of decision tree and other post-pruning means pruning after making a decision tree. So forward pruning stop the pruning process before reaching its maturity level and in a post-pruning button-up, approach is used to cut off the node. The Minimum Description Length Principle, Expected Error Rate Minimization Principle and Principle of Occam's Razor are used for pruning.

A. *ID3*

ID3 stands for Iterative Dichotomize 3. It is built by J.R Quinlan in [19]. It is the core algorithm to build a decision tree. It generates all possible decision tree. It simply classifies the training and testing set for the dataset. It does not require much more computation as compared to another approach for creating a decision tree. It is an iterative approach. It chooses the training set randomly and makes the decision tree. If it answers all object then it terminates the process it not then it add to again in training data for further process. It iterates the process and makes the decision tree correctly up to thirty thousand instance and fifty attributes. This algorithm based on the information gain of candidates attributes. If any attribute has more gain information then it selected for decision tree and less gain information is discorded.

The effectiveness of this approach also depends on the computational requirement based on the gain of untested attributes and non-leaf nodes of the decision tree. The total computational power of the ID3 is relative to the size of the training set, several attribute, and non-leaf nodes. The similarity in attributes extends the computational requirement. In ID 3 time and space are not grow exponentially so it can be used for larger and complex tasks.

ID3 algorithm has some advantages like i.e. easily understandable rule for classification, it is fastest and provides a short tree. It calculation time is a linear function not exponential as well as it has some disadvantages i.e. data may be overfitted or over-classified due to the small sample and for the continuous value it computation time may be more due to make many trees to find where to break the continuum.

B. *C4.5*

Quinlan et al. proposed the extended version of ID3 that is known as C4.5 in [20]. It is also developed for making a tree. It is developed by Quinlan in 1993. Quinlan described many issues for decision tree-like handling missing value, pruning and converting trees to rule and how C4.5 handle it. Decision tree algorithms used some cases and make a tree-like structure in which the main node is called the root node and other node are test node and leave node. Every decision node used a test and leave node show the class label.

C4.5 algorithm creates a small, accurate and fast decision tree and it is known as a reliable classifier. These are the best and popular properties for making the classification. This algorithm extracts the best information from a set of cases and takes only one attributes for the test. For this purpose information gain and gain, the ratio is used for the selection of best attributes. Some dataset may contain unknown information so Quinlan used C4.5 approach. Information gain for unknown value can be ignored. And known value attribute information gain can be calculated. So information on this test case may be quite small. The unknown value may affect the decision tree making process.

An every decision tree cannot be considered as a good classifier for every data set in respect of making a smaller tree that may not fit for all training data. So avoid by overfitting, many decision tree algorithm used the pruning method. In this method, growing the decision is stopped while deleting the portions of the tree. C4.5 pruning method based on error rate. The error rate of every subtree is calculated if the error rate is low then it will be treated as a leaf node. This process used bottom-up approach. If C4.5 algorithm indicates that tree will be treated as accurate even children of concern node deleted than algorithm considered concern node as a leaf node. If this method proved as good then this decision tree is considered the best decision tree.

Quinlan discusses some shortcoming of c4.5. It has a built-in bias, t take only a single attribute for testing that takes more time computation. It makes the value of the given attribute in the same group and considered as a single value. It may use for single training set once and not used for other training set for binary classification. Suppose attributes for a chemical element that can be classified into the light and heavy element and other training set having an electric conductor that can be classified in conductor and non-conductor. So these groups may overlap with each other. This algorithm cannot is used for both groups. C4.5 used greedy approach for the grouping, so it gives the unsatisfactory result and remains an open problem.

C. Adaboost.NC

AdaBoost.NC is a negative correlation learning algorithm proposed by Wang et al. in [21]. It is used for classification ensemble. AdaBoost.NC algorithm is used for multiclass imbalance data. It provides the solution of two class imbalance problem. AdaBoost.NC provides the best accuracy with random oversampling on the minority class as compared to another balancing approach. The accuracy is achieved by the less border classification and overfitting in the minority class.

AdaBoost.NC is the advance version of AdaBoost for negative correlation but it based on AdaBoost training framework. It provides better classification boundaries and creates lower error correlation as compared to AdaBoost. This is used to improve the performance of the original AdaBoost algorithm. This algorithm is used for better classification in control of upper bound on the generalization error of Traditional AdaBoost. AdaBoost.NC provided the best performance in respect of the distribution of better margin.

AdaBoost is a very simple and effective ensemble algorithm. It is not only used to emphasize to misclassified example, but also provide the mechanism to control the error of misclassification of the same example. Due to this reason, it provides the best accuracy and diversity.

AdaBoost.NC does not show good performance in overall and in minority class working with class decomposition scheme. This algorithm receives and learns from all data information of all classes. It learns from several decomposition problems for partial knowledge. It provides the best performance in analyzing subproblem as compared to combine the whole problem. So it needs to better technique to combine the subproblem to acquire knowledge from AdaBoost.NC.

D. CART

CART stands for classification and regression tree. CART is proposed by Breiman et al. in [22]. It is an algorithm used to construct a decision tree from the categorical and continuous form of data. Classification is used for a categorical form of data and regression tree is constructed from a continuous form of data. The first time Morgan and Sonquist proposed a method to construct a tree by quantitative variable. They gave the name Automatic Interaction Detection. Each cluster is grouped into two clusters. Each predictor is tested on every cluster. Their model naturally incorporates interaction among all predictor.

A classification tree is dependent on discrete or categorical value. Kass (1980) proposed a modification in AID model called CHAID for the creation of a tree from the dependent and independent variable. This model limited to categorical predictor so it cannot be used for the quantitative variable.

These two models have a problem where to stop the tree. Breiman et al. (1984) method show that node that cannot contribute to prediction eliminate from the tree.

CART is a mechanism to construct a decision tree. It makes the solution in a tree-like structure. It starts from the root node and split into a test node on the base of selected attributes. This process ends on the leaf node that cannot be further divided. To make the best and effective tree it used pruning method i.e. Complexity based pruning. Pruning is started from the bottom toward the root node.

CART algorithm may a structure of question and answer of these question lead to the next question. So, the result of these question make a tree structure where to question is not more. CART uses the basic rule for making a decision tree i.e. splitting data rule and stopping rule where the terminal cannot be split and prediction of the leaf node. CART has some advantages like can handle missing value automatically.

IV. Support Vector Machine based Classification

Support Vector Machine was introduced after in the 1990s and used for many engineering application [23]. Support Vector Machine is an algorithm developed for binary classification by Cortes & Vapnik. The objective of this algorithm to find hyper-plane and classification of data points. It is used for separating the two classes with a maximum margin between two points called support vector. SVM algorithm is used for class separation, nonlinearity and overlapping classes where a data point lies in the opponent class [24].

Support Vector Machine classifies the data by using hyper-plane. The hyper-plane can be chosen by either of the sides but optimal hyper-plane is that maximizes the margin between two support vectors. Support vector is the data point that closer to the hyper-plane. Hyper-plane has different features on different location and deleting the support vector can influence the position of the hyper-plane [25].

The main purpose of the Support Vector Machine is to choose the best hyper-plane that classifies the data point correctively with maximum margin. It is easy to find the best hyper-plane in linear form but non-linear hyper-plane is hard to find as compared to linear form. For this purposes, a function called Kernel is used that find the best hyper-plane in no linear form. In non-linear form classification kernel trick, it mapped the input from low dimensional feature space to high dimension feature space.

Support Vector Machine algorithm provides a solution for a limited number of training data in more time and they consume more time for large databases [23]. It is used for text and hypertext categorization, classification of images, image segmentation, and hand-written recognition of character.

The following subsection describes the SVM based classification methods selected for the empirical study in this thesis. The naming convention for the methods is used of KEEL implementations.

A. C-SVM

The C-SVM is a new type of support vector machine proposed by Cortes and Vapnik in [26]. It used non-linear mapping to map the input vector to high dimension space and using this space, it constructs the decision surface to ensure the generalized ability of the network. The main purpose of support vector machine to separate the training data without an error when it is impossible in this scenario. It must find the optimal hyperplane to separate the training data. Optimal hyperplane maximizes the margin between two classes. C parameter makes the best classification between two classes with the optimal hyperplane. More support vectors are required to separate the training data that optimize the margin between classes.

In another case, soft margin hyperplane is used when training data is not possible to separate without minimum error. So, training data can be separated with a decreased error. In soft margin analysis, to minimize the expense of error rate the C parameter is used with less value to separate the training data with minimum error. Sometime dataset have positive and negative instance overlapped with each other so it is difficult to classify the data. On the other hand, it may be over-fitted that cause computational complexity. This problem may be solved with C-SVM algorithms.

In support vector machine Coefficient C is used as a parameter that tolerates the systematic outlier in other class C-SVM tolerates less outlier in opponent classification. It holds a uniform value of C parameter for the positive and negative instances that help to satisfy of similar class distribution. Parameter C holds a value for positive and negative instance that satisfy the data set for distribution in classification. SVM interface depends upon the position of support vector. If a support vector found in opposition class then it influences the SVM interface, for this problem an error interface was built for a tolerance of support vector in opposite class. Value of parameter C allows less support-vector in the opposite class.

B. NU-SVM

NU-SVM is classification approach provided by Schölkopf et al. in [27]. It is used to control a large number of support vectors as well as training errors. Parameter v used upper bound and lower bound on the fraction of training errors and support vector respectively. Its range is between 0 and 1.

C. SMO

SMO stands for Sequential Minimal Optimization. The SMO method is proposed by Keerthi et al. [28]. This machine learning classifier used to train the Support Vector Machine. This new learning SVM learning algorithm is very simple, faster, easy to implement and having better-scaling properties. SMO perform well for sparse data set either it is binary or non-binary input data.

Sequential Minimal Optimization algorithm is used to solve the quadric programming problem. This algorithm decomposed the Quadric programming problems into sub-problems. It chooses the smallest optimization problem with two Lagrange multipliers to optimize jointly and find the optimal value for these multipliers. All Quadric programming problems solved quickly due to fast sub-problem.

SMO algorithm is best for avoiding extra use storage memory to store the 2 x 2 matrix. It solves the two Lagrange multipliers analytically. So a very large training problem can be solved in a personal computer that having less memory.

Three components for SMO like two Lagrange multipliers through the analytical method, a heuristic method for multiplier optimization and computing b method. For solving two Lagrange multiplier, this algorithm first computes the constraints and makes a solution for constrained maximum. Multiplier gives the name as script 1 and script 2 to multiplier 1 and 2 respectively that displayed on two-dimension. Constrained maximum lies on the diagonal line and this constraint explains why Lagrange multiplier is optimized.

Sequential Multiplier Optimization always maintains a feasible Lagrange multiplier vector. It increases overall objective function and converges asymptotically. SMO uses a heuristic approach to jointly optimize the Lagrange multipliers. One heuristic approach is for 1st Lagrange multiplier and one for 2nd Lagrange multiplier. The first heuristic approach provides an outer loop for 1st Lagrange multiplier that checks the overall objective function of the training set. If the first approach violates the KKT condition then check second multiplier KKT condition because it jointly optimizes the Lagrange multipliers.

V. Experimental Setup

A. Data Sets Description

Table I describes the biological munch mark databases used for the performance analysis of the decision tree based classifiers and SVMs in this empirical research study. Table I provide the information of data sets in terms of number of attributes, attribute type, number of instances and either missing values exist or not in the corresponding data set. The data sets are selected with significantly variant in database size, number of attributes and number of instances.

TABLE. I. DATA SET DESCRIPTION

Data Sets Name	No. of Attributes	Attributes Type	Missing Value	No. of instance
Lung-cancer	56	Integer	2	32
Lymphography	19	Categorical	None	148
Primary-Tumor	17	Categorical	N/A	339
Breast Cancer Dataset	9	Categorical	None	286
Dermatology	35	Categorical, Integer	Yes	366
Herbarman	3	Integer	None	165
Statlog	13	Categorical, Integer	None	270
Hepatitis	19	Categorical, Integer, Real	Yes	155

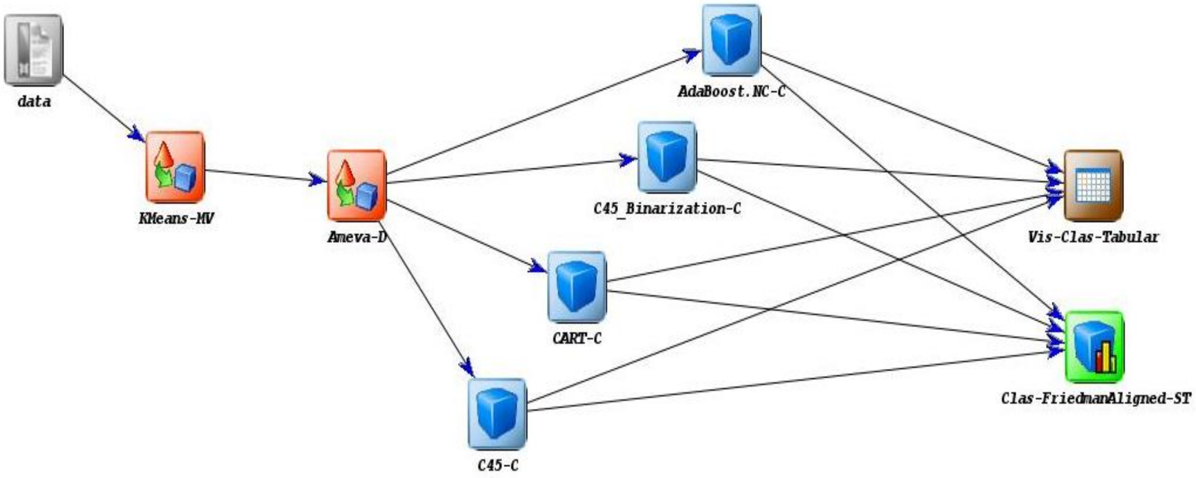


Fig. 1. Experimental Graph.

B. KEEL

Knowledge Extraction based on Evolutionary Learning (KEEL) is a data mining tool possessing various facilities for data preprocessing and different types of classification approaches for the comparison of new proposed classification approaches. It is a freeware java software tool. It provides a user-friendly GUI interface. It contains many built-in dataset and algorithm for data analysis. It provides many preprocessing techniques like feature selection, a method for missing value and hybrid models and statistical method for experiment.it use for educational and research purposes [8].

The current version of Keel has many advance features like multi-instance learning, subgroup discovery, semi-supervised learning and imbalanced classification. These features make versatility of the Keel improved and better deal with new data mining problems [29].

C. Experimental Graph

Fig. 1 shows the experimental graph generated in the KEEL. First stage data set loading, the second stage provide the facility of the imputation of missing values, the third stage provides the module for data discretization, the fourth stage shows the algorithms exploited the empirical study in this paper and final module provide the results of classifiers for the specific databases.

VI. RESULT AND DISCUSSION

This section provides the performance analysis of decision tree based classification approaches and support vector machines on medical databases in terms of accuracy and variance. Furthermore, the performance of a specific classifier is investigated in two fold; on a specific medical database and among the classification approaches.

A. Performance Analysis of Decision Tree based Classifiers

Table II shows the comparative performance analysis of AdaBoost.NC-C C4.5 –C, C4.5_Binarization–C and CART-C tree based classifiers that are chosen in this empirical research study. We compare the performance of these algorithms in Table II on different datasets in term of accuracy. The results show that C45-C and C45_Binarization-C provide equal accuracy on lung-cancer dataset. Moreover, C45-C also perform better on lymph, primary-tumor breast cancer dataset as compared to other algorithms in terms of accuracy. C45_Binarization provide the best performance in term of accuracy on Dermatology and Heart-statlog dataset. The AdaBoost.NC-C provide promising results on Hepatitis dataset; CART-C provides the best performance on Haberman dataset while the C45-C classifier provides 75.19% average accuracy on all datasets that is more promising comparatively w.r.t other classification algorithms. The C45-C_Binerization provide minimum accuracy of 6.06% and maximum accuracy 96.05 in percentage. Table III shows the comparative performance of the selected classifiers in terms of win/lose/draw. The win/lose/draw provides information, how many times a specific algorithm best performs to others.

From Table III, C45-C provides best accuracy on 4 selected datasets with respect to other classifiers. AdaBoost.NC and CART-C provide best accuracy only on one dataset and remaining 7 dataset loose by others algorithm. So AdaBoost.NC and C45_Binarization draw in one dataset.

The application of decision tree based classifier on selected dataset also provides performance in term of variance parallel. More variance on dataset provides lower performance result. CART-C provide bad performance on Lung-Cancer as well as on Dermatology and Hepatitis datasets as compared to AdaBoost.NC-C, C45_C, and C45_Binarization. C45-C provide variance on two selected dataset such as lymph and primary tumor and AdaBoost. NC-C classifier provides more variance on Breast cancer, Haberman, and Heart-Statlog. C45_Binarization –C classifier provides the best performance on selected dataset because there is no more variation as compared to other proposed classifier. CART-C provide 1.01% average variance and maximum 4.27% variance on selected dataset. C45-C provide minimum variance of 0.13% that is more than the other three classifiers. Fig. 2 provides more understandability of this decision tree based classifier's variance.

TABLE. II. DECISION TREE BASED CLASSIFIERS PERFORMANCE IN TERM OF ACCURACY (%)

Data Sets	**AdaBoost.NC-C**	**C45-C**	**C45_Binarization-C**	**CART-C**
lung-cancer	80.30	**83.33**	83.33	81.82
lymph	54.63	**78.10**	6.06	77.97
primary-tumor	19.56	**41.29**	37.79	35.93
breast-cancer	70.47	**74.64**	74.33	70.49
dermatology	45.91	94.05	**96.05**	53.92
haberman	66.51	71.83	71.83	**75.67**
heart-statlog	78.79	81.14	**81.48**	71.38
hepatitis	**80.08**	77.16	77.16	76.10
Average	62.03	**75.19**	66.00	67.91
Min	19.56	41.29	**6.06**	35.93
Max	80.30	94.05	**96.05**	81.82

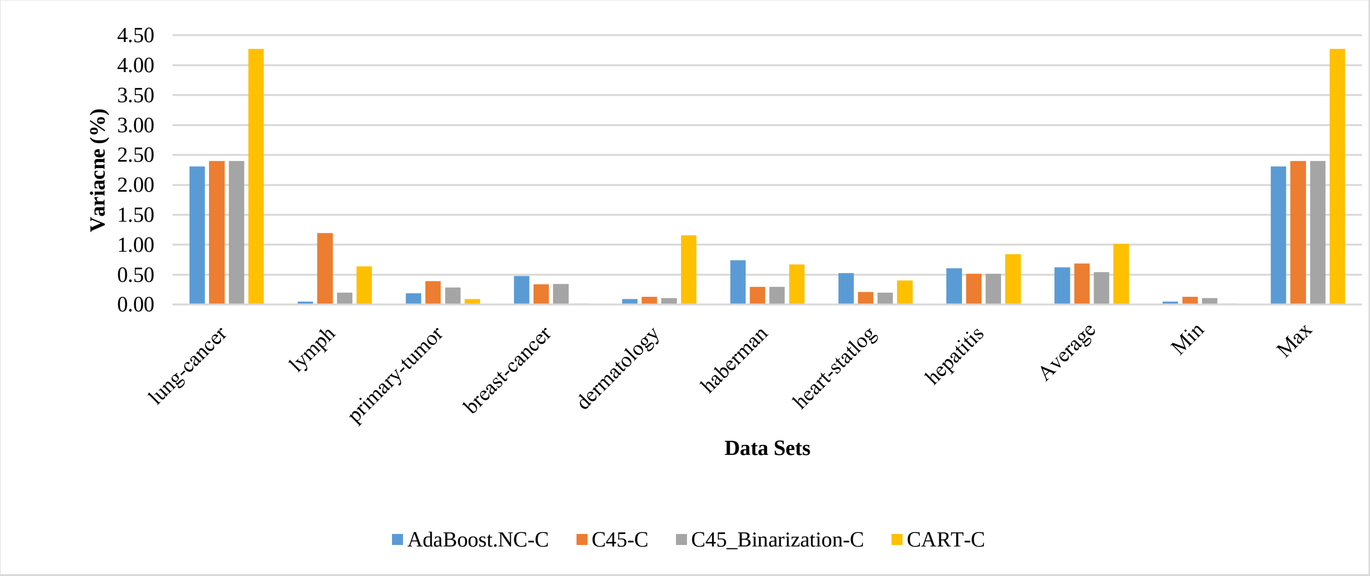


Fig. 2. Performance Analysis of Decision Tree based Classifiers in Term of Variance.

TABLE. III. STATUS COMPARISON OF DECISION TREE BASED CLASSIFIER

	Decision Tree Based Classifiers			
	AdaB.NC	**C45**	**C45_Bin**	**CART**
Win	1	**4**	2	1
Loose	7	**3**	5	7
Draw	0	1	1	0

B. Performance Analysis of SVM based Classifiers

Support Vector Machine performs classification tasks on the base of hyper-plane by using data point that is called support vectors. We used three support vector-based classifier on selected data set by using KEEL software. Table IV presents the results of comparative performance analysis of selected SVMs on corresponding medical databases. Support vector machine based classifier like SMO-C, NU_SVM-C and C_SVM-C are used in this proposed thesis on selected datasets. NU_SVM-C and C_SVM-C classifier provide best performance in term of accuracy on lung-cancer dataset. But C-SVM-C also provide best accuracy on primary tumor and breast cancer datasets 46.12% and 72.11% respectively. SMO-C provide best performance in term of accuracy on dermatology, haberman, heart-statlong, hepatitis and arrhythmia as compared to other two classifier but also proved average accuracy. C-SVM-C classifier provide minimum accuracy 46.12 and maximum accuracy 97.28 % accuracy as compared to other proposed SVM based classifiers. Table V provides the comparative performance of SVM based classifier in terms of win/lose/draw.

Table V shows the status of SVM based Classifiers with their performance for comparison. C_SVM-C classifier give best accuracy four time which is greater from other classifier. NU_SVM-C does not give best performance as compared to other even in one of the selected dataset. SMO-C gives performance in three classifiers.

Fig. 3 provides the comparative performance analysis of SVMs in terms of variance. SMO-C provide more variance on primary-tumor, Dermatology and hepatitis than NU_SVM-C and C_SVM-C and also provide minimum variance of selected variance as compared to other two classifiers. NU_SVM-C provide more variance on six datasets that make the performance bad on selected dataset as compared to other datasets. It also make more value of average variance on selected datasets that reach 1.26. NU_SVM-C and C_SVM-C provide equal maximum variance on selected datasets; as well as equal variance on lung-cancer dataset. All the information is highlighted in Table V.

Table VI provides the combined performance behavior of both categories Decision Tree-based classifiers and SVMs based classifiers in terms of accuracy. The performance of AdaBoost.NC-C classifier is lower than other methods on selected datasets. The C45-C provide the best performance on based of accuracy on lung-cancer and breast cancer datasets. C45_Binarization-C provide best accuracy result on Lung-cancer dataset equal to C45-C and Minimum average accuracy 6.06% as compared to remain six classifiers. CART-C provided the best performance on based of accuracy on Haberman dataset as compared to another dataset. Support Vector based algorithm SMO-C provided the best performance on based of accuracy on lymph, heat-statlog and hepatitis dataset and provided average accuracy as compared to other Decision tree and SVM based algorithms. NU_SVM-C accuracy is low to other both classifiers. C_SVM-C SVM based classifier provides the best performance on based of accuracy on primary tumor dermatology and provides maximum accuracy as compared to other classifiers on selected datasets.

TABLE. IV. PERFORMANCE ANALYSIS OF SVM BASE CLASSIFEIRS IN TERMS OF ACCURACY (%)

Data Set	**SMO**	**NU_SVM**	**C_SVM-C**
lung-cancer	70.45	**73.48**	**73.48**
lymph	**81.08**	71.90	75.54
primary-tumor	44.76	33.80	**46.12**
breast-cancer	69.25	61.95	**72.11**
dermatology	95.79	97.03	**97.28**
haberman	**75.09**	50.36	73.62
heart-statlog	**84.85**	73.06	82.49
hepatitis	**85.91**	81.78	85.38
arrhythmia	**62.03**	49.18	51.42
Average	**74.36**	65.84	73.05
Min	44.76	**33.80**	46.12
Max	95.79	97.03	**97.28**

TABLE. V. COMPARISION IN TERMS OF WIN/LOSE/DRAW

	Support Vector Machine Based Classifiers		
	SMO	**NU_SVM**	**C_SVM**
Win	**6**	0	4
Loose	3	7	3
Draw	0	1	1

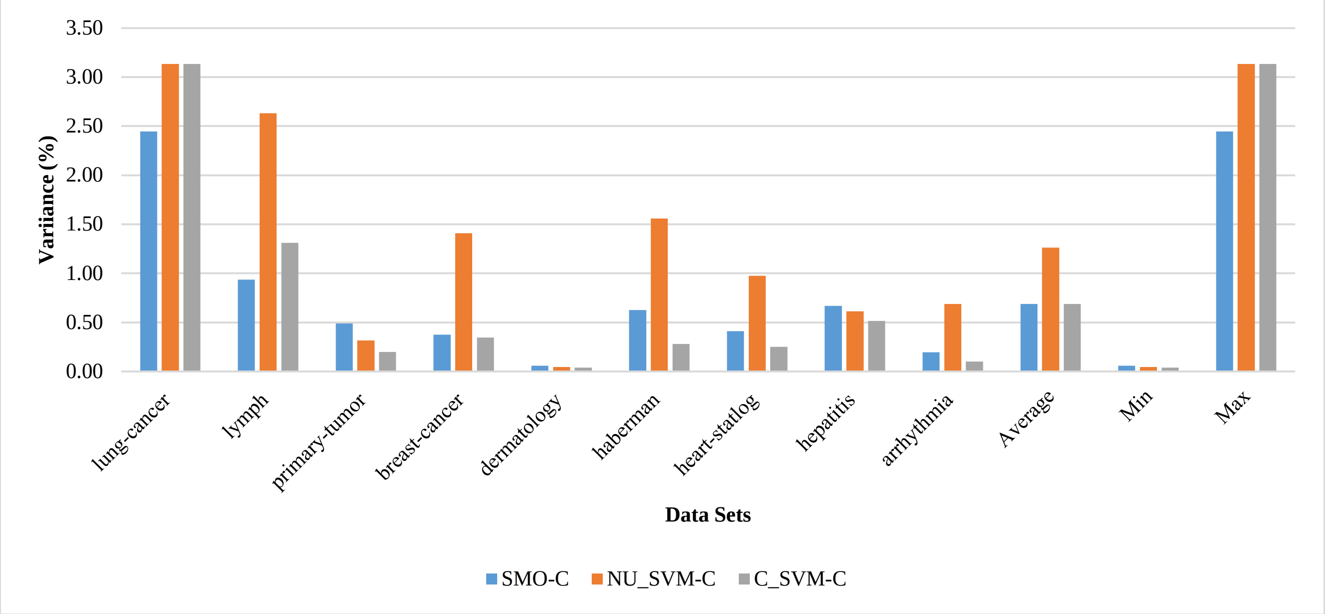


Fig. 3. Performance Analysis of Support Vector Machines in Term of Variance.

TABLE. VI. COMBINED RESULTS OF PROPOSED ALGORITHMS IN TERM OF ACCURACY

Data Sets	Decision Tree Algorithms				Support Vector Machines		
	AdaBoost.NC-C	C45-C	C45_Binarization	CART-C	SMO-C	NU_SVM-C	C_SVM-C
lung-cancer	80.30	**83.33**	83.33	81.82	70.45	73.48	73.48
lymph	54.63	78.10	6.06	77.97	**81.08**	71.90	75.54
primary-tumor	19.56	41.29	37.79	35.93	44.76	33.80	**46.12**
breast-cancer	70.47	**74.64**	74.33	70.49	69.25	61.95	72.11
dermatology	45.91	94.05	96.05	53.92	95.79	97.03	**97.28**
haberman	66.51	71.83	71.83	**75.67**	75.09	50.36	73.62
heart-statlog	78.79	81.14	81.48	71.38	**84.85**	73.06	82.49
hepatitis	80.08	77.16	77.16	76.10	**85.91**	81.78	85.38
Average	62.03	75.19	66.00	67.91	**75.90**	67.92	75.75
Min	19.56	41.29	6.06	35.93	44.76	33.80	46.12
Max	80.30	94.05	96.05	81.82	95.79	97.03	**97.28**

VII. CONCLUSION

Classification Rule Discovery from medical databases is a very hot and challenging problem in the field of Data Mining. There are several classification approaches proposed for the discovery of classification rules and prediction of diseases from medical databases. The choice of a classification method for the discovery of classification rules from specific medical databases still requires investigation of the suitability of classifiers in terms of performance analysis. This study investigates the performance of decision tree-based classifiers and Support Vector Machines on specific medical databases. The empirical performance analysis results reveal that C45-C performs better in terms of a total number of datasets while the overall average performance of C45_Binarization-C is better than other decision tree-based classifiers. The performance of SVM based classifiers, SMO-C is results are promising to NU_SVM-C and C_SVM-C in terms of accuracy. This research work provides the empirical performance analysis of decision tree-based classifiers and SVM on a specific dataset. Moreover, this paper provides a comparative performance analysis of classification approaches in terms of statistics.

In the future, this research work can be enhanced by increasing the number of medical databases with other statistical and evolutionary classifiers.

REFERENCES

[1] O. Trelles, P. Prins, M. Snir, and R. C. Jansen, “Big data, but are we ready?,” Nature Reviews Genetics, vol. 12, no. 3, pp. 224, 2011.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "Knowledge Discovery and Data Mining: Towards a Unifying Framework." pp. 82-88.

[3] D. A. Benson, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank,” Nucleic acids research, vol. 42, no. D1, pp. D32-D37, 2013.

[4] T. W. Harris, J. Baran, T. Bieri, A. Cabunoc, J. Chan, W. J. Chen, P. Davis, J. Done, C. Grove, and K. Howe, “WormBase 2014: new views of curated biology,” Nucleic acids research, vol. 42, no. D1, pp. D789-D793, 2013.

[5] C. E. Bouton, A. Shaikhouni, N. V. Annetta, M. A. Bockbrader, D. A. Friedenberg, D. M. Nielson, G. Sharma, P. B. Sederberg, B. C. Glenn, and W. J. Mysiw, “Restoring cortical control of functional movement in a human with quadriplegia,” Nature, vol. 533, no. 7602, pp. 247, 2016.

[6] F. Thabtah, and D. Peebles, “A new machine learning model based on induction of rules for autism detection,” Health informatics journal, pp. 1460458218824711, 2019.

[7] R. Saravanan, and P. Sujatha, "A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification." pp. 945-949.

[8] J. Alcalá-Fdez, L. Sánchez, S. Garcia, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit, and V. M. Rivas, “KEEL: a software tool to assess evolutionary algorithms for data mining problems,” Soft Computing, vol. 13, no. 3, pp. 307-318, 2009.

[9] Z. Iqbal, R. Ilyas, W. Shahzad, and I. Inayat, "A comparative study of machine learning techniques used in non-clinical systems for continuous healthcare of independent livings." pp. 406-411.

[10] S. O. Akinola, and O. J. Oyabugbe, “Accuracies and training times of data mining classification algorithms: An empirical comparative study,” Journal of software Engineering and Applications, vol. 8, no. 09, pp. 470, 2015.

[11] M. H. Tafish, and A. M. El-Halees, "Breast Cancer Severity Degree Predication Using Data Mining Techniques in the Gaza Strip." pp. 124-128.

[12] M. Ramasamy, S. Selvaraj, and M. Mayilvaganan, "An empirical analysis of decision tree algorithms: Modeling hepatitis data." pp. 1-4.

[13] Y. Yan, M. Chen, M.-L. Shyu, and S.-C. Chen, "Deep learning for imbalanced multimedia data classification." pp. 483-488.

[14] A. Anand, M. A. Haque, J. S. R. Alex, and N. Venkatesan, "Evaluation of Machine learning and Deep learning algorithms combined with dimentionality reduction techniques for classification of Parkinson’s Disease." pp. 342-347.

[15] S. Belarouci, F. Bekaddour, and M. A. Chikh, "A comparative study of medical data classification based on LS-SVM and metaheuristics approaches." pp. 548-553.

[16] S. Tharaha, and K. Rashika, "Hybrid artificial neural network and decision tree algorithm for disease recognition and prediction in human blood cells." pp. 1-5.

[17] D. A. Aljawad, E. Alqahtani, A.-K. Ghaidaa, N. Qamhan, N. Alghamdi, S. Alrashed, J. Alhiyafi, and S. O. Olatunji, "Breast cancer surgery survivability prediction using bayesian network and support vector machines." pp. 1-6.

[18] P. Hamsagayathri, and P. Sampath, "Priority based decision tree classifier for breast cancer detection." pp. 1-6.

[19] J. R. Quinlan, “Induction of decision trees,” Machine learning, vol. 1, no. 1, pp. 81-106, 1986.

[20] J. Quinlan, “C4. 5: Programs for machine learning. Morgan Kaufmann, San Francisco,” C4. 5: Programs for machine learning. Morgan Kaufmann, San Francisco., pp. -, 1993.

[21] S. Wang, and X. Yao, “Multiclass imbalance problems: Analysis and potential solutions,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1119-1130, 2012.

[22] L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and regression trees (Chapman y Hall, Eds.),” Monterey, CA, EE. UU.: Wadsworth International Group, 1984.

[23] R. Gholami, and N. Fakhari, "Support Vector Machine: Principles, Parameters, and Applications," Handbook of Neural Computation, pp. 515-535: Elsevier, 2017.

[24] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, C.-C. Lin, and M. D. Meyer, “Package ‘e1071’,” The R Journal, 2019.

[25] R. Gandhi, “Support Vector Machine—Introduction to Machine Learning Algorithms,” Towards Data Science, 2018.

[26] C. Cortes, and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, pp. 273-297, 1995.

[27] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, “New support vector algorithms,” Neural computation, vol. 12, no. 5, pp. 1207-1245, 2000.

[28] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, “Improvements to Platt's SMO algorithm for SVM classifier design,” Neural computation, vol. 13, no. 3, pp. 637-649, 2001.

[29] I. Triguero, S. González, J. M. Moyano, S. García López, J. Alcalá Fernández, J. Luengo Martín, A. Fernández Hilario, J. Díaz, L. Sánchez, and F. Herrera, “KEEL 3.0: an open source software for multi-stage analysis in data mining,” 2017.

Computer-based Approach to Detect Wrinkles and Suggest Facial Fillers

Amal Alrabiah[1], Dr. Mai Alduailij[2]
College of Computer and Information Sciences
Princess Nourah bint Abdulrahman University
Riyadh, Saudi Arabia

Dr. Martin Crane[3]
School of Computing
Dublin City University
Dublin, Ireland

***Abstract*—Modern medical practice has embraced facial filler injections as part of the innumerable cosmetic procedures that characterize the current age of medicine. This study proposed a novel methodological framework. The Inception model is the core of the framework. By carefully detecting the classification of wrinkles, the model can be built for different applications to aid in the detection of wrinkles that can objectively help in deciding if the forehead area needs to have filler injections. The model achieved an accuracy of 85.3%. To build the Inception model, a database has been prepared containing face forehead images, including both wrinkled and non-wrinkled face foreheads. The face image pre-processing is the first step of the proposed framework, which is important for reliable feature extraction. First, in order to detect the face and facial landmarks in the image, a Multi-task Cascaded Convolutional Networks model has been used. Before feeding the images into the deep learning Inception model for classifying whether the face foreheads have wrinkles or no wrinkles, an image cropping process is required. Given the bounding box and the facial landmarks, face foreheads can be cropped accurately. The last step of the proposed methodology is to retrain an Inception model for the new categories (Wrinkles, No Wrinkles) to predict whether a face forehead has wrinkles or not.**

***Keywords*—Deep learning; classification; facial fillers; wrinkle detection**

I. Introduction

Filler injections are a modern cosmetic procedure and have been widely embraced by women and men alike because of their wonderful ability to create fuller cheeks, lips, and other facial features. Filler injections are also used to reduce the effects of wrinkles around the mouth, eyes, and eyebrows and to hide any scars that may be causing an individual to feel self-conscious and unattractive.

Human beings can without much difficulty see what the image represents. As an example, humans can easily see that the image contains a number of objects and can detect faces in an image as well as distinguish between the different features of the face. Computer Systems, on the other hand, can have difficulties. Computers cannot easily see whether the image contains objects or not. Also, cannot easily detect human faces and facial features.

As many researchers are trying to set up computers with artificial intelligence capabilities to be able to serve patients with different health problems such as diabetes, blood pressure, and obesity [1]. Artificial intelligence databases can help in easily identifying whether there are similar cases that have been registered before, and then returns the diagnosis and treatment of these similar cases [1]. There is still a dearth in research studies regarding how computer vision can be deployed in facial filler injections.

This study presents a novel methodological framework. The Inception model is the core of the framework. With carefully detecting the classification of wrinkles, the model can be built for different application to aid the detection of wrinkles to that can objectively help in determining if the forehead area needs to have filler injections.

II. Related Work

Previous studies have also proposed a new algorithm referred to as Hessian Line Tracking (HLT) for detecting wrinkles [2]. The researchers began with a group of seeds that the researchers extracted from the Hessian Matrix’s ridge area before proceeding to apply a multi-scale tracking system recursively to all the seeds. After completion, the researchers validated each pixel confidence over the scales with the objective of producing an initial map of wrinkles. The last step involved post-processing in which the researchers carried out a series of mini-steps including median and directional filtering as well as area thresholding in order to reduce noise [2]. In the experimental set-up, the researchers employed the services of three number of coders instructed to manually annotate the wrinkle on 100 cropped images of the forehead– the images were extracted from the Bosphorus dataset [3]. The dataset contains forehead wrinkles of varying sizes. Particularly, the dataset that was employed in the study contained 106 subjects from whom the researchers used an ordinary camera to fetch 2-D facial images under good flat, illuminated conditions. Although the researchers took several images of varying poses and facial expressions of each subject, the experiment only used frontal images. The researchers reported both intra and inter-reliability with regard to the manual annotation process – reliability was 94% and above [2]. Wrinkles typically appear in a wide variation in both images, in pattern, length, and width as well as within the same image. This significantly challenges the generation of an automatic wrinkle detection operator. This justified the development of a multiscale HLT – an approach based on seed extraction by Hybrid Hessian Filter (HHF) as well as multiscale tracking for overcoming the weakness of HHF while also making it possible to capture wrinkle variability in the entire image.

In another study researchers also proposed a new algorithm for automatic tracking of linear, fixed and chaotic forms of transient wrinkles [4]. For the automatic analysis of wrinkles, the researchers came up with two clusters of wrinkles including transient and permanent. While the latter are usually found on the faces of older individuals, the former on the other hand often appear in relatively wider regions in the course of generating an expression. The research was divided into two parts whereby while the first part explored an algorithm for detecting transient wrinkles, the second part dwelled on its application. The proposed wrinkle detector was made up of three steps. The first step involved the Canny edge detector that the authors applied to the input face for detecting pairs of continuous wrinkles. This was followed by applying an active appearance model to locate all candidate wrinkle lines. This generated data for constructing the structure of the wrinkle. In the third step, the researchers defined quantitative metrics which they subsequently used for Support Vector Machine (SVM) classification – this step was critical in helping to discriminate regions of the face with wrinkles from those without. Despite the fact that competitive results are achieved by the proposed transient wrinkle detector model that the researchers propose which is also the case with improved wrinkle mapping, there is a number of areas in which future studies should focus. A case in point regards long wrinkles in the forehead, for which [4] employed the five points in the wrinkle structure–future studies should add on more points which will go a long way towards improving the approximation of accuracy of the wrinkle edge.

Batool and Chellapa [5], in their work presented a quick deterministic algorithm based on image morphology as well as Gabor filters with the aim of improving localization results. The researchers proposed features derived from Gabor filter bank–the aim here was to shed light on the subtle curvilinear discontinuities in the texture of the skin attributed to wrinkles. The researchers then employed image morphology for integrating geometric constraints to localize curvilinear wrinkle shapes at the locations of wrinkles of pronounced Gabor filter responses. Experiments were carried out at two sets of images including those with high and low resolutions before the researchers compared the results to those generated from Marked Point Processes (MPP). Experiments illustrate that the suggested algorithm is not only faster compared to the MPP framework, but also generates the merit of visually satisfactory results.

In the last study analysed in the review, researchers compounded texture orientation fields with Gabor filter responses to detect wrinkles [6]. In the experiment, a bimodal Gaussian Mixture Model (GMM) described the distribution of normal skin verse skin imperfection of Gabor features [6]. The researchers then proceeded to employ a Markov random field model to integrate the spatial relationships for their texture orientations between adjacent pixels as well as GMM distribution. To classify skin versus skin imperfections, the study employed an expectation-maximization algorithm. As opposed to blending or blurring the detected wrinkles, the study removed them completely. The exemplar-based constrained texture synthesis algorithm, as the researchers conclude is the most ideal tool for in-painting irregularly shaped gaps that are left behind in the form of scars of removed wrinkles [6]. Overall, the experiment illustrates that most skin imperfections and wrinkles are usually detected and in-painted. Nevertheless, there are a few areas on the face that have less contrast to the skin around it that fails to be detected. A case in point includes small parts of the upper parts of an individual's forehead. Moreover, the experiment also reported challenges of the impact of sagging skin and aging, illumination, as well as artefacts resulting from the repetition of patches. Particularly, facial images of subjects with sagging skin that appears alongside wrinkles according to the researchers posed significant challenges. Such is the case since the same patch of wrinkled skin is selected as a source of skin texture resulting in repeated patterns of imperfection.

III. Background and Methodology

A. Deep Learning

Deep learning is one of the major theories of machine learning that is founded on learning data representation and not task-related algorithms. It was first introduced in 1986 to the community of machine learning by Rina Dechter [7]. Learning using this theory might be unsupervised, supervised or a hybrid of both. As a branch of machine learning, deep learning is inspired by the human brain's main function and structure. The brain constitutes of neural networks whose interconnected neurons plays a crucial role in processing and transmitting signals from one neuron to the other. On the basis of this operation, the founder of deep learning, Geoffrey Hinton [8], made some artificial neural networks comprising of man-made neurons that could easily conduct operation as well as process the required information. The three layers of the neural network in deep learning include the input, hidden and output layers. The input layer is responsible for accepting a variety of input using formats such as audio, picture, number or text. The hidden layer conducts mathematical functions, feature extraction and manipulation of data. The output layer, on the other hand, is essential in getting the desired final output [8].

B. Transfer Learning

In 1993, Lorien Pratt came up with algorithm that was founded on discriminability-based transfer [9]. In so doing, he gave a platform through which the transfer learning theory was born. In this method, a model that has been formulated for an activity is normally reutilized as a beginning point for some other second task. For instance, the knowledge that one gets when learning to distinguish a jet could be useful in recognizing a helicopter or a spacecraft. Therefore, this model concentrates on the storage of acquired knowledge while at the same time striving to be utilized in another similar scenario or problem [10]. This theory is used in deep learning particularly in instances where the pre-trained models are employed as beginning points on the natural language and computer vision. In comparisons to the conventional machine learning, transfer learning utilizes these pre-trained models that were useful in another instance to kick start the process of development of the new problem or task. As stated earlier, this model depends on the task and domain concepts [10].

C. Multi-Task Cascaded Convolutional Neural Network

Multi-Task Cascaded Convolutional Neural Network model or simply MTCCNN is a deep learning algorithm-based model that comprises of three stages that identify faces' and bounding box in a particular image as well as the main five-point landmarks on an individual's face [11]. Back in 2001, researchers attempted to put forward a method of forward-cascade detection founded on the features of AdaBoost and Haar to conduct cascade classifiers [12]. Kaipeng Zhang later introduced the MTCNN theory which in addition to having three phases, helped in the detection of the bounding box, points, and landmarks [11].

For every MTCNN phase that the image goes through, the investigator sees an improvement in quality. The input goes to the CNN which not only give it a particular score but also return a bounding box. The initial stage ensures that the input is scaled downwards. The CNN also ensures the facilitation of the MTCNN in making a pyramid of a picture in question. The next stages involve the extraction of the patches of this picture for every bounding box. It is then resized and resized even more in the third stage. Other than bounding the image onto the box and later assigning some score, this stage also computes the points of the five face landmarks for every bounding box [11].

D. Inception-v3 Model

Inception V3 is one of the most popular image-recognition deep learning models. It has been a culmination of multiple ideas by a pool of researchers in the years. Inception V3 was theorized by Szegedy in a paper that rethought the Inception concept in computer vision [13]. In itself, this model is composed of both asymmetric and symmetric building blocks. They include convolutions, max, and average pooling as well as fully linked layers. Throughout the Inception –v3 model, Batchnorm is extensively utilized in addition to being applied to the activation inputs. On the other hand, Softmax is usually used in computing the loss. The two parts of Inception-v3 model include the feature extraction component and the classification component. The former components are reliant on the convolutional neural networks while the latter is reliant on softmax and fully-linked layers [13].

E. Dataset

The FERET database [14], [15] is used in this study to experiment and evaluate the performance of the proposed methodology. The FERET database was established to support machine learning algorithms in both development and evaluation. The database contains 14,126 images associated with 1,199 people.

IV. Experiments and Results

A. Overview of the Proposed Methodology Framework

As shown in Fig. 1, the core of the framework is the Inception model. To build the Inception model, a database has been prepared containing 618 cleaned face forehead images, including both wrinkled and non-wrinkled face foreheads. The face image pre-processing is the first step of the proposed framework, which is important for reliable feature extraction. First, the face and facial landmarks are detected in the image using a Multi-task Cascaded Convolutional Networks model [11], known for its strong and accurate abilities to detect faces and facial landmarks quickly. Before feeding the images into the deep learning Inception model for classifying whether the face foreheads have wrinkles or no wrinkles, an image cropping process is required. Given the bounding box and the facial landmarks, face foreheads can be cropped accurately. Next, k-means algorithm has been used to separate 1,199-image data set of face foreheads into two clusters; then went through each cluster and separated wrinkled foreheads from non-wrinkled foreheads. Also removed the foreheads covered by hair (581 images), then labelled images Wrinkles (309 images) and No-Wrinkles (309 images). The last step of the proposed methodology is to retrain an Inception model for the new categories (Wrinkles, No Wrinkles) to predict whether a face forehead has wrinkles or not.

B. Dataset and Pre-Processing Details

The FERET database [14], [15] is used in the implementation experiments to evaluate the performance of the proposed methodology. The FERET database was established to support machine learning algorithms in both development and evaluation. The database contains 14,126 images associated with 1,199 people. In the experiments only the frontal face images have been used from the database, which amounted to 1,199 unlabelled images.

To detect faces in an image and discover the location of different facial features a pre-trained Multi-task Cascaded Convolutional Networks has been used [11]. Face alignment and face detection are executed jointly in a multi-task training method, enabling the model to properly detect faces and locate five points of facial landmarks.

By default, the pretrained Multi-task Cascaded Convolutional Networks packaged with a face detection weights model. The detector passes a list of Javascript object notation objects. Each JavaScript object notation object carries three main keys: 'confidence', 'keypoints', and 'box':

- The confidence is how probable it is that a bounding box will match a face.
- The box is put into a format [x, y, W, H] which can produce the bounding box around the face.
- The keypoints are arranged into a Javascript object notation object with the keys 'eyeLeft', 'eyeRight', 'nose', 'leftMouth', and 'rightMouth'. Every keypoint is recognized by the position of the pixel (x, y).

Fig. 2 shows a result of an image with a bounding box around the face and five landmarks located on the eyes, mouth, and nose after it has been detected by the MTCNN model. An image cropping process is required. Given the bounding box and the facial landmarks, face foreheads can be cropped accurately.

As regards the labelling process and cleaning data, the manual labelling of images would be a hurdle, so the k-means unsupervised learning clustering algorithm has been used to help with this process.

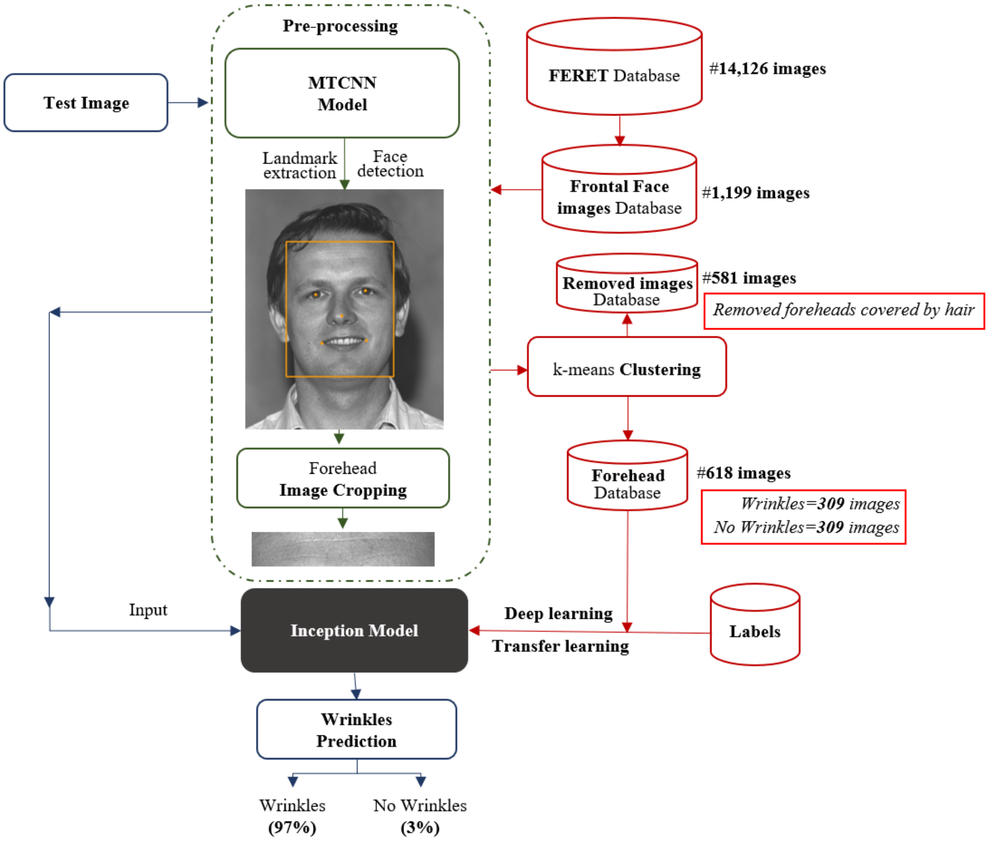


Fig. 1. Overview of the Proposed Methodology.

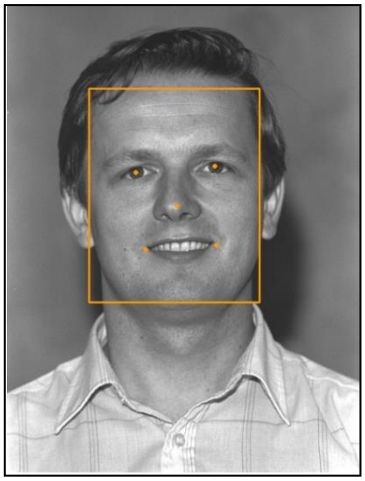


Fig. 2. The Result of MTCNN Detector.

Data clustering is the process of placing data in similar clusters, which is a branch of data mining. The clustering algorithm divides a data set into several groups since the similarity between points within a particular cluster is greater than the similarity between two points within two different clusters [15]. The idea of data clustering is simple in nature and very close to human in its way of thinking. Whenever people deal with a large amount of data, they tend to summarize the huge amount of data into a few groups or categories in order to facilitate analysis.

So, to label 1,199-image data set of face forehead images into two categories (1- Wrinkles, 2- No Wrinkles), the k-means clustering method has been used to do the job. The k-means algorithm has been used to separate 1,199-image data set of face foreheads into two clusters. After that went through each cluster and separated wrinkled foreheads from non-wrinkled foreheads. Also removed the foreheads covered by hair (581 images), then labelled images into Wrinkles (309 images), and No-Wrinkles (309 images). Finally, an experienced dermatologist reviewed the labelling of the images to ensure they were labelled correctly.

C. Wrinkle Detection

Advanced image classification models have hundreds of parameters. Also, training convolutional neural networks from scratch requires a set of labelled training data and a huge computing power capacity (GPU). Transfer learning is a method that shortcuts many of this by using a model that has previously been trained on a similar task and data and then reusing it in a new model.

In this study, a deep learning image classifier has been retrained for new categories (Wrinkles, No Wrinkles) and made reuse of the capabilities of feature extraction from the powerful image classifier Inception-v3 [13] which was trained on ImageNet and then retrained using a different classification layer on top. The power of transfer learning is that lower layers that become trained to detect among some objects can be reused for many classification tasks without any modification.

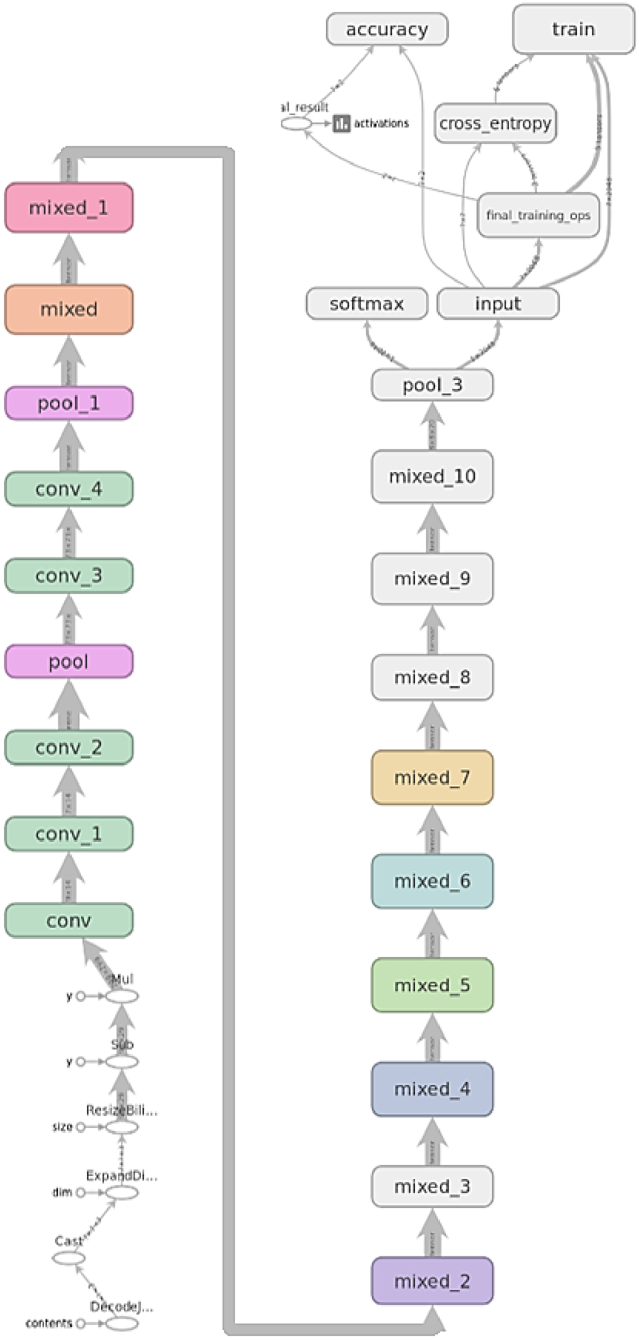


Fig. 3. Retrained Inception Model main Graph.

The aforementioned pretrained model is useful for various applications, works with reasonable training data amounts (hundreds rather than thousands of labelled data), and can be run in as short as ten minutes on a regular laptop without a GPU.

The image data set has been divided up into three separate sets. The largest was the training set, which implies all of the images were fed in the Inception convolutional neural network. The results were utilized to adjust the weights of the Inception model during training. The first stage analyses all of the images for each class in the folders and then the bottleneck values are calculated and stored for each of them. 'Bottleneck' is a non-formal phrase that is often meant for the layer that carries out the classification before the last output layer. This second-to-last layer has been trained to produce a collection of values that is fair enough to be used by the classifier to discriminate between all of the classes it has been requested to classify [13]. By the time that the bottlenecks values are completed, the real training of the neural network's top layer begins. The learning process is carried out for 500 training steps. Every step takes ten images randomly from the data set for the training, gets their bottlenecks out of the cache, and feeds them inside the last layer to make predictions. Then by the back-propagation process, the last layer's weights get updated. The retrained Inception model graph is shown in Fig. 3.

Some of the image data has been kept outside of the training process so that the Inception model could not memorize them. After that, these data images have been used as an investigation to make sure that overfitting was not happening; if it shows good accuracy on these it's a genuine sign that the network is not overfitting. The data set has been split, with 80% of the images put into the training set and 10% put aside to run as validation regularly during training, ending up with 10% used as a testing set which can predict the real-world performance of the Inception model classifier. Sample results of wrinkle predictions done on test data are shown in Fig. 4.

The final test accuracy = 85.3% with error rate = 14.7%. Test evaluation is the best estimation of how the re-trained Inception model will perform on the classification task.

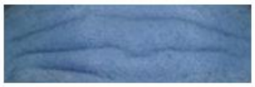
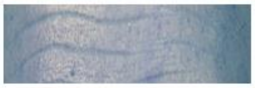
Input Image (Forehead)	Wrinkles Prediction Score	No Wrinkles Prediction Score
	100%	0%
	20.8%	79.2%
	88.6%	11.4%
	18.5%	81.5%
	98%	2%
	1.5%	98.5%

Fig. 4. Sample Results of Wrinkle Predictions Done on Test Data.

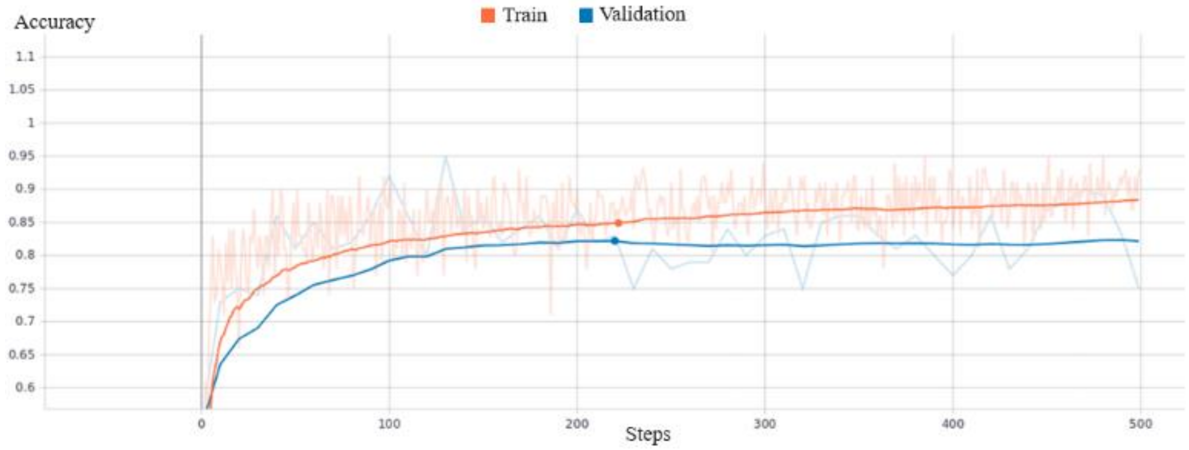


Fig. 5. The Learning Curves for the Training and the Validation of Inception Model.

In order to ensure that overfitting is not occurring in the re-trained Inception model with the new dataset and labels, as shown in Fig. 5, a line plot that displays the learning curves for the training and the validation of the Inception model. The curve for the training set is slightly more accurate than the test set to avoid overfitting.

V. Discussion

One of the main limitations of this study that have been encountered during the investigations of the proposed research is the lack number of previous studies that explore how computer vision methods can help in the facial cosmetic healthcare field.

Furthermore, the resolution quality of FERET images database used in this study is low compared to today's image resolutions also the time constraint allowed only to study one area of the face which is forehead.

VI. Conclusion and Future Work

Experiments accomplished in the study investigate a proposed novel methodological framework. The Inception model is the core of the framework. By carefully detecting the classification of wrinkles, the model can be built for different applications to aid in the detection of wrinkles that can objectively help in deciding if the forehead area needs to have filler injections. The model achieved an accuracy of 85.3%. To build the Inception model, a database has been prepared containing face forehead images, including both wrinkled and non-wrinkled face foreheads. The face image pre-processing is the first step of the proposed framework, which is important for reliable feature extraction. First, the face and facial landmarks are detected in the image using a Multi-task Cascaded Convolutional Networks model known for its strong and accurate abilities to detect faces and facial landmarks quickly. Before feeding the images into the deep learning Inception model for classifying whether the face foreheads have wrinkles or no wrinkles, an image cropping process is required. Given the bounding box and the facial landmarks, face foreheads can be cropped accurately. The last step of the proposed methodology is to retrain an Inception model for the new categories (Wrinkles, No Wrinkles) to predict whether a face forehead has wrinkles or not.

Since there is a lack in the previous studies there is much to do in the future research studies. For the presented research, there is only one part has been investigated so the other area of the face to be investigated such wrinkles around the eyes and mouth.

Semantic segmentation methods can be used to label each pixel of the wrinkles, and fine lines in the face so that the person can know exactly the location of the areas that need to be filled with filer injection instead of classifying whether an area of the face has wrinkles or not. For more complicated applications it could also be used two frontal face pictures of the same person, one image with a blank expression and the other image with happy facial expression. The model will then compare those two images against each other and then decide if the wrinkles show only in the happy facial expression then it's transient wrinkles and have to be filled with Botox, if the wrinkles appear in the blank facial expression then it needs to be filled with natural substances such as body fat or collagen and artificial substances such as hyaluronic acid. Essentially that would be like an artificial intelligent cosmetic consultant.

Acknowledgment

"Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office" [14], [15].

References

[1] P. Hamet and J. Tremblay, "Artificial intelligence in medicine", Metabolism, vol. 69, pp. S36-S40, 2017. Available: 10.1016/j.metabol.2017.01.011.

[2] C. Ng, M. Yap, N. Costen and B. Li, "Wrinkle Detection Using Hessian Line Tracking", IEEE Access, vol. 3, pp. 1079-1088, 2015. Available: 10.1109/access.2015.2455871.

[3] A. Savran, B. Sankur and M. Taha Bilge, "Regression-based intensity estimation of facial action units", Image and Vision Computing, vol. 30, no. 10, pp. 774-784, 2012. Available: 10.1016/j.imavis.2011.11.008.

[4] W. Xie, L. Shen and J. Jiang, "A Novel Transient Wrinkle Detection Algorithm and Its Application for Expression Synthesis", IEEE Transactions on Multimedia, vol. 19, no. 2, pp. 279-292, 2017. Available: 10.1109/tmm.2016.2614429.

[5] N. Batool and R. Chellappa, "Fast detection of facial wrinkles based on Gabor features using image morphology and geometric constraints",

Pattern Recognition, vol. 48, no. 3, pp. 642-658, 2015. Available: 10.1016/j.patcog.2014.08.003.

[6] N. Batool and R. Chellappa, "Detection and Inpainting of Facial Wrinkles Using Texture Orientation Fields and Markov Random Field Modeling", IEEE Transactions on Image Processing, vol. 23, no. 9, pp. 3773-3788, 2014. Available: 10.1109/tip.2014.2332401.

[7] R. Dechter, "Learning while searching in constraint-satisfaction-problems", Proceeding AAAI'86 Proceedings of the Fifth AAAI National Conference on Artificial Intelligence, pp. 178-183, 1986.

[8] A. Krizhevsky, I. Sutskever and G. Hinton, "ImageNet classification with deep convolutional neural networks", Communications of the ACM, vol. 60, no. 6, pp. 84-90, 2012. Available: 10.1145/3065386.

[9] L. Y. Pratt, "Discriminability-based transfer between neural networks", Proc. Adv. Neural Inf. Process. Syst., pp. 204-211, 1993.

[10] K. Weiss, T. Khoshgoftaar and D. Wang, "A survey of transfer learning", Journal of Big Data, vol. 3, no. 1, 2016. Available: 10.1186/s40537-016-0043-6.

[11] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks", IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, 2016. Available: 10.1109/lsp.2016.2603342.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001. Available: 10.1109/cvpr.2001.990517.

[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. Available: 10.1109/cvpr.2016.308.

[14] P. Phillips, H. Wechsler, J. Huang and P. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms", Image and Vision Computing, vol. 16, no. 5, pp. 295-306, 1998. Available: 10.1016/s0262-8856(97)00070-x.

[15] D. MacKay, Information theory, inference, and learning algorithms. Cambridge: Cambridge University Press, 2003, pp. 284–292.

On Some Methods for Dimensionality Reduction of ECG Signals

Monica Fira[1]
Institute of Computer Science
Romanian Academy Iasi, Romania

Liviu Goras[2]
Institute of Computer Science, Romanian Academy
“Gheorghe Asachi” Technical University of Iasi, Romania

***Abstract*—Dimensionality reduction with two methods, namely, Laplacian Eigenmap (LE) and Locality Preserving Projections (LPP) is studied for normal and pathological noisy and noiseless ECG patterns. Besides, the possibility of using compressed sensing (CS) as a method of dimensionality reduction is also analyzed. The classification rate for the initial domain as well as in manifolds of various dimensions for the three cases are presented and compared.**

Keywords*—Dimensionality reduction; compressed sensing; electrocardiography (ECG)

I. Introduction

In the last years, dimensionality reduction methods have been widely investigated. The main idea about these methods is to represent high-dimensional raw data on intrinsic lower dimensional spaces. The main targets are either to reduce the computation costs for the raw data or to represent the data in a friendlier manner.

Wireless biomedical sensors are easy to use for long-term monitoring, especially outside the hospital, offering the possibility of better results able to substantially improve the patient's health and quality of life. Since electrocardiography (ECG) signals recorded from the electrical activity of the heart over a period of time have been used for diagnosis in many diseases, ECG tele-monitoring is accepted as an encouraging method in tele-medicine. Considering the importance of ECG signal recording, we are confronted with the problem of storing, transmitting and processing/classification the signals preferably in real time. One of the solutions to all these problems is the application of the dimensionality reduction of data.

Several reasons for using dimensionality reduction are [1]:

- The space required for storing data is reduced as the number of dimensions decreases.
- Reduced dimensions lead to less computing / training time.
- Some algorithms do not work well for large size data. Thus, the diminution of these dimensions must be in order for the algorithm to be useful.
- It considers multi-collinearity by removing redundant features.
- In numerous cases, for datasets with high dimensional not all measured variables are “important” for comprehension the basic phenomena of interest.
- Helps view data. It is difficult or impossible to view data in dimensions larger than 3D, therefore, reducing the space dimension to 2D or 3D, permits representing the data and thus plot and observe more clearly the models that appear and/or even better understand their spatial representation.

In this paper we will use three methods for dimensionality reduction of ECG signals, namely, Laplacian Eigenmaps (LE) and Locality Preserving Projections (LPP), as well as a third method, the compressed acquisition (compressed sensing – CS). For testing these methods, we used ECG signal segments belonging to 8 distinct classes.

LPP is a typical graph-based dimensionality reduction method, consisting of projective maps based on solving a variational problem that optimally conserve the neighborhood structure of the data set. When the high dimensional data lies on a low dimensional manifold embedded in the ambient space, the LPP are obtained by finding the optimal linear approximations to the eigen functions of the Laplace Beltrami operator on the manifold [2,3].

The main advantage of LE is that it reduces dimensionality by keeping the local and global structure even when the data is on a manifold [4,5].

To increase precision in classifications while reducing dimensionality all ECG segments have been processed to have the R wave centered. In order to conserve classes on the manifold, the main idea is to preserve neighbors as unchanged as possible. For all cases of dimensionality reduction with the three tested algorithms, the qualitative/quantitative evaluation is the classification rate. Therefore, classification rate obtained with the original ECG patterns with the classification rates obtained on new ECG patterns on which dimensionality reductions with Laplacian Eigenmaps, Locality Preserving Projections and compressed sensed were applied will be compared.

II. Background

A. Dimensionality Reduction Techniques

In the linear case the dimensionality problem can be stated as follows: Starting from a dataset of n vectors.

$$X = [x_1 x_2 \dots x_n],\ x_i \in R^N$$

find, a projection matrix $P \in R_{N,M}$ which leads to the low dimensional dataset

$$Y = [y_1 y_2 \dots y_n],\ y_i \in R^M$$

M<N through the linear projection

$$Y = P^T X$$

Depending on the specific target represented, P can be learned through various dimension reduction techniques. In our case the above approach corresponds to LPP and CS. The LE is closely related to the above techniques except for the non-linear character, both techniques conserving the local manifold structure [1].

a) Locality Preserving Projection: LPP is a linear projective map that arises by solving a variational problem which optimally preserves the neighborhood structure of the data set.

The LPP algorithm is proposed for the purpose of deriving a linear subspace with manifold data structure. The LPP algorithm can be viewed as linear approximate version derived by Laplacian Eigenmap algorithm.

The first step in LLP is the construction of an adjacency graph G with n nodes, the number of testing signals, each node corresponding to a signal in the dataset; the ith node corresponds to the data x_i. We put an edge between any x_i and all points within a K-neighbourhood [2-3]. Based on the distance between nodes choose the weights w_{ij} with the formula.

$$w_{ij} = e^{-\|x_i - x_j\|^2/\sigma}$$

where, σ is a positive constant. The similarity matrix w_{ij} of the graph G models the local structure of the data manifold in the original space. The last step is to solve the generalized eigen-decomposition problem:

$$XLX^T\mathbf{a} = \Lambda XDX^T\mathbf{a},$$

Where:

D = Diagonal matrix

L = Laplacian matrix

Λ = a diagonal matrix with eigenvalue in ascending order

a = the corresponding eigenvector matrix.

To finish, P matrix is constructed with the first M columns in **a**. Note that the projected data to the subspace from different classes can be separated in general with nonlinear decision boundary [2-3].

b) Laplacian Eigenmap: The first two stages of the LE method are the same as those of LLP while the last stage consists of computing the eigenvectors and eigenvalues for the generalized eigenvector problem [4, 5]:

$$L\mathbf{f} = \lambda D\mathbf{f}$$

where, D = (d_{ij}) is an (n×n) diagonal matrix with,

$$d_{ii} = \sum_{j \in N_i} w_{ij}$$

L = D−W being the Laplacian matrix that can be thought as an operator on functions defined on the vertices of G.

At finish, we eliminate the eigenvector f_0 corresponding to the 0 eigenvalue and utilize the next m eigenvectors corresponding to the next m eigenvalues in increasing order for embedding in an m dimensional Euclidean space:

$$x_i \rightarrow (f_1(i), \dots, f_m(i))$$

where, $f_0, \dots, f_{k-1}$ are the solutions of generalized eigenvector problem [1,4,5].

B. Compressive Sensing

CS is a method that can be used to acquire signals with smaller number of measurements than the Nyquist rate in order to approximate sparse signals. According to CS theory a signal x can be characterized using the projections:

$$y = \emptyset x$$

where, $x \in R^N$, $y \in R^M$ is the measurement vector and $\emptyset \in R_{M,N}$ is the CS measurement or projection matrix whose entries are independent identically distributed (i.i.d) samples [6,7]. According to CS theory, the signal x can be approximately reconstructed from its projection by using an appropriate dictionary. However, in our approach we do not aim at signal reconstruction but only use the low dimensional projection vector y for classifications [6,7].

III. Experimental Results

To order to study the possibilities of dimension reduction using LE, LPP and CS method, we have considered 44 ECG signals from the MIT-BIH Arrhythmia databank acquired at a sampling frequency of 360Hz, with 11 bits / sample [8]. The databank also contains annotation files with the index of the R wave and the class to which each ECG pattern belongs.

According to the annotations files, eight major classes have been recognized, i.e., seven classes of pathological classes: atrial premature beat (A), left bundle branch block beat (L), right bundle branch block beat (R), premature ventricular contraction (V), fusion of ventricular and normal beat (F), paced beat (/), fusion of paced and normal beat (f) and a class of normal beats (N).

We used the same segmentation as in [9] to increase classification rate. So, a cardiac pattern begins in the middle of the RR interval and ends in the middle of the next RR interval. In the cardiac pattern thus obtained the R wave will be placed in the middle by resampling the waveforms on both sides of R. In this way patterns with the cardiac R wave centered have been obtained. So, all cardiac patterns are of dimension 301, with the R wave placed on the 150-th sample.

We constructed a database, namely a data set with 5608 patterns, having 701 patterns for each of the above classes. All patterns were normalized to unity norms. However, sensitivity to normalization was observed only in the case of LPP.

To classification in the initial 301 dimensional signal space we have used the KNN classifier using the Euclidean distance and the membership decision was based on the nearest neighbor.

For the original normalized ECG data the classification rate for the eight classes analyzed has been found to be 94.92% [11].

Fig. 1 shows the results for all tested methods for dimensionality reduction. It can be seen that for LE and CS the classification rate is increasing and once a maximum is reached, the classification rate stabilizes around that value. For very small values of the space dimension the best results are obtained with LE. Thus, for space dimension equal to 2 a classification rate of 82.61% is obtained, while for space dimension 8 the results are comparable for all three tested methods. If we refer to the maximum values achieved in terms of classification, then LPP offers the best results, i.e., for space dimension equal to 25, a 94.37% classification rate was obtained [10]. Several results regarding dimensionality reduction for the three tested methods are presented in Table I.

Since many techniques are noise-sensitive, we tested all three algorithms with waveforms with 8% added noise normally distributed.

Fig. 2 shows the classification rate obtained by LPP for ECG segments with and without noise. It is found that the locality-preserving character of the LPP method makes it relatively insensitive to noise because the classification rate varies significantly in the presence of 8% noise.

Fig. 3 shows the classification rate obtained by Laplacian Eigenmaps for ECG segments with and without noise. There are some small differences, but they are not significant so that LE can be considered almost insensitive to the presence of noise.

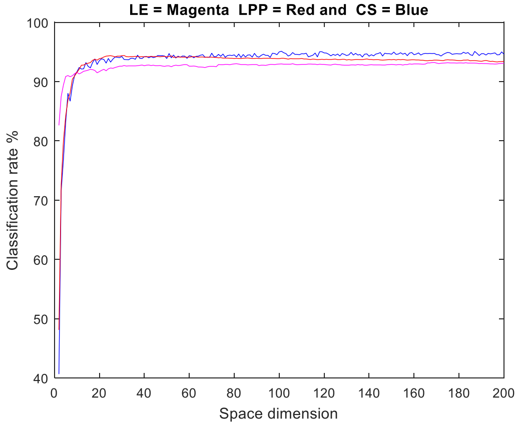


Fig. 1. Classification Rate vs. Space Dimension with LPP, LE and CS (Sigma = 5, Neighborhood k = 9) for Noiseless and Normalised ECG Patterns.

TABLE. I. CLASSIFICATION RATE % VS. SPACE DIMENSION FOR LLP, LE AND CS (SIGMA = 5, NEIGHBORHOOD K=9)

Space dimension	LE	LPP	CS
2	82,61	48,10	40,64
3	87,46	72,55	71,47
4	89,40	79,87	76,79
5	90,85	84,14	83,06
6	91,01	86,83	87,99
7	90,82	88,11	86,72
8	90,98	90,41	89,29
9	91,35	90,96	91,07
10	91,57	91,54	91,65
12	91,48	92,76	92,12
14	91,82	93,01	93,20
16	92,07	93,21	92,32
18	91,87	93,76	93,81
20	91,68	93,87	93,54
22	92,07	94,15	93,73
24	92,15	94,34	93,90
26	92,23	94,29	93,78
28	92,48	94,29	94,09
30	92,59	94,37	94,23
32	92,62	94,21	93,70
34	92,65	94,21	93,92
36	92,65	94,26	93,76
38	92,73	94,18	93,92
40	92,68	94,18	94,20
42	92,79	94,21	94,01
44	92,76	94,23	94,03
46	92,68	94,23	94,39
48	92,62	94,23	94,37
50	92,65	94,18	93,90
75	92,84	93,96	94,64
100	92,95	93,87	95,00
125	92,93	93,79	94,42
150	92,90	93,68	94,62
175	93,17	93,68	94,92
200	93,09	93,43	94,62

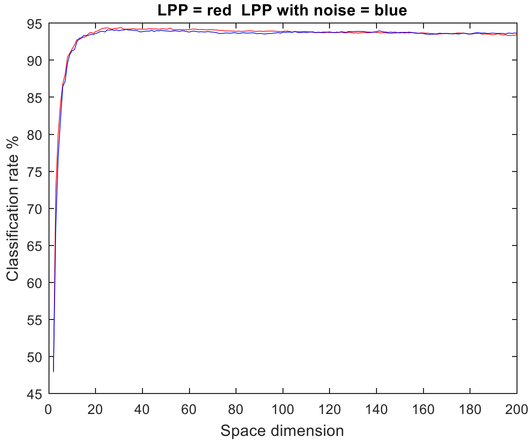


Fig. 2. Classification Rate % vs. Space Dimension with LPP (Sigma = 5, Neighborhood k = 9) Noisy and Noiseless ECG Normalised Patterns.

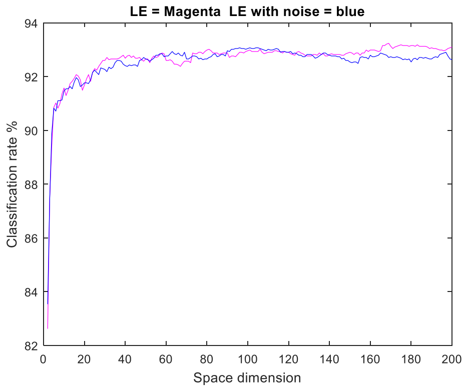


Fig. 3. Classification Rate % vs. Space Dimension with LE (Sigma = 5, Neighborhood k = 9) Noisy and Noiseless ECG Normalised Patterns.

For dimensions less than 10, it has been found that the classification rate for CS may differ, depending on the projection matrix. For this we tested several projection matrices; Table II presents these results and the average classification rates. For all three methods, the sensitivity of the algorithm to data normalization was analyzed as well.

Fig. 4 shows the classification rates obtained by compressed sensed ECG patterns with and without noise. The results are similar, so CS is not noise-sensitive as well.

Fig. 5 shows ECG patterns transformed into a 3-dimensional space for LE (87.46% classification rate), LPP (72.55% classification rate) and CS techniques (66.5% classification rate).

Fig. 6 shows the classification rate obtained by LPP for ECG segments with various levels of noise and without noise for non-normalized signals. Interestingly, for dimensions less than 40, the classification rates are practically unaffected by noise. However, for larger space dimensions, the results are worse but are improved by noise. Observe that, if we refer to the maximum values achieved in terms of classification, then LPP offers the best results for space dimension equal to 27 with a classification rate of 94%, even higher than that obtained for the initial non-normalized ECG signals (space dimension equal to 301 and classification rate 92.5%).

In Table III, several results for the LPP method applied to noiseless and noisy ECG waveforms are presented.

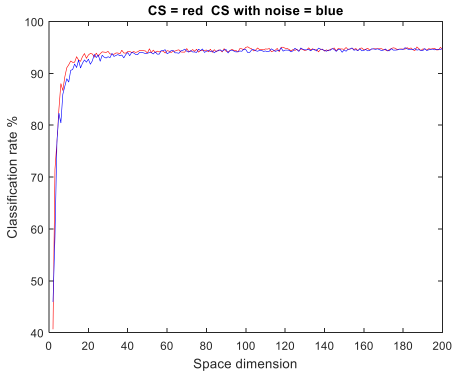


Fig. 4. Classification Rate % vs. Space Dimension with CS Noisy and Noiseless ECG Normalised Patterns.

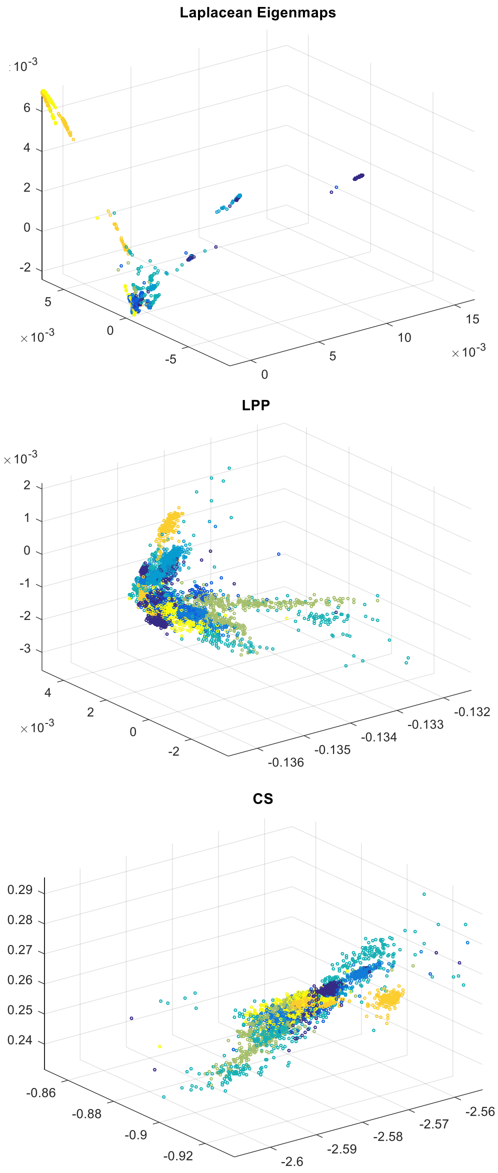


Fig. 5. ECG Patterns Reprezented into a 3-Dimensional Space with LE, LPP and CS Techniques.

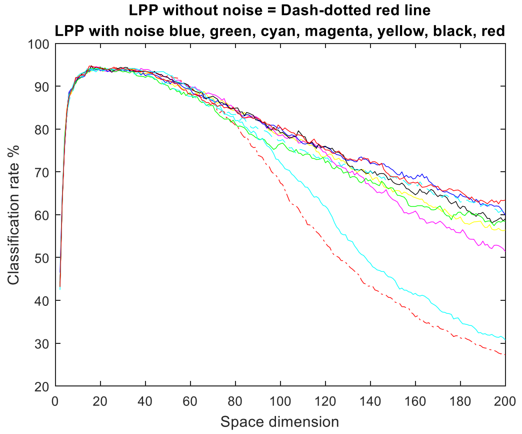


Fig. 6. Classification Rate % vs. Space Dimension with LPP (Sigma = 5, Neighborhood k = 9) for Original and Noisy Non-Normalised ECG Patterns.

TABLE. II. CLASSIFICATION RATE VS. SPACE DIMENSION FOR COMPRESSED SENSED WITH A FEW PROJECTION MATRICES

DIM	**mean**	**CS 1**	**CS 2**	**CS 3**	**CS 4**	**CS 5**	**CS 6**	**CS 7**	**CS 8**	**CS 9**	**CS 10**
2	**39,09**	44,64	33,19	36,56	37,50	40,33	40,75	38,47	36,53	42,25	40,69
3	**53,22**	53,31	54,94	50,42	54,81	49,72	48,44	61,31	48,03	64,39	46,86
4	**66,03**	66,19	63,50	66,81	68,83	62,50	68,83	63,78	74,22	62,06	63,56
5	**73,20**	74,53	80,19	74,31	71,97	70,08	69,75	74,56	68,00	72,50	76,11
7	**79,63**	83,47	81,17	74,28	78,89	77,94	83,58	80,44	77,92	76,33	82,22
9	**84,54**	84,89	84,31	84,17	84,25	85,31	84,92	85,03	84,78	82,42	85,31
11	**85,68**	85,56	82,44	85,47	84,94	86,28	87,25	87,42	87,28	85,94	84,19
13	**86,86**	85,11	88,39	89,03	85,72	85,44	87,75	85,56	86,97	88,11	86,50
15	**87,98**	87,94	87,06	87,56	88,36	88,00	87,31	88,28	87,83	88,94	88,53
17	**88,21**	89,50	88,56	89,33	87,78	87,67	87,67	88,22	87,69	88,42	87,22
20	**88,99**	89,75	89,00	90,03	88,72	88,75	89,92	88,86	88,14	88,75	87,94
25	**89,12**	89,72	87,94	89,00	87,94	89,14	89,94	89,81	90,42	88,72	88,58
30	**89,33**	89,69	88,83	89,31	88,92	89,00	88,75	89,86	90,50	89,00	89,42
35	**89,95**	90,14	89,86	89,36	89,64	90,56	90,19	90,56	89,81	89,19	90,19
40	**89,76**	90,14	89,75	89,44	89,11	90,83	89,25	89,17	90,17	89,78	89,94
45	**89,81**	89,83	89,50	89,94	90,17	89,14	90,19	89,97	89,44	89,72	90,17
50	**90,14**	90,50	89,31	89,94	90,22	89,86	90,64	89,89	90,25	90,83	90,00

TABLE. III. CLASSIFICATION RATE VS. SPACE DIMENSION FOR LPP WITH SEVERAL TYPES OF NOISE (FOR NON-NORMALISED SIGNALS)

Space Dim.	normal distribution Mean = 2 SD = 2	uniform distribution (-5, 5)	normal distribution Mean = 1 SD = 2	normal distribution Mean = 5 SD = 2	normal distribution Mean = 1 SD = 1	LPP without noise
	Red color	yellow color	black color	green color	blue color	dash-dotted red line
50	91.35	90.67	91.35	92.03	92.03	92.47
80	84.38	83.45	84.26	84.88	84.38	80.90
100	80.40	77.29	80.40	79.78	78.47	67.39
125	74.30	72.37	74.05	74.61	71.44	50.28
150	68.89	68.64	70.44	67.64	63.53	40.20
175	66.09	64.09	65.09	63.04	57.06	32.55
200	63.41	60.30	59.93	59.05	51.40	27.26

IV. CONCLUSIONS

The results presented in this paper concerned LE, LPP and CS dimension reduction techniques for ECG signals without and with noise.

Interestingly, for dimensions up to 10, the results obtained with LE are best even for very small dimensions like 2D or 3D. Even though the classification rates are smaller, it is possible to make an intuitive image on data separation. However an inconvenience of LE (unlike LPP) is that for each new data the computation should be taken from the beginning.

The results obtained with CS are close to those obtained with LE, CS offering the advantage of very low complexity in the compression stage. Another major advantage is that if a representation in which the signal is sparse is known, then from the reduced space it is possible to reconstruct (with some error) the initial data.

An advantage of LPP is that once the projection subspace has been found any new data entry will be projected on it with no other computation. However, although LPP has been successfully applied in numerous practical problems of pattern recognition the method may have some problems since the LPP results depend mainly on its underlying neighborhood graph whose construction suffers as the graph is constructed using the nearest neighbor criterion which tends to work weakly for high-dimensional original space and it is generally uneasy to assign appropriate values for the neighborhood size and heat kernel parameter implicated in the graph building.

In the case of LPP, we have found that the ECG classification results are influenced by the normalization of the signals for high space dimension-the influence of normalization is insignificant for space dimension less than 27. This can be justified by the fact that for the test signals the ECG signals are projected onto the new found subspace (the projection matrix represented by the corresponding eigenvector matrix).

Last, but not least, we found that of all three tested algorithms, LPP is the most robust to noise but sensitive to data normalization, while CS is sensitive to small dimensions of space at the projection of the matrix.

In the future, we will analyze the influence of data normalization on classification rates for dimensionality reduction methods.

References

[1] L.J.P. van der Maaten, E. O. Postma , H. J. van den Herik, Dimensionality Reduction: A Comparative Review, 2008

[2] Xiaofei He, Partha Niyogi, Locality Preserving Projections, Advances in Neural Information Processing Systems, volume 16, page 37, Cambridge, MA, USA, The MIT Press, 2004

[3] Xiaofei He, *Locality Preserving Projections*. Ph.D. Dissertation. University of Chicago, Chicago, IL, USA. Advisor(s) Partha Niyogi. AAI3195015.

[4] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15:1373–1396, 2003

[5] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, Journal of Machine Learning Research, 7, 2399–2434, 2006

[6] D. Donoho, Compressed sensing, IEEE Trans. Info. Theory, vol. 52, no. 4, pp. 1289–1306, September 2006

[7] E. J. Candes, J. Romberg, Quantitative robust uncertainty principles and optimally sparse decompositions, Foundations of Computational Mathematics, vol. 6, no. 2, pp. 227–254, April 2006.

[8] https://www.physionet.org/content/mitdb/1.0.0/ (18 september 2019)

[9] M. Fira, L. Goras, N. Cleju, C. Barabasa, On the classification of compressed sensed signals, ISSCS 2011, Iasi, 2011

[10] M. Fira, L. Goras, Dimensionality Reduction for ECG Signals; Laplacian Eigenmaps and Locality Preserving Projections, ISSCS 2019, Iasi, 2019

[11] M. Fira, L. Goras, M. Fira, L. Goras, On Biomedical Signals Dimensionality Reduction with Laplacian Eigenmaps, ISMAC - CVB India, 2019

Fraud Detection using Machine Learning in e-Commerce

Adi Saputra[1], Suharjito[2]
Computer Science Department, BINUS Graduate Program–Master of Computer Science
Bina Nusantara University Jakarta, Indonesia 11480

***Abstract*—The volume of internet users is increasingly causing transactions on e-commerce to increase as well. We observe the quantity of fraud on online transactions is increasing too. Fraud prevention in e-commerce shall be developed using machine learning, this work to analyze the suitable machine learning algorithm, the algorithm to be used is the Decision Tree, Naïve Bayes, Random Forest, and Neural Network. Data to be used is still unbalance. Synthetic Minority Over-sampling Technique (SMOTE) process is to be used to create balance data. Result of evaluation using confusion matrix achieve the highest accuracy of the neural network by 96 percent, random forest is 95 percent, Naïve Bayes is 95 percent, and Decision tree is 91 percent. Synthetic Minority Over-sampling Technique (SMOTE) is able to increase the average of F1-Score from 67.9 percent to 94.5 percent and the average of G-Mean from 73.5 percent to 84.6 percent.**

***Keywords*—Machine learning; random forest; Naïve Bayes; SMOTE; neural network; e-commerce; confusion matrix; G-Mean; F1-score; transaction; fraud**

I. Introduction

Insight of previous research results on internet users in Indonesia as released on October 2019 edition of Marketeers Magazine [1], according to the research the number of internet users in Indonesia on 2019 alone, had reached 132 million users, an increase from the previous year at 143.2 million users show in Fig. 1.

The increasing number of internet users in Indonesia has triggered market players in Indonesia to try opportunities to develop their business through internet media. One method used is to develop an E-Commerce business [3].

Based on statistical data obtained by Statista.com, it is shown that the number of retail e-Commerce (electronic commerce) sales in Indonesia will grow 133.5% to the US $ 16.5 billion or around IDR 219 trillion in 2022 from the position in 2017. This growth is supported by the rapid advances in technology that provide convenience for consumers to shop.

Huge number of transactions in e-commerce raises the potential for new problems namely fraud in e-commerce transactions shows in Fig. 2. The number of e-commerce-related frauds has also increased every year since 1993. As per a 2013 report, 5.65 cents lost due to a fraud of every $ 100 in e-commerce trading turnover. Fraud has reached more than 70 trillion dollars until 2019 [5]. Fraud detection is one way to reduce the amount of fraud that occurs in e-commerce transactions.

Fraud detection that has developed very rapidly is fraud detection on credit cards ranging from fraud detection using machine learning to fraud detection using deep learning [6] but unfortunately fraud detection for transactions on e-commerce is still small, fraud detection research on e-commerce commerce is still not much so far, fraud detection research on e-commerce is only limited to the determination of features or attributes [7] which will be used to determine the nature of fraud or non-fraud transactions in e-commerce.

The dataset used in this paper has a total of 151,112 records, the dataset classified as fraud is 14,151 records, the ratio of fraud data is 0.093 percent. Datasets that have very small ratios result in an imbalance of data. Imbalance data results in accuracy results that are more inclined to majority data than minority data. The dataset used results more in the classification of the majority of non-fraud than fraud. Accuracy results that are more inclined to majority data make the classification results worse; handling imbalance data using the SMOTE (Synthetic Minority Oversampling Technique).

Recent research about fraud detection in e-commerce transactions still determine feature extraction [8], purpose of this paper is to find the best model to detect fraud in e-commerce transactions.

In this paper research fraud transaction in ecommerce, research use dataset from Kaggle, improve classification machine learning using SMOTE, SMOTE using to handling unbalance data, after using SMOTE, dataset will be training using machine learning. Machine learning is decision tree, Naïve Bayes, random forest, and neural network machine learning to determine accuracy, precision, recall, G-mean, F1-Score.

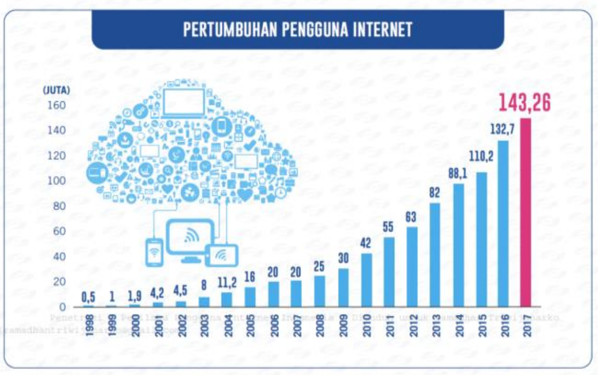


Fig. 1. Growth of Internet users [2].

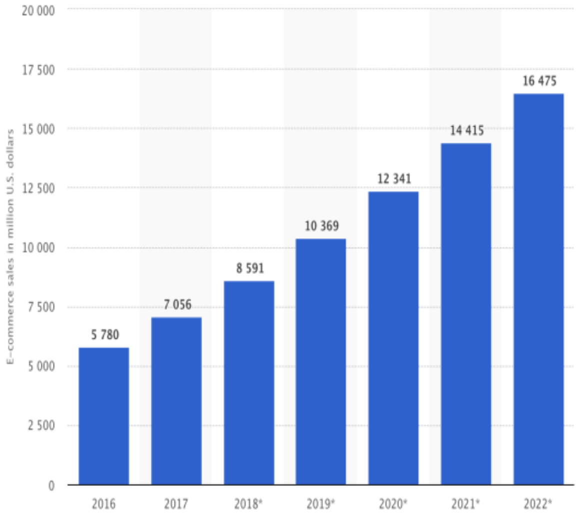


Fig. 2. Sales of e-Commerce, Statista.com [4].

II. Related Works

Fraud detection that has developed very rapidly is fraud on credit cards. Many studies discuss the fraud method. One of the studies carried out using deep learning is auto-encoder and restricted Boltzmann machine [9]. Deep learning is used to build a fraud detection model that runs like a human neural network, where data will be made in several layers that are tiered for the process, starting from the Encoder at layer 1 hinge decoder at layer 4. The researcher compares the deep learning method with other algorithms such as Hidden Markov Model (HMM) [10].

Credit card fraud detection research was also using machine learning [11] machine learning used as a decision tree algorithm, naïve Bayes, neural networks, and random forests.

Decision tree is one algorithm that is widely used in fraud detection because it is easy to use. Decision tree is a prediction model using tree structure or hierarchical structure.

Naïve Bayes is used in fraud detection credit cards because Naïve Bayes is a classification with probability and statistical methods. Naïve Bayes is very fast and quite high inaccuracy in real-world conditions neural network on fraud detection credit cards uses genetic algorithms to determine the number of hidden layer architectures on neural networks [12] with genetic algorithm, the genetic algorithm produces the most optimal number of hidden layers [13]. Fraud detection on credit cards also uses random forest [14]. Random forest uses a combination of each good tree and then combined into one model. Random Forest relies on a random vector value with the same distribution on all trees where each decision tree has a maximum depth [15].

Research on fraud detection in e-commerce is still not much so far. Fraud detection research on e-commerce is only limited to the determination of features or attributes that will be used to determine the nature of the fraud or non-fraud transactions [16]. The study describes the extraction attribute/feature process used to determine behavior in e-commerce transactions. This attribute is used as fraud detection in e-commerce. This attribute determines the transaction conditions.

Another research on fraud detection in e-commerce is a reason transaction based on the attributes or features that exist in e-commerce transactions. The features/attributes used are features of the transaction, namely invalid rating, confirmation interval, average stay time on commodities, a feature of buyer namely real name, positive rating ratio, transaction frequency.

Imbalance of data results in suboptimal classification results. The dataset on the paper has a total number of 151,112 records, the dataset classified as fraud is 14,151 records, and the ratio of fraud data is 0.093 percent. Synthetic Minority Oversampling Technique (SMOTE) is one of the methods used to make data into balance, Synthetic Minority Oversampling Technique (SMOTE) [17] is one of the oversampling methods that work by increasing the number of positive classes through random replication of data, so that the amount of data positive is the same as negative data. The way to use synthetic data is to replicate data in a small class. The SMOTE algorithm works by finding k closest neighbor for a positive class, then constructing duplicate synthetic data as much as the desired percentage between randomly and positively chosen k classes.

Recent paper about fraud detection only limited to the determination of features or attributes. Improvement fraud detection in e-commerce is used machine learning. Machine learning used is the Decision Tree, Naïve Bayes, Random Forest, and Neural Network.

III. Research Methodology

This paper aims to classify e-commerce transactions that include fraud and non-fraud using machine learning, namely Decision Tree, Naïve Bayes, Random Forest, and Neural Network. The research process is carried out as shown Fig. 3.

The classification process begins with the feature selection process in the dataset. After the feature is determined, what is done is preprocessing data using PCA, the process is carried out by transformation, normalization, and scaling of features so that the features obtained can be used for classification after the classification process is done by the SMOTE (Synthetic Minority Oversampling Technique) process. SMOTE is useful for making imbalance data into balance. The SMOTE (Synthetic Minority Oversampling Technique) process is useful for dealing with data imbalance problems in fraud cases, because fraud cases are usually below 1 percent, so as to reduce the majority class in the dataset. The majority class can make the classification more directed to the majority class so that the predictions of the classification are not as expected; the results of the SMOTE dataset transaction fraud process will be balanced [18]

Machine learning used in the classification process is decision tree, random forest, artificial neural network, and naïve Bayes. This machine-learning algorithm will be compared to find the best accuracy results from the transaction dataset in e-commerce.

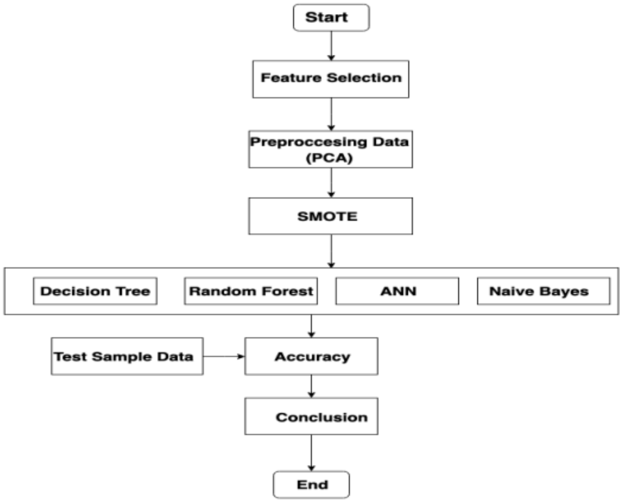


Fig. 3. Research Steps.

A. Preprocessing Data

Preprocessing is used to extract, transform, normalize and scaling new features that will be used in the machine learning algorithm process to be used. Preprocessing is used to convert raw data into quality data. In this study preprocessing uses PCA (Principle Component Analysis) with the features [19] of extraction, transformation, normalization and scaling.

PCA is a linear transformation commonly used in data compression and is a technique commonly used to extract features from data at a high-dimensional scale. PCA can reduce complex data to smaller dimensions to display unknown parts and simplify the structure of data. PCA calculations involve calculations of covariance matrices to minimize reduction and maximize variance.

B. Decision Tree

Decision trees are useful for exploring fraud data, finding hidden relationships between a number of potential input variables and a target variable. Decision tree [20] combines fraud data exploration and modeling, so it is very good as a first step in the modeling process even when used as the final model of several other techniques [21].

Decision tree is a type of supervised learning algorithm; a decision tree is good for classification algorithm. Decision tree divides the dataset into several branching segments based on decision rules, this decision rule is determined by identifying a relationship between input and output attributes.

- Root Node: This represents the entire population or sample, and this is further divided into two or more.
- Splitting: This is the process of dividing a node into two or more sub-nodes.
- Decision Node: When a sub-node is divided into several sub nodes.
- Leaf / Terminal Node: Unspecified nodes are called Leaf or Terminal nodes.
- Pruning: When a sub-node is removed from a decision.
- Branch / Sub-Tree: Subdivisions of all trees are called branches or sub-trees.
- Parent and Child Node: A node, which is divided into sub-nodes [22].

The fraud detection architecture using a decision tree consists of the root node, internal node and leaf node of the decision tree architecture as shown Fig. 4.

C. Naïve Bayes

Naïve Bayes predicts future opportunities based on past experience [23], it uses the calculation formula as below.

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \quad (1)$$

Where:

B: Data with unknown classes

A: The data hypothesis is a specific class

P(A|B): Hypothesis probability based on conditions (posterior probability)

P (A): Hypothesis probability (prior probability)

P(B|A): Probability-based on conditions on the hypothesis

P (B): Probability A

By using the formula above can be obtained opportunities from fraud transactions and non-fraud transactions

D. Random Forest

Random forest (RF) is an algorithm used in the classification of large amounts of data. Random Forest (RF) is a development of the Classification and Regression Tree (CART) method by applying the bootstrap aggregating (bagging) method and random feature selection Architecture Random forest as shown in Fig. 5.

Random forest is a combination of each good transaction fraud tree which is then combined into one model. Random Forest relies on a random vector value with the same distribution on all trees, each decision tree in e-commerce fraud detection which has a maximum depth. The class produced from the classification process is chosen from the most classes produced by the decision tree.

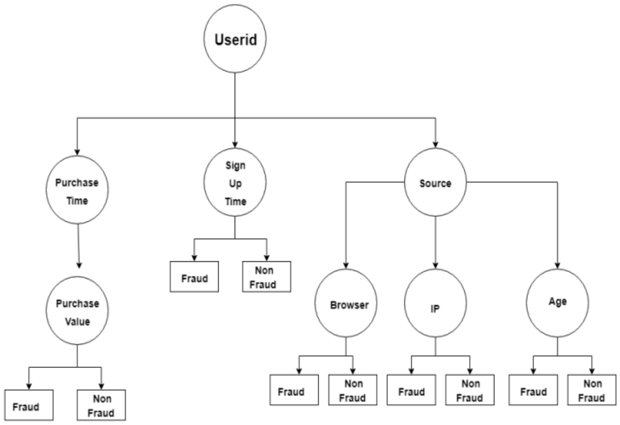


Fig. 4. Architecture of Decision Trees.

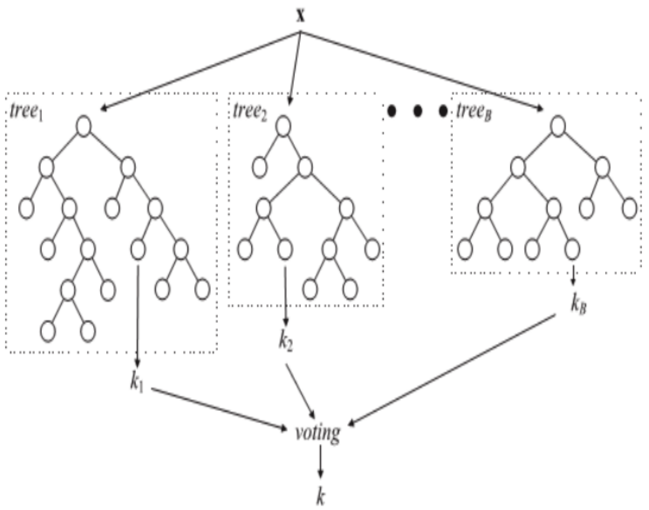


Fig. 5. Architecture of Random Forest.

E. Neural Network

The algorithm neural network is an artificial intelligence method whose concept is to apply a neural network system in the human body where nodes are connected to each other, architecture neural network as shown in Fig. 6.

The number of input layers before training is 11 input layers, after preprocessing the input layer to 17 input layers, in addition to determining the hidden layer, genetic algorithms on the neural network is used [24]. The GA-NN [25] algorithm process for this forecasting process is as follows:

- This forecasting is as follows:
- Initialization count = 0, fitness = 0, number of cycles
- Early population generation. Individual chromosomes are formulated as successive gene sequences, each encoding the input.
- Suitable network design
- Assign weights
- Conduct training with backpropagation Looks for cumulative errors and fitness values. Then evaluated based on the value of fitness.
- If the previous fitness <current fitness value, save the current value
- Count = count +1
- Selection: Two mains are selected using a wheel roulette mechanism
- Genetic Operations: crossover, mutation, and reproduction to produce new feature sets
- If (number of cycles <= count) return to number four
- Network training with selected features
- Study performance with test data.

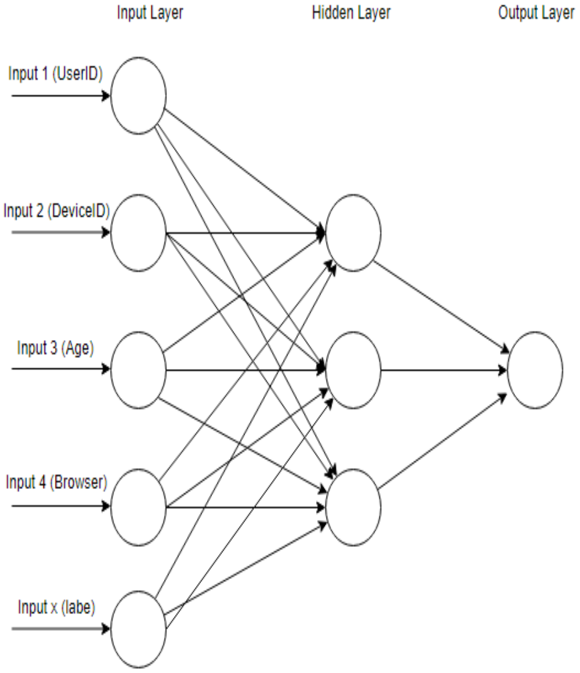


Fig. 6. Architecture of Neural Network.

F. Confusion Matrix

Confusion matrix is a method that can be used to evaluate classification performance. Table I shows a dataset with only two types of classes [26].

True Positive (TP) and True Negative (TN) are the number of positive and negative classes that are classified correctly, False Positive (FP) and False Negative (FN) is the number of positive and negative classes that are not classified correctly. Based on the confusion matrix, performance criteria such as Accuracy, Precision, Recall, F-Measure, G-Mean can be determined.

Accuracy is the most common criteria for measuring classification performance, but if working in an imbalanced class, this criterion is not appropriate because the minority class will have a small contribution to the accuracy criteria. The recommended evaluation criteria are recall, precision F-1 Score and G-Mean. F-1 Score is used to measure the classification of minority classes in unbalanced classes, and the G-mean index is used to measure overall performance (overall classification performance).

In this study, classification performance using Recall, Precision, F-1 Score and G-Mean:

$$\text{Accuracy} \quad = \frac{\text{TP+ TN}}{\text{TP+T N+FN+FP}} \tag{2}$$

$$\text{Recall} \quad = \frac{\text{TP}}{\text{TN+FP}} \tag{3}$$

$$\text{Precision} \quad = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

$$\text{G} - \text{Mean} \quad = \sqrt{\text{TP} - \text{TN}} \tag{5}$$

$$\text{F1} - \text{Score} \quad = \frac{\text{2 x Precision x Recall}}{\text{Precision + Recall}} \tag{6}$$

TABLE. I. CONFUSION MATRIX

Class	Predictive Positive	Predictive Negative
Actual Positive	TP	TN
Actual Negative	FP	FN

IV. RESULTS AND DISCUSSION

A. Dataset

This study uses an e-commerce fraud dataset sourced from Kaggle. The dataset consists of 151,112 records, a dataset classified as fraud is 14,151 records, and the ratio of fraud data is 0.093. SMOTE (Synthetic Minority Oversampling Technique) [27] minimizes class imbalance in the fraud transaction dataset by generating synthesis data, so that the total data consists of 151,112 records, dataset classified as fraud is 14,151 records, fraud data ratio is 0.093, as shown in Fig. 7.

After oversampling at the picture Fig. 8

The SMOTE (Synthetic Minority Oversampling Technique) process makes the synthesis data so that the data becomes balance.

B. Decision Trees

The experimental process using the decision tree model is done by preparing data that has been done by the preprocessing process. After preprocessing, the data will be carried out by oversampling the classification using the decision tree will be done using the oversampling data, and also the decision tree will be done by using the data that has not been oversampled. The results of these two experiments will show the results of the classification using a comparison of decision trees and the SMOTE (Synthetic Minority Oversampling Technique) oversampling process.

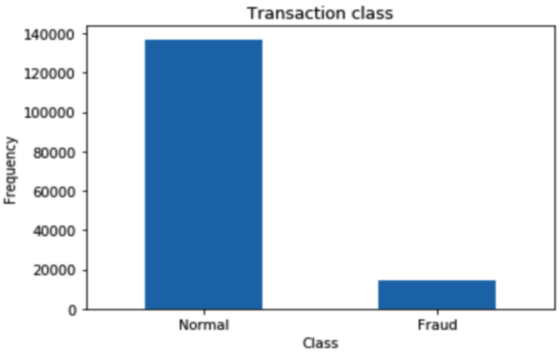


Fig. 7. Ratio Fraud.

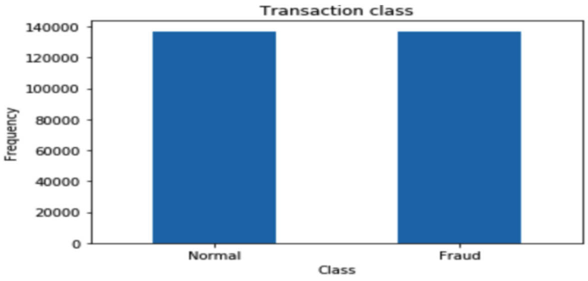


Fig. 8. Ratio Fraud after over Sampling.

Decision tree without SMOTE produce Accuracy is 91%, recall is 59.8%, Precision is 54.1%, F1-Score is 56.8%, G-Mean is 75.2%. Table II shows result from confusion matrix decision tree without SMOTE.

Decision tree with SMOTE produce Accuracy is 91%, recallis 60.4%, Precisionis 91.6%, F1-Score is 91.2%, G-Mean is 75.3%. Table III shows result from confusion matrix decision tree with SMOTE.

C. Naïve Bayes

The process of testing using the Naïve Bayes model is done by preparing data that has already been done in the preprocessing process. After preprocessing, the data will be carried out oversampling using Naïve Bayes classification will be done using data that has been oversampling, and also Naïve Bayes will be done using data that is not oversampling. The results of these two experiments will show the results of the classification using the comparison of Naïve Bayes and the SMOTE (Synthetic Minority Oversampling Technique) oversampling process.

Naïve Bayes without SMOTE produce Accuracy is 95%, recall is 54.1%, Precision is 91.1%, F1-Score is 67.9%, G-Mean is 73.3%. Table IV shows result from confusion matrix naïve Bayes without SMOTE.

Naïve Bayes with SMOTE produce Accuracy is 95%, recall is 54.2%, Precision is 94.9%, F1-Score is 94.5%, G-Mean is 73.4%. Table V shows result from confusion matrix Naïve Bayes with SMOTE.

TABLE. II. CONFUSION MATRIX DECISION TREE WITHOUT SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	38782	38782
Actual Negative	1746	2595

TABLE. III. CONFUSION MATRIX DECISION TREE WITH SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	38651	2342
Actual Negative	1724	2617

TABLE. IV. CONFUSION MATRIX NAÏVE BAYES WITHOUT SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	40764	229
Actual Negative	1993	2348

TABLE. V. CONFUSION MATRIX NAÏVE BAYES WITH SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	40760	233
Actual Negative	1988	2353

D. Random Forest

The trial process using the Random Forest model is carried out by preparing data that has already been done by the preprocessing process. After preprocessing, the data will be carried out classification oversampling using Random Forest will be done using data that has been oversampled, and also Random Forest will be done using data that is not oversampling. The results of these two experiments will show the classification results using the Random Forest comparison and the SMOTE (Synthetic Minority Oversampling Technique) oversampling process.

Random forest without SMOTE produce Accuracy is 95%, recall is 55%, Precision is 95.5%, F1-Score is 69.8%, G-Mean is 74.0%. Table VI shows result from confusion matrix random forest without SMOTE.

Random Forest with SMOTE produce Accuracy is 95%, recall is 58.1%, Precision is 80.5%, F1-Score is 94.3%, G-Mean is 75.7%. Table VII shows result from confusion matrix random forest with SMOTE.

E. Neural Network

Research using the Neural Network model is done by preparing data that has already been done by the preprocessing process. After preprocessing, the data will be carried out classification oversampling using Neural Network will be done using data that has been oversampling, and also Random Forest will be done using data that is not oversampling. The results of these two experiments will show the results of classification using the Neural Network comparison and the SMOTE (Synthetic Minority Oversampling Technique) oversampling process.

Neural network without SMOTE produce Accuracy is 96%, recall is 54%, Precision is 97.1%, F1-Score is 97.1%, G-Mean is 73.5%. Table VIII shows result from confusion matrix neural network without SMOTE.

Neural network with SMOTE produce Accuracy is 85%, recall is 76.7%, Precision is 92.5%, F1-Score is 85.1%, G-Mean is 84.6%. Table IX shows result from confusion matrix neural network with SMOTE.

TABLE. VI. CONFUSION MATRIX RANDOM FOREST WITHOUT SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	40881	112
Actual Negative	1954	2387

TABLE. VII. CONFUSION MATRIX RANDOM FOREST WITH SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	40383	610
Actual Negative	1820	2521

TABLE. VIII. CONFUSION MATRIX NEURAL NETWORK WITHOUT SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	41113	24
Actual Negative	1932	2265

TABLE. IX. CONFUSION MATRIX NEURAL NETWORK WITH SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	38566	2539
Actual Negative	9585	31487

Experiments using several algorithms produce accuracy values as shown in Fig. 9. The highest accuracy value in the neural network algorithm is 96%.

Experiments using several algorithms produce recall values as shown in Fig. 10, recall values increase using machine learning algorithms and also the Synthetic Minority Over Sampling Technique (SMOTE) compared only using the decision tree algorithm, random forest, Naïve Bayes, and neural networks only, the highest increase occurred in the neural network algorithm and the SMOTE (Synthetic Minority Over Sampling Technique).

Experiments using several algorithms produce precision values as shown in Fig. 11, the value decreases using machine learning algorithm and the Synthetic Minority Over Sampling Technique (SMOTE) compared only using the decision tree algorithm, random forest, Naïve Bayes, and neural networks, highest occurs in neural network algorithms and SMOTE (Synthetic Minority Over Sampling Technique).

Experiments using several algorithms produce F1-Score values as shown in Fig. 12, F1-Score values are increased by using machine learning algorithms and also Synthetic Minority Over Sampling Technique (SMOTE) compared only using algorithms. F1-Score is used to measure the classification of minority classes in unbalanced classes.

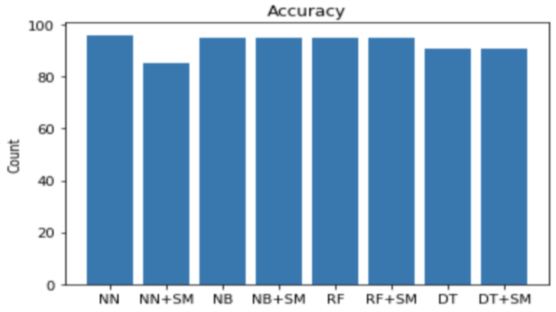


Fig. 9. Accuracy Result.

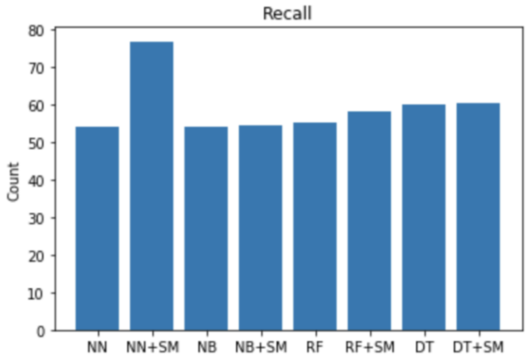


Fig. 10. Recall Result.

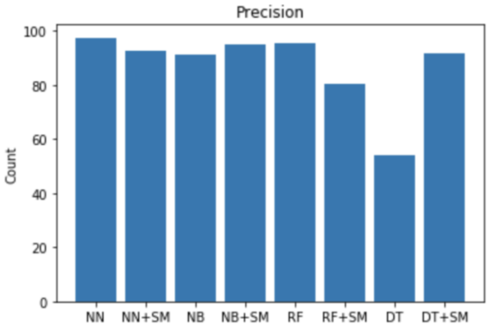


Fig. 11. Precision Result.

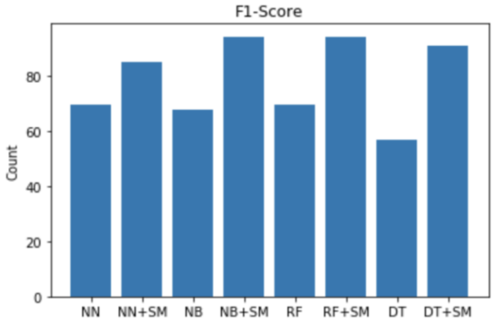


Fig. 12. F1-Score Result.

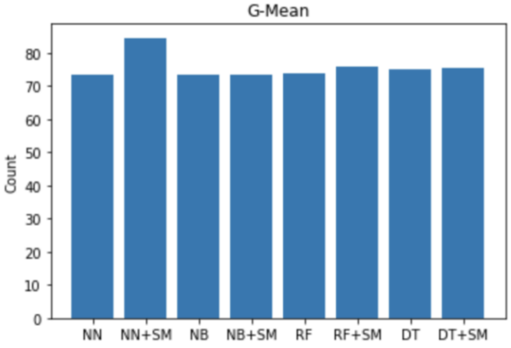


Fig. 13. G-Mean Result.

The G-mean value increased by using machine learning algorithm values as shown in Fig. 13, Synthetic Minority Over Sampling Technique (SMOTE) compared only using the G-mean algorithm used to measure overall performance (overall classification performance).

V. Conclusion

The e-commerce transaction fraud dataset is a database that has a class imbalance. This study applies the Synthetic Minority Over Sampling Technique (SMOTE) method to deal with class imbalances in the e-commerce transaction fraud dataset, the algorithm used is the decision tree, Naïve Bayes. random forest and neural network.

The results showed that the highest accuracy was 96% neural network, then random forest, and Naïve Bayes were 95%, for decision trees accuracy was 91%. Neural network has best accuracy because GA (genetic algorithms). Genetic algorithms can be used for improving ANN performance. Genetic algorithm can determine the number of hidden nodes and hidden layers, select relevant features, neural network. The SMOTE method in the experiment showed an increase in the value of recall, f1-score and also G -Mean, Neural network recall increased from 54% to 76.7%, Naïve Bayes recall increased from 41.2% to 41.3%, recall random forest from 55% to 58%, and recall decision tree from 59.8% to 60.4%. The value of f1-score also increased for all machine learning methods for neural networks increased from 69.8% to 85.1%, f1-score Naïve Bayes increased from 67.9% to 94.5%, f1-score random forest 69.8% to 94.3%, the f1-score for the decision tree also increased from 56.8% to 91.2%. By using SMOTE the value of G-Mean also increased for neural networks increased from 73.5% to 84.6%, G-Mean Naïve Bayes increased from 73.3% to 73.4%, G-Mean random forest 74% to 75 7%, the G-Mean for decision tree also increased from 75.2% to 75.3%.

Based on the results of the above experiment, it was concluded that the application of SMOTE on neural networks, random forests, decision trees, and Naïve Bayes was able to handle the imbalance of the e-commerce fraud dataset by producing higher G-Mean and F-1 scores compared to neural networks, random forest, decision tree, and Naïve Bayes. This proves that the SMOTE method is effective in increasing the performance of unbalanced data classification.

VI. Future Work

In Future studies, it is expected to be able to use other algorithms or deep learning for fraud detection in e-commerce and other future study to improve neural network accuracy when using the SMOTE (Synthetic Minority Over Sampling Technique) process.

References

[1] Asosiasi Penyelenggara Jasa Internet Indonesia, " Magazine APJI(Asosiasi Penyelenggara Jasa Internet Indonesia)" (2019): 23 April 2018.

[2] Asosiasi Penyelenggara Jasa Internet Indonesia, "Mengawali integritas era digital 2019 - Magazine APJI(Asosiasi Penyelenggara Jasa Internet Indonesia)" (2019).

[3] Laudon, Kenneth C., and Carol Guercio Traver. E-commerce: business, technology, society. 2016.

[4] statista.com. retail e-commerce revenue forecast from 2017 to 2023 (in billion U.S. dollars). (2018). Retrieved April 2018, from Indonesia: : https://www.statista.com/statistics/280925/e-commerce-revenue-forecast-in-indonesia/.

[5] Renjith, S. Detection of Fraudulent Sellers in Online Marketplaces using Support Vector Machine Approach. International Journal of Engineering Trends and Technology (2018).

[6] Roy, Abhimanyu, et al. "Deep learning detecting fraud in credit card transactions." 2018 Systems and Information Engineering Design Symposium (SIEDS). IEEE, 2018.

[7] Zhao, Jie, et al. "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce." Decision support systems 86 (2016): 109-121.

[8] Zhao, Jie, et al. "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce." Decision support systems 86 (2016): 109-121.

[9] Pumsirirat, Apapan, and Liu Yan. "Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine." International Journal of advanced computer science and applications 9.1 (2018): 18-25.

[10] Srivastava, Abhinav, et al. "Credit card fraud detection using hidden Markov model." IEEE Transactions on dependable and secure computing 5.1 (2008): 37-48.

[11] Lakshmi, S. V. S. S., and S. D. Kavilla. "Machine Learning For Credit Card Fraud Detection System." International Journal of Applied Engineering Research 13.24 (2018): 16819-16824.

[12] Aljarah, Ibrahim, Hossam Faris, and Seyedali Mirjalili. "Optimizing connection weights in neural networks using the whale optimization algorithm." Soft Computing 22.1 (2018): 1-15.

[13] Bouktif, Salah, et al. "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches." Energies 11.7 (2018): 1636.

[14] Xuan, Shiyang, Guanjun Liu, and Zhenchuan Li. "Refined weighted random forest and its application to credit card fraud detection." International Conference on Computational Social Networks. Springer, Cham, 2018.

[15] Hong, Haoyuan, et al. "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)." Catena 163 (2018): 399-413.

[16] Zhao, Jie, et al. "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce." Decision support systems 86 (2016): 109-121.

[17] Sharma, Shiven, et al. "Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance." 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018.

[18] Kim, Jaekwon, Youngshin Han, and Jongsik Lee. "Data imbalance problem solving for smote based oversampling: Study on fault detection prediction model in semiconductor manufacturing process." Advanced Science and Technology Letters 133 (2016): 79-84.

[19] Sadaghiyanfam, Safa, and Mehmet Kuntalp. "Comparing the Performances of PCA (Principle Component Analysis) and LDA (Linear Discriminant Analysis) Transformations on PAF (Paroxysmal Atrial Fibrillation) Patient Detection." Proceedings of the 2018 3rd International Conference on Biomedical Imaging, Signal Processing. ACM, 2018.

[20] Harrison, Paula A., et al. "Selecting methods for ecosystem service assessment: A decision tree approach." Ecosystem services 29 (2018): 481-498.

[21] Randhawa, Kuldeep, et al. "Credit card fraud detection using AdaBoost and majority voting." IEEE access 6 (2018): 14277-14284.

[22] Lakshmi, S. V. S. S., and S. D. Kavilla. "Machine Learning For Credit Card Fraud Detection System." International Journal of Applied Engineering Research 13.24 (2018): 16819-16824.

[23] Li, Tong, et al. "Differentially private Naïve Bayes learning over multiple data sources." Information Sciences 444 (2018): 89-104.

[24] Suganuma, Masanori, Shinichi Shirakawa, and Tomoharu Nagao. "A genetic programming approach to designing convolutional neural network architectures." Proceedings of the Genetic and Evolutionary Computation Conference. ACM, 2017.

[25] Ruehle, Fabian. "Evolving neural networks with genetic algorithms to study the string landscape." Journal of High Energy Physics 2017.8 (2017): 38.

[26] Ting, Kai Ming. "Confusion matrix." Encyclopedia of Machine Learning and Data Mining (2017): 260-260.

[27] Siringoringo, Rimbun. "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor." Journal Information System Development (ISD) 3.1 (2018).

Enhancing Visualization of Multidimensional Data by Ordering Parallel Coordinates Axes

Ayman Nabil[1]
Faculty of Compuuter Science
Misr International University
Cairo,Egypt

Karim M. Mohamed[2], Yasser M. Kamal[3]
College of Computing and Information Technology
AASTMT, Cairo Branch
Cairo,Egypt

***Abstract*—Every year business is overwhelmed by the quantity and variety of data. Visualization of Multi-dimensional data is counter-intuitive using conventional graphs. Parallel coordinates are proposed as an alternative to explore multivariate data more effectively. However, it is difficult to extract relevant information through the parallel coordinates when the data are Multi-dimensional with thousands of lines overlapping. The order of the axes determines the perception of information on parallel coordinates. This paper proposes three new techniques in order to arrange the axes in the most significant relation between the datasets. The datasets used in this paper, for Egyptian patients, with many external factors and medical tests. These factors were collected by a questionnaire sheet, made by medical researchers. The first Technique calculates the correlation between all features and the age of the patient when they get diabetes disease. The second technique is based on merging different features together and arranging the coordinates based on the correlations values. The Third Technique calculates the entropy value for each feature and then arrange the parallel coordinates in descending order based on the positive or negative values. Finally based on the result graphs, we conclude that the second method was more readable and valuable than the other two methods.**

Keywords*—Parallel coordinates; visualization; correlation coefficient; entropy function

I. Introduction

In the recent studies of computer science and technologies, an accelerating information explosion is being witnessed. In digital universe today about 2.7 Zeta bytes executed continuously [1]. Based on the Estimations and studies presented by the International Data Corporation (IDC), they suspect that by 2020 business transactions on the internet-business-to-business and business-to-consumer will reach 450 billion per day [2]. Moreover, analysis and knowledge are power and in order to analysis and interpret these huge amounts of data, Users have to use tools to visualize this data. These visualization tools can assist in retrieving valuable information, which may effectively help in solving many different types of problems. One of these important tools is the Parallel coordinates, method of visualizing high dimensional geometry and analyzing multidimensional data [3].

These days the data and its dimensions' increase rapidly which results too much interference in the coordinates and timelines of the parallel coordinates, lead to obstacles in analyzing. For this reason, many papers are presented to solve these difficulties and complexities to interpret this data [4] [5] [6] [7]. This interference could lead to a complexity in reading or interpreting the data.

Previous research has proposed exploratory techniques to enhance the visualization of multidimensional data. Within the last 20 years researches focused on Techniques to reduce the number of poly-lines or reducing or reordering the parallel axes [8] [9]. This paper introduces novel techniques for reordering the factors of the data based on the correlation coefficient calculations. The goal of these techniques is to facilitate the readiness and the complexity of the parallel coordinates. The paper categorized into different sections, the proposed methods, a detail explanation about the new techniques proposed. The results and discussion the comparison between the three techniques and finally the conclusion section.

II. Background and Related Work

The Parallel coordinates is an interactive visualization, and is the most used for multidimensional data visualization. It was developed and popularized by Alfred Inselberg [10]. Improving the parallel coordinates plot is a highly active research topic. There are some techniques proposed in previous research that attempted to enhance the readability of the results by applying clustering techniques or sampling polynies [11] [12] [13] [14]. Moreover the readiness and effectiveness of the parallel coordinates depends on ordering the dimensions and factors, different dimension ordering techniques were presented [15] [16] [17].

Other papers proposed new methods for interpreting the readiness of the parallel coordinate by dividing the dimensions of the datasets input into groups of lower dimensions based on the correlations calculations; the conclusion of this technique can represent various groups of correlated dimensions in high dimensional data space [8].

Furthermore, another paper proposed the automated assistance to rearrange the order of the variable; this automation was done using a system called V-miner. Motorola engineers were affected by the new powerful enhancements and also facilitate the use of the parallel coordinates [4].

Also techniques were proposed to simplify the representation of the parallel coordinate visualization, where a new study proposed using the eye tracking. The main idea is to understand whether the parallel coordinate visualizations are easy to be perceived or not.

From the results of this study, the users were able to interpret and realize the parallel coordinate easily by concentration on the correct areas for the chart [18].

III. The Proposed Methods

This section will discuss the proposed methods to enhance the visualization in the parallel coordinates. The goal of using the Parallel coordinates is one of the most important techniques to visualize dataset with multidimensional datasets, the better visualization becomes obvious, and more information can be retrieved [19]. The results of the parallel coordinate visualization always confuse the reader, and could lead to difficulties to read. Past studies proved that the correlation coefficient affects the result and the interpretation of the parallel coordinate visualization [20]. The effectiveness on the interpretation and the readiness, also has an effect on the visualization between two coordinates, for instance, the parallel coordinates plot for data that have negative 1 correlation different from the parallel coordinates for data that have 1 correlation is as follows:

In Fig. 1 and 2, the correlation affects the visualization of the parallel coordinate chart, and incase the two features are correlated or not the lines interfere or move in parallel path. For this reason, this paper proposed two of the new methods based on the correlation coefficient. In order to simplify the complexity of the intervention between lines that may lead to difficulties in tracking the parallel coordinate's graphs.

Moreover, these techniques give the user a better chance to interpret and analyze the datasets more professionally. The used datasets are for Egyptian people suffering from the diabetes disease. This data was collected by the Egyptian National Research Center and was based on standard medical questionnaire. This questionnaire was prepared by specialized doctors in the diabetic field.

The goal of implementing these two methods on the diabetic patients' dataset is to reach the most significant features that affect the health of these patients and assist in triggering the diabetic disease faster in younger ages.

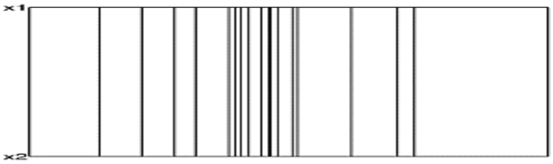


Fig. 1. Parallel Coordinates Plot for Data with Correlation Coefficient of 1.

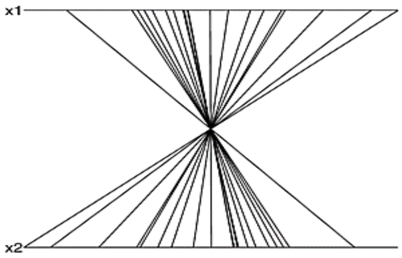


Fig. 2. Parallel Coordinates Plot for Data with Negative Correlation Coefficient.

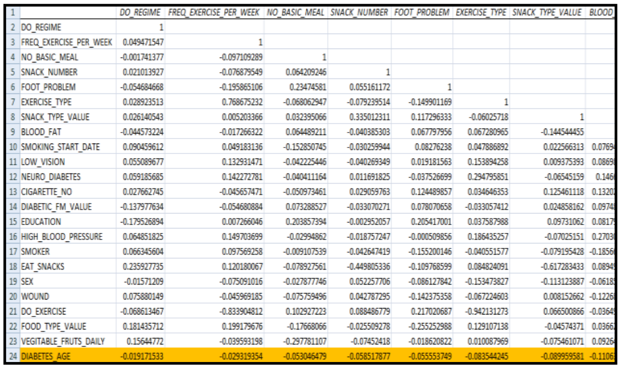
A. Datasets

The Egyptian National Research Center compiled these Datasets based on a medical questionnaire which contains 348 patients. This questionnaire is comprised of questions regarding the risk factors that cause diabetes disease and were questions for diabetes patients. After that these forms were extracted into a statistical tool called SPSS, for doing statistical analysis on this data and finally they were exported into an excel sheet, in order to be used in experiments as shown in Fig. 3.

The datasets were collected 6-years ago. The Dataset comprised of 23 features; these features are summarized in Table I.

TABLE. I. Description Of Dataset Feature

No.	Feature name	Type	Range
1	Diabetes Age	Numeric	Real Values
2	Gender	Numeric	Categorical
3	Education	Numeric	Categorical
4	Diabetic Family member	Numeric	Categorical
5	Smoker	Numeric	Categorical
6	Cigarette number	Numeric	Real Values
7	Smoking Start Date	Date	Real Values
8	Exercising Status	Numeric	Categorical
9	Frequent Exercise per week	Numeric	Real Values
10	Exercise Type	Numeric	Categorical
11	Food Type	Numeric	Categorical
12	Healthy Food status	Numeric	Categorical
13	No of Basic Meals	Numeric	Real Values
14	Snacks Status	Numeric	Categorical
15	Snacks Number	Numeric	Real Values
16	Snack Type	Numeric	Categorical
17	Regime Status	Numeric	Categorical
18	Blood Pressure Status	Numeric	Categorical
19	Blood Fat Status	Numeric	Categorical
20	Foot Complications	Numeric	Categorical
21	Neuro Complications	Numeric	Categorical
22	Low Vision status	Numeric	Categorical
23	Wound Recovery Status	Numeric	Categorical



1		DO_REGIME	FREQ_EXERCISE_PER_WEEK	NO_BASIC_MEAL	SNACK_NUMBER	FOOT_PROBLEM	EXERCISE_TYPE	SNACK_TYPE_VALUE	BLOOD_
2	DO_REGIME	1							
3	FREQ_EXERCISE_PER_WEEK	0.049471547	1						
4	NO_BASIC_MEAL	-0.001741377	-0.097109289	1					
5	SNACK_NUMBER	0.021013927	-0.076879549	0.064209246	1				
6	FOOT_PROBLEM	-0.054684668	-0.195865106	0.23474581	0.055161172	1			
7	EXERCISE_TYPE	0.028923513	0.768675232	-0.068062947	-0.079239514	-0.149901169	1		
8	SNACK_TYPE_VALUE	0.026140543	0.005203366	0.032395066	0.335012311	0.117296333	-0.06025718	1	
9	BLOOD_FAT	-0.044573224	-0.017266322	0.064489211	-0.040385303	0.067797956	0.067280965	-0.144544455	
10	SMOKING_START_DATE	0.090459612	0.049183136	-0.152850745	-0.030259944	0.08276238	0.047886892	0.022566313	0.0769
11	LOW_VISION	0.055089677	0.132931471	-0.042225446	-0.040269349	0.019181563	0.153894258	0.009375393	0.0869
12	NEURO_DIABETES	0.059185685	0.142272781	-0.040411164	0.011691825	-0.037526699	0.294795851	-0.06545159	0.146
13	CIGARETTE_NO	0.027662745	-0.045657471	-0.050973461	0.029059763	0.124489857	0.034646353	0.125461118	0.1320
14	DIABETIC_FM_VALUE	-0.137977634	-0.054680884	0.073288527	-0.033070271	0.078070658	-0.033057412	0.024858162	0.0974
15	EDUCATION	-0.179526894	0.007266046	0.203857394	-0.002952057	0.205417001	0.037587988	0.09731062	0.0817
16	HIGH_BLOOD_PRESSURE	0.064851825	0.149703699	-0.02994862	-0.018757247	-0.000509856	0.186435257	-0.07025151	0.2703
17	SMOKER	0.066345604	0.097569258	-0.009107539	-0.042647419	-0.155200146	-0.040551577	-0.079195428	-0.1856
18	EAT_SNACKS	0.235927735	0.120180067	-0.078927561	-0.449805336	-0.109768599	0.084824091	-0.617283433	0.0894
19	SEX	-0.01571209	-0.075091016	-0.027877746	0.052257706	-0.086127842	-0.153473827	-0.113123887	-0.0618
20	WOUND	0.075880149	-0.045969185	-0.075759496	0.042787295	-0.142375358	-0.067224603	0.008152662	-0.1226
21	DO_EXERCISE	-0.068613467	-0.833904812	0.102927223	0.088486779	0.217020687	-0.942131273	0.066500866	-0.0364
22	FOOD_TYPE_VALUE	0.181435712	0.199179676	-0.17668066	-0.025509278	-0.255252988	0.129107138	-0.04574371	0.0366
23	VEGITABLE_FRUTS_DAILY	0.15644772	-0.039593198	-0.297781107	-0.07452418	-0.018620822	0.010087969	-0.075461071	0.0926
24	DIABETES_AGE	-0.019171533	-0.029319354	-0.053046479	-0.058517877	-0.055553749	-0.083544245	-0.089959581	-0.1106

Fig. 3. The Correlation Coefficient for Each Variable with Respect to the Age.

B. First Method

In the first method the datasets are categorized into independent variable and dependent variables, the dependent variable in this case is the age of the patients when they got the diabetes disease. Then calculate the correlation between all the features with the dependent variable (age). These features will be organized based on the correlation values ascending on both sides of the age variable. The positive correlation features arranged on the right hand side and the negative correlation values on the left hand side. Then a parallel coordinate chart is drawn using the TIBCO spotfire software.

The correlation was calculated based on the Pearson's correlation function; the used function is:

$$\hat{r}(x, y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})}^2} \qquad (1)$$

This function will measure the strength of the linear association between the two variables. n is the number of pairs data. The X and the Y variables represent the independent and the dependent variables. r is such that -1 < r < +1. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

1) Positive correlation: If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increase, values for y also increase.

2) Negative correlation: If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.

3) No correlation: If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables.

In Fig. 5, it illustrates the arrangement of the positive or negative values regarding to the dependent variable.

C. Second Method

In this section a new method is proposed for rearranging the coordinates. On the first method calculated the correlations of all the features with the output value only. But the values of the most two significant correlation values are merged with the age and then calculate the correlation of these merged factors with the rest of the features to get the most two significant values to the new merged value as shown in Fig. 4.

For example, the result of the first calculation for correlation with the age factors were the high blood pressure and the smoking variable, subsequently multiply the values of these three factors, and recalculate the correlation again. Other positive and negative correlations will be resulted; hence multiply the five factors together and recalculate the correlations and so on until getting a final arrangement based on this methodology, then draw the parallel coordinate chart based on these arrangements.

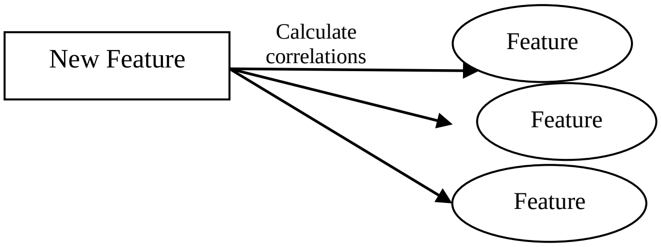


Fig. 4. Merge Features.

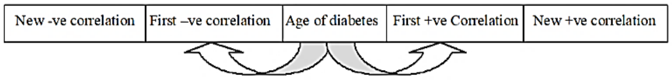


Fig. 5. Ordering Features.

D. Third Method

Third method used the entropy function, which characterizes the impurity of an arbitrary collection. The entropy always uses the information theory and is used in the decision tree algorithm to calculate the homogeneity of the datasets [21]. In this method the entropy value is being calculated for all the independent and dependent variables, then rearranging them in a descending order.

Furthermore, these features will be arranged based on the sign, where positive values on the right and the negative values on the left. Finally, plot the parallel coordinates chart with the result in ordering. The following query is used to calculate the entropy:

$$E(s) = \sum_{i=1}^{c} -p_i \ \log_2 p_i \qquad (2)$$

Pi: the probability of class i. Compute proportion of i in the set. The higher E(s) the more information gain.

IV. Results and Discussion

The first experiment is comparing the three methods in general without using the brushing tool. The differences between the three figures are obvious. Fig. 9 is readable and easily to interpret comparing to Fig. 8 and Fig. 10, for instance in Fig. 9 there are many negative correlations easily to be tracked or analysis other than the other two figures. These negative correlations became visible after applying the 2nd new method on the parallel coordinate chart.

Moreover, some features aren't correlated in Fig. 8 and 10, where the lines crossing all around forming disorganization and complexity, for example the Neuro and cigarettes numbers factors.

Fig. 9 shows all the features are correlated to each other, this could result in enhancing the readability of the charts by reorganizing the factors or the parallel coordinates based on the correlations combination method. On the next section a comparison between two snap shots Fig. 6 represent the parallel coordinates using the 1st method and Fig. 7 represent the 2nd method. If analyzed, the lines in Fig. 7 between Cigarette, Family Diabetes History and High blood pressure features are organized and correlated.

Fig. 6, the lines are highly interventions which lead to difficulties in interpreting.

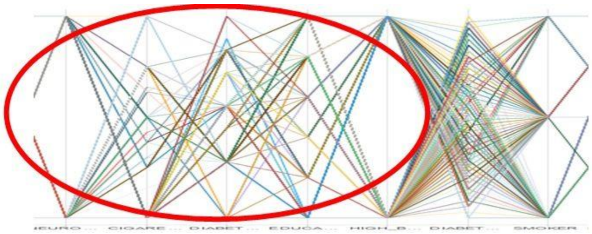


Fig. 6. 1st Method Showing Lines between Features.

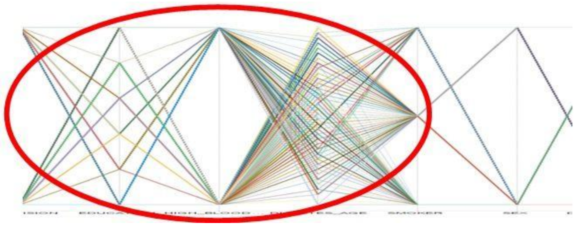


Fig. 7. 2nd Method Showing Lines between Features.

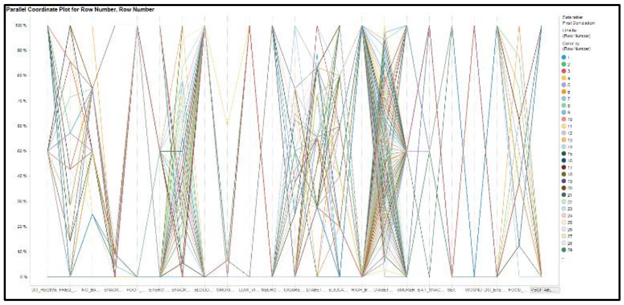


Fig. 8. Parallel Coordinates Chart based on the 1st Method Calculating the Correlation for Each Feature with Respect to Age Feature.

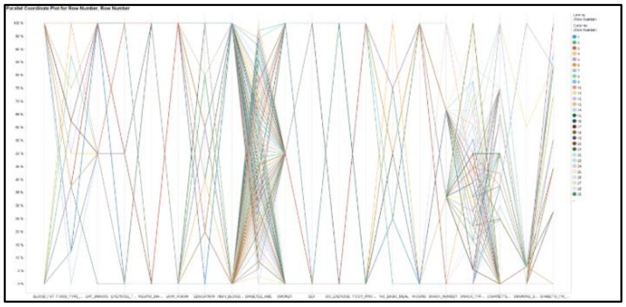


Fig. 9. Parallel Coordinates Chart based on the 2nd Method Calculating the Correlation by Merging the Previous Features with the Respect to Age Feature.

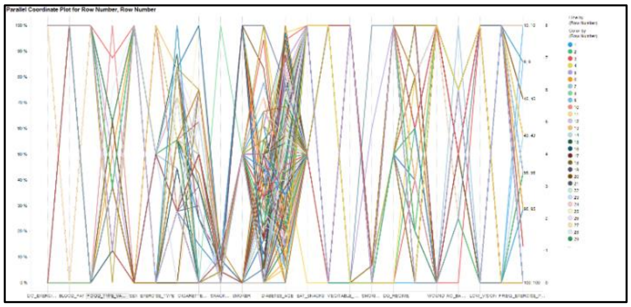


Fig. 10. Parallel Coordinates Chart based on the 3rd Method Calculating the Entropy Function for Each Feature then Rearrange the Coordinates Accordingly.

In this section, Fig. 11, 12 and 13, brushed data to focus on the Education feature and specifically the highest level of educational patients, as we can see the difference between the three graphs, where the features aren’t correlated between each other in the first graph, forming random lines between different features. On the other hand, most of the features are either positive or negative correlations in the second chart.

Furthermore, as a quick notice can be reached from the second figure, all people with high level of education are much stressed and most of them suffer from high blood pressure, for this reason they probably may have a high risk to be candidate of diabetes disease at a younger age. Also the second figure is still better than the third chart.

Another example, Fig. 14, 15 and 16, when brushing the data for people who are smoking, the same result like previous charts, most of the features are significant correlated with the second method other than the first and third method. Also extracting useful information from the second chart, for instance most of the male patients are smoking and they are a positive correlation to be a candidate of the diabetes disease.

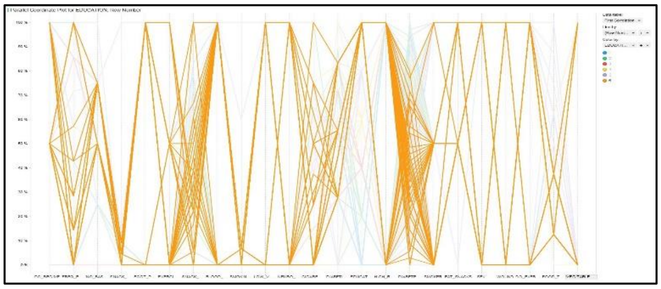


Fig. 11. 1st Method Focus on the Education Feature and Specially the Highest Level of Education.

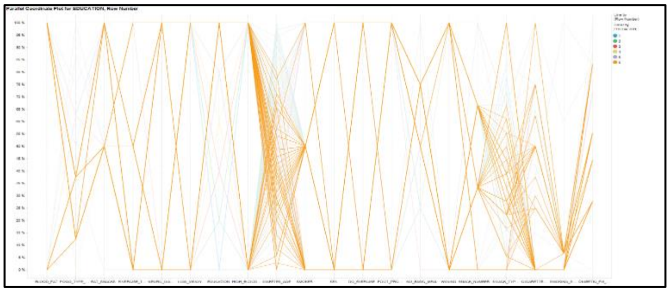


Fig. 12. 2nd Method Focus on the Education Feature and Specially the Highest Level of Education.

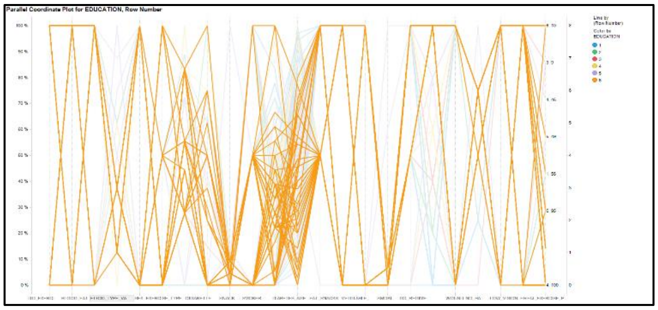


Fig. 13. 3rd Method Focus on the Education Feature and Specially the Highest Level of Education.

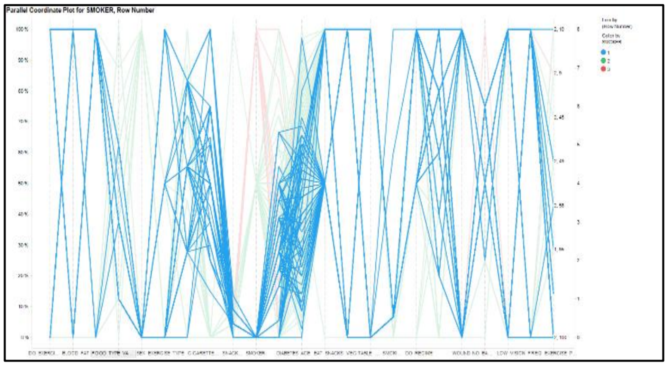


Fig. 14. 1st Method Focus on Smoking Feature and Specially the Smokers.

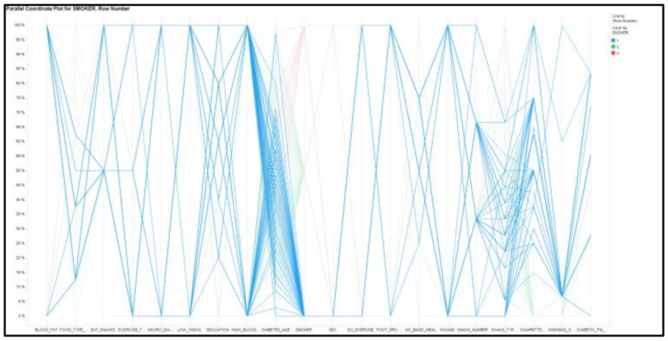


Fig. 15. 2nd Method Focus on Smoking Feature and Specially the Smokers.

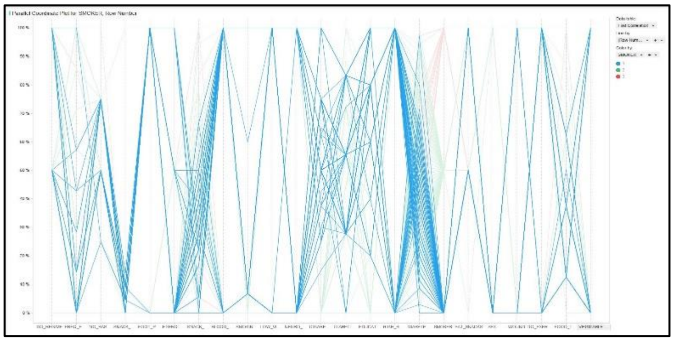


Fig. 16. 3rd Method Focus on Smoking Feature and Specially the Smokers.

V. Conclusion

In this paper, three techniques to reorder the coordinates of the charts were introduced. Two of these techniques based on the correlation coefficient and the third one based on the entropy function. The goals of these techniques to enhance the parallel coordinate visualization and facilitate the interpretation of data.

Concluding based on the analysis and by comparison, the second method results a better visualization than others. New information was interpreted and extract from the charts. In the future work a plan to merge between the three techniques with the clustering methodology. Moreover, further analysis and discussion will be held between the old and the new charts.

References

[1] G. Noseworthy, "Infographic: Managing the Big Flood of Big Data in Digital Marketing," [Online]. Available: http://analyzingmedia.com/2012/infographic-big-flood-of-big-data-in-digitalmarketing/.

[2] V. T. M. S. Carrie MacGillivary, "IDC's Worldwide Internet of Things Taxonomy," IDC, 2015.

[3] M. G. B. Akbar, "Data Analytics Enhanced Data Visualization and Interrogation with Parallel Coordinates Plots," in 26th International Conference on Systems Engineering, ICSEng 2018 , 2019.

[4] T. M. T. A. S. Kaidi Zhao, "Detecting Patterns of Change Using Enhanced Parallel Coordinates Visualization," in ICDM '03 Proceedings of the Third IEEE International Conference on Data Mining, ,2003.

[5] X. Y. Z. G. X. Huamin Qu, "Scattering Points in Parallel Coordinates," IEEE Transactions on Visualization & Computer Graphics, vol. 15, pp. 1001-1008,, 2009.

[6] W. Sun and S. Wang, "A new data mining method for early warning landslides based on parallel coordinate," in Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services, 2011.

[7] G. R. ,. T. J. ,. F. L. D. A. a. R. B. Joris Sansen *, "Visual Exploration of Large Multidimensional Data Using Parallel Coordinates on Big Data Infrastructure," Informatics, vol. 7, 2017.

[8] Z. Y. T. I. F. Haruka Suematsu, "Arrangement of Low-Dimensional Parallel Coordinate Plots for High-Dimensional Data Visualization," in 2013 17th International Conference on Information Visualisation, 2013.

[9] K. Zhao, B. Liu, T. Tirpak and A. Schaller, "Detecting Patterns of Change Using Enhanced Parallel Coordinates Visualization," in Third IEEE International Conference on Data Mining, 2003.

[10] B. D. Alfred Inselberg, "Parallel coordinates: a tool for visualizing multi-dimensional geometry," in Proceedings of the First IEEE Conference on Visualization: Visualization `90, 1990.

[11] M. W. E. R. Y. Fua, "Hierarchical parallel coordinates for exploration of large datasets," in Proceedings Visualization '99 (Cat. No.99CB37067), 1999.

[12] P. L. M. J. M. C. J. Johansson, "Revealing Structure within Clustered Parallel Coordinates Displays,," in INFOVIS '05 Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization, 2005.

[13] F. C. P. A. A. M. V. Elena Geanina ULARU, "Perspectives on Big Data and Big Data Analytics," Database Systems Journal, Vols. vol. III, no. 4/2012, pp. 3-14, 2012.

[14] X. Y. H. Q. W. C. B. C. H. Zhou, "Visual Clustering in parallel Coordinates," in EuroVis'08 Proceedings of the 10th Joint Eurographics / IEEE - VGTC conference on Visualization, 2008.

[15] S. T. S. J. Hemant Mekwana, "Axes Re-ordering in parallel coordinate for pattern Optimization," International Journal of Computer Applications , vol. Volume 40– No.13, pp. 42-47, 2012.

[16] L. F. Lu, M. L. Huang and T.-H. Huang, "A New Axes Re-ordering Method in Parallel Coordinates visualization," in 11th International Conference on Machine Learning and Applications, 2012.

[17] J. Z. B. H. R. Rosenbaum, "Progressive Parallel Coordinates," in IEEE Pacific Visualization Symposium, 2012.

[18] H. Siirtola, T. Laivo, T. Heimonen and K.-J. Räihä, "Visual Perception of Parallel Coordinate Visualizations," in 13th International Conference Information Visualisation, 2009.

[19] Tran Van Long, "Visualizing High-density Clusters in Multidimensional Data," Jacobs University, 2009.

[20] J.-B. M. a. J. J. V. W. J. Li, "Judging correlation from scatter plots and parallel coordinate plots," Information Visualization, vol. Volume 9 Issue 1, pp. 13-30, 2010.

[21] C. Kamath, Scientific Data Mining : A practical Perspective,, Society for Industrial and Applied Mathematics, 2009.

A Map-based Job Recommender Model

Manal Alghieth[1], Amal A. Shargabi[2]
Information Technology Department
College of Computer, Qassim University, Qassim, Saudi Arabia

***Abstract*—Location is one of the most important factors to consider when looking for offering a new job. Currently, there exist many job recommender systems to help match the right candidate with the right job. A review of the existing recommender systems, included within this article, reveals that there is an absence of appropriate mapping support offering for job recommendation. This article aims to propose a general map-based job recommender model, which is implemented and applied within a system for job seekers in Saudi Arabia. The system adapts content-based technique to recommend jobs using the cosine similarity and will help Saudi job seekers finding their desired job in an efficient way using interactive maps. This ultimately will contribute to Saudi Arabia moving forward to the digital transformation which is one of the major objectives to fulfill the Saudi vision 2030.**

Keywords—Recommender systems; content-based recommendation; location-based search; maps

I. Introduction

Finding a job in today's market is a major challenge. A common way to look for a job is to use job search websites. Rather than taking the time to search newspapers, company web sites, and other traditional job postings. A job search website can do it all with the click of a button. A job search engine facilitates the matching and communication of job opportunities between job seekers and employers.

The location of a job has the potential to significantly affect an individual's lifestyle. People often tend to focus their job search on a particular area and the job location can play an important part in the decision to apply for a job. Thus, location is one of the most important factors to consider when looking for a new job. As such, every job search website worth using has the ability to search for jobs based on location.

In the most popular job search websites, e.g. Indeed (www.indeed.com) and Monster (www.monster.com), the representation for the retrieved information may not be appropriate for job seekers with respect to the job location. The reason is because the results returned to a job seeker about the job locations are in textual form with no provision of a map for the employers' geospatial location.

As such, users of such websites may find difficult searching for their desired job, and not efficient because the lack of the website support in displaying the available jobs on a map might lead to them spending a significant amount of their time reading and reviewing their options regarding job location.

Two job search websites, namely Glassdoor (https://www.glassdoor.com/Job/explorer/index.htm) and Pathwayjobs (www.pathwayjobs.com), provide mapping tools for job search but their tools are specifically designed for American job seekers only and they need to be personalized. They require recommender systems to meet job seekers preferences as thousands of jobs are posted on these websites daily, and it takes a great deal of effort to find the right position.

Currently, there exist many personalized job search systems, i.e. job recommender systems to help match the right candidate with the right job. Examples include CASPER [1], Proactive [2], FES [3], PROSPECT [4], eRecruiter [5], iHR [6], RésuMatcher [7] and the work of [8]. The work of [9] and [10] provides comprehensive review on job recommender systems.

Table I shows a general comparison among these systems based on the following characteristics: recommendation input, recommendation technique, and the offering of mapping support in these systems.

All the recommender systems of Table I support personalization in the job search, although they differ in the recommendation technique used. Most of the system use collaborative filtering recommendation (CFR) and content-based recommendation (CBR). Unfortunately, most of these systems are poor in personalized search because their search functionality is limited to keyword-based search, often resulting in poor, irrelevant search results. For example, a job search using the keyword "Java" to search for jobs within a limited geographical location (New York, NY) on www.indeed.com returned over 8000 jobs.

In the context of this research, it is worth mentioning that none of them offer mapping support tools. There are a number of international job search web sites, including Indeed (www.indeed.com), Monster (www.monster.com), Glassdoor (www.glassdoor.com),CareerBuilder(www.careerbuilder.com), SimplyHired(www.simplyhired.com),Pathwayjobs(www.path wayjobs.com), and LinkedIn (www.linkedin.com), that provide geospatial search. To the best of our knowledge, only two of these websites, namely Glassdoor and Pathwayjobs, provide mapping tools for job search and both of them are for American job seekers only. Anyway, none of them are recommender systems.

This research aims is to bridge the above two mentioned gaps in the current job search websites and systems, and proposes a personalized model based on job seeker preferences and also support location-based search with interactive mapping tool. The proposed model and prototype will help job seekers, especially in Saudi Arabia, to find the right job that meets their qualifications in an efficient way using interactive maps.

TABLE. I. A SUMMARY OF EXISTING JOB RECOMMENDATION SYSTEMS

Job Recommender System	**Year**	**Reference**	**Recommendation Input (User Profile)**	**Recommendation Technique**	**Mapping Support**
CASPER	2000	[1]	Personal information User behaviour (revisited data, read time data, activity data, feedback) Query (job description, salary, location, education background)	CFR CBR	No
Proactive	2000	[2]	Personal information (includes preferences and interests either predefined or general)	CBR (Ontology) KBR	No
FES	2007	[3]	Personal information (Qualification, class of degree, years of experience, certifications, age, course of study)	Fuzzy-based	No
PROSPECT	2010	[4]	Personal information	CBR (Resume Mining)	No
eRecruiter	2011	[5]	Personal information (include interest) User behaviour (include feedback in real-time)	CBR KBR (Ontology)	No
iHR	2013	[6]	Personal information User behaviour	CBR CFR Hybrid Cluster-based	No
Social JRS	2013	[11]	Users data, users friends data, users profile	SVM	No
iHR+	2015	[12]	Basic information only	Lucene IKAnalyzer TF-ADF	No
SKILL	2015	[13]	Skills in a resume , requirements in job poster	Skills taxonomy, skills tagging	No
RésuMatcher	2016	[7]	Resume document	Statistical similarity FST	No
Graph-based recommender	2017	[8]	Personal information User behaviour (active and passive users)	CBR Graph-based	No
DNN Model for Job Matching	2018	[14]	Personal information (age, gender, major, profession, educational level and employment history)	DNN	No

These research work main research contributions are:

- Construction of a map-based job recommender model.
- Evaluation of the constructed job model using a simple prototype.

The rest of the article is organized as follows: Section II presents the proposed map-based job recommender model. The section starts with a subsection on the theoretical base of the proposed model and then explains the proposed model elements in detail. Section III presents the several modules of a prototype implemented based on the proposed model and applied for Saudi job seekers. Finally, Section IV provides a simple comparison between the proposed prototype and other Saudi job recruitment systems. Finally Section V summarizes and concludes the article.

II. A PROPOSED MAP-BASED JOB RECOMMENDER MODEL

Despite the differences among the different job recommender systems, most of them have the same general architecture, as the proposed model does. Therefore, before describing in detail the proposed model, its general architecture is first introduced in the following subsection.

A. The General Architecture of Recommendation Systems

The general architecture of recommendation systems consists of three basic elements: the input data, the recommendation technique, and the recommendation output (Fig. 1).

The input is the users' preferences used for recommendation such as user profile. The input captures the main preferences of users and is the content of the user profile, which can contain one or more of three types of data: 1) individual information such as educational experience, working experience and skills; 2) user's historical behaviors regarding job application and collecting job posts, e.g. CASPER and eRecruiter; and 3) user preference captured based on a description of a preferred job, e.g. CASPER, or based on mining the resume of the users, e.g. PROSPECT and eRecruiter. The input data also include the available jobs provided by recruiters.

Input
Recommendation Technique
Output

Fig. 1. The General Architecture of Recommendation System.

The recommendation technique is the core of the recommender system, and it refers to the recommendation strategy to use. In the literature, several techniques have been introduced based on the following filtering techniques:

1) Collaborative filtering makes a user-to-user comparison in order to suggest preferred items from similar users. For suggesting items, collaborative filtering recommender systems search for users with similar taste. Most often, such recommender systems are based on ratings. These ratings are then used to match similar users; for example, two users have the same rating for the same article. Afterwards, the preferences of similar users are re-used for recommending items. An example of collaborative filtering recommenders is [15].

2) Content-based filtering tries to find items that are similar to items the user likes. Content-based recommender systems analyze item descriptions to identify items of particular interest to the user. This type of recommender system may be used in a variety of different domains, such as web page recommendations, television programs, news articles, and social media content. This approach has its roots in the information retrieval area, as methods for searching for documents are involved. Compared to traditional information retrieval applications, these recommender systems require user profiles, which encode the user's preferences. An example of collaborative filtering recommenders is [4].

3) Knowledge-based uses additional knowledge in order to infer items that best match the user's needs. Knowledge-based recommender systems use knowledge about the items, the users and on how to map users' needs to items' features, the so-called functional knowledge. Knowledge-based recommender systems suggest products based on inferences about a user's needs and preferences that are derived explicitly from their mapping to product features [14]. An example of such recommenders is [16].

4) Hybrid filtering combines two or more of the above techniques to achieve better performance. An example of these filtering is [17].

Different systems may employ different recommendation strategies approaches based on their own user profiles. For example, PROSPECT uses content filtering techniques while CASPER applies hybrid techniques based on content and collaborative filtering techniques.

After taking the user profile as the input and applying the recommendation technique, the recommendation system outputs the recommendation results that satisfy the desires of users. The output is usually in the form of a list of recommended jobs. It may also include "You Maybe Also Like" and "What Others Looking" jobs.

B. The Proposed Model

The proposed model is based on the general architecture of a recommendation system presented in the previous section (see Fig. 2). The details of its three elements are provided below.

- Input

This element consists of the job seeker and job vacancies data. The job seekers data are the preferences of the people who are looking for jobs. In the proposed model two main preferences have been identified: job title and location. The location in the proposed model is central because the proposed model is a map-based model. The job vacancies data are the jobs postings to be retrieved from common job boards' sites.

- Recommender

This element gets the job seeker preferences, i.e., job title and location as well as the job retrieved from job boards' sites and recommend the most similar jobs. The recommendation technique in the proposed model is content-based.

The basic approach in content-based filtering is to match similarity between two texts based on the count of common words between the two documents. This provides a measure of the 'Euclidean distance' between the two documents. As the size of documents increases, the number of common words would increase, even if the documents are related to the same topic/idea. The Euclidean distance, therefore, stops being a good measure of the match between the documents. However, a measure that can indicate the closeness of two documents irrespective of their size is the cosine of the angle between two associated vectors, i.e. arrays having word counts in the two documents, projected in a multidimensional space (Fig. 3). The axes correspond to search words being compared. It is obvious that we might have a large Euclidean distance between very similar vectors, i.e. between vectors with a small angle (hence a larger value of cosθ). The cosine gives a normalized measure of the match: 0 for no match to 1 for a perfect match.

- Output

The output is a visualizer element which is responsible for providing a map support for the recommended jobs. The visualizer gets the recommended jobs locations and processes them and then shows them on an interactive map.

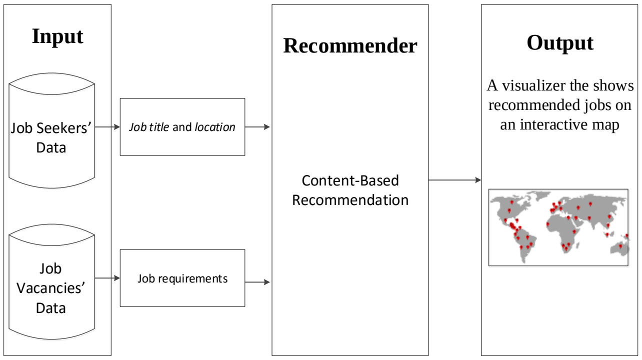

Input
Job Seekers' Data
Job title and *location*
Job Vacancies' Data
Job requirements
Recommender
Content-Based Recommendation
Output
A visualizer the shows recommended jobs on an interactive map

Fig. 2. The Architecture of the Proposed Model.

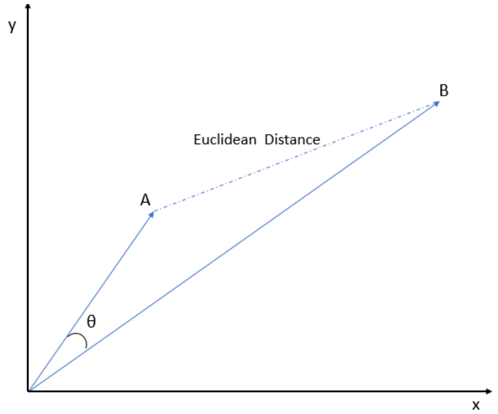


Fig. 3. Cosine Similarity.

III. A Prototype based on the Proposed Model

To realize and evaluate the proposed model in the previous section, a corresponding prototype is developed. The prototype was implemented for Saudi job seekers and consists of three modules (Fig. 4): job grabber, job recommender, location geocoder and geometry visualizer.

As shown in the figure, there are three modules: job grabber, job recommender, location geocoder and geometry visualizer. The following subsections describe each module in detail.

A. The Job Grabber

The proposed system is meant for Saudi job seekers; thus, the module extracts a list of available jobs posted online in Saudi Arabia. Specifically, the module is based on Mihnati.com, one of the most popular job boards in Saudi Arabia and Gulf countries. This module scrapes the jobs posted on Mihnati.com and saves then save them as a Comma-separated values file, i.e., CSV file. The jobs posted are grabbed either they were in English or Arabic languages.

Three main elements of each job are extracted: job title, address and URL. The address here is essential to be used later to the recommended jobs on a map.

To implement this module, a web scrapper is developed using Python language. Specifically, the Beautiful Soup library is used for pulling jobs out. A sample of the grabbed jobs is shown in Fig. 5.

B. Job Recommender

The job recommender module is a content-based recommender system. This module reads the scrapped jobs from the CSV file produced from the job extractor module and recommends jobs based on job seeker filters. These filters are job title and location.

The entered job title filter is checked for cosine similarity against jobs saved in the CSV file and shortlisted in decreasing order based on their cosine-similarity values, with job location being used as a secondary ordering criterion of the sorting. The jobs within a certain distance from the preferred location are included in the shortlist, and sorting is done based on increasing distance from the preferred location. The algorithm used in this module is described.

Algorithm 1: Recommendation Algorithm used in job recommender module

```
Define the minimum cosine-similarity threshold for a match;
Define the maximum distance threshold between job location and preferred location;
Read user job title and preferred location;
Define a data-frame with the columns Job title, Job location and Job link;
Load job records from the CSV file into the jobdata data frame;
Define a job_shortlist data frame with columns: jobdata-index, cosine similarity and distance;
For each job title in the data-frame
        Create a vector of common words between job-search title and job title on file;
        Calculate the cosine-similarity;
For records with cosine >= cosine threshold
        Calculate distance between location preference and job location;
If distance <= distance threshold
        Move job-data index, cosine-similarity, distance to shortlist data frame;
Sort the shortlist data-frame on decreasing cosine-similarity and increasing distance;
Use the index in the short-list record to retrieve the job-data record;
Display the job title, job location and job link in the sorted order;
```

The cosine similarity was chosen as the metric to rank the recommendations, as the job-search text is usually small and hence the number of common words (Euclidean distance) is not a good indicator of a match in this content-based recommendation. A minimum threshold of 0.8 was used for the cosine similarity, although this is customizable. The vector orientation is more important in this work than the magnitude. The search is further narrowed down by job-location comparison. The results from the recommendation can also be plotted on a map as highlighted locations within a radius of the specified maximum distance from the preferred location.

The module is implemented in Python and uses packages CSV to read comma-separated values for the database file, Counter and CountVectorizer to create vectors (word count arrays), and cosine_similarity to calculate cosine similarity. To calculate the distance between two cities, the package Nominatim is imported to get geocoordinates, i.e., latitude and longitude, of a place, and a package distance is imported to calculate the distance between two geo-coordinates. Package pandas are imported to use the Python data-frame structure.

The module can be further extended by including cosine-similarity calculations on other parameters like work experience, expected vs offered compensation, etc. It can also be adapted to take input directly from a resume rather than a user.

The recommended jobs are shown in Fig. 6.

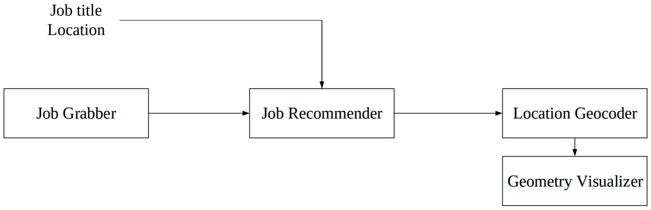


Fig. 4. Prototype Architecture.

C. Location Geocoder

Geocoding is the process of converting addresses into geographic coordinates, i.e. latitude and longitude.

This module reads the addresses of the recommended jobs and formats the address to be complete if they are not. The geographic coordinates of the formatted addresses are then found. As an example, the address of the third recommended job in Fig. 6 was extracted as:

Al Othman Agriculture Production and Processing Company (NADA), Khobar, Saudi Arabia

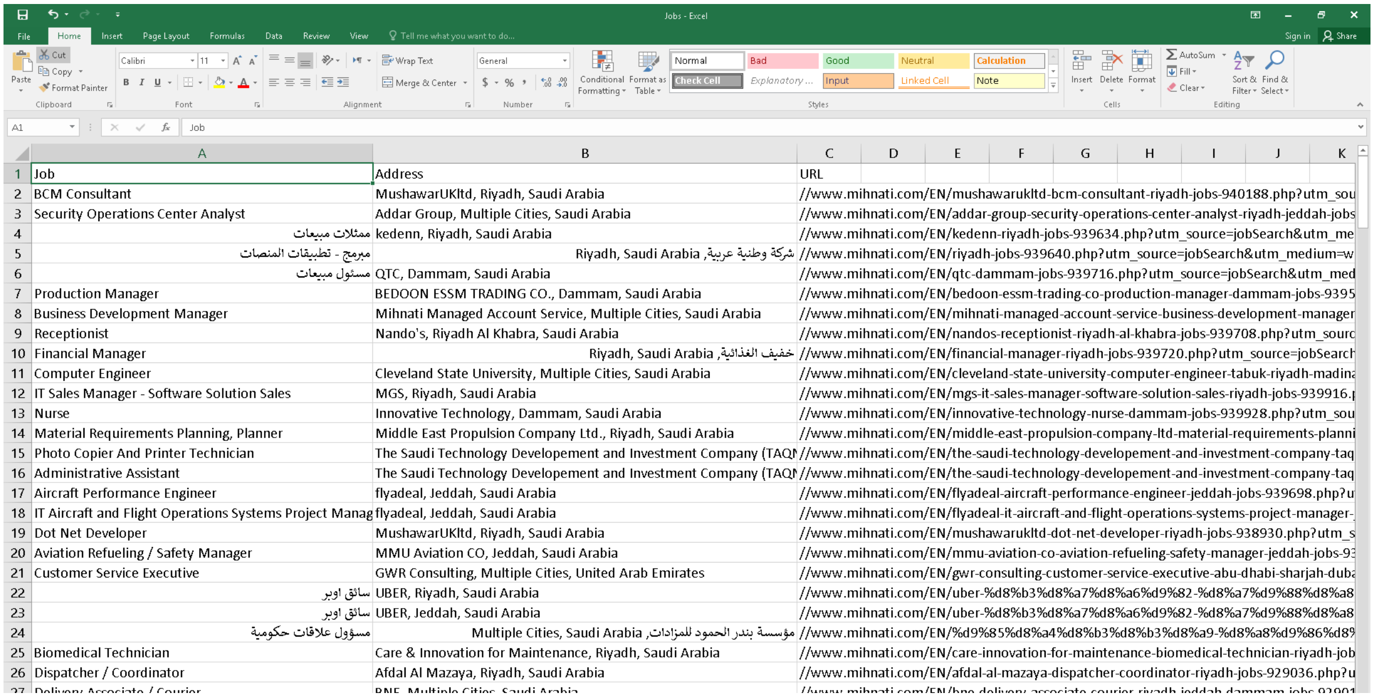
In order to be able to visualize the address on a map, the Location Geocoder module finds the complete address, i.e. formatted address as well as the geographic coordinates, i.e. geometry, in terms of latitude and longitude.

The results got form; this module is shown in Fig. 7. To implement this module, Google Geocoding API is used. In order to use Google APIs, an API key is needed. This key should be later embedded in the code.

D. Geometry Visualizer

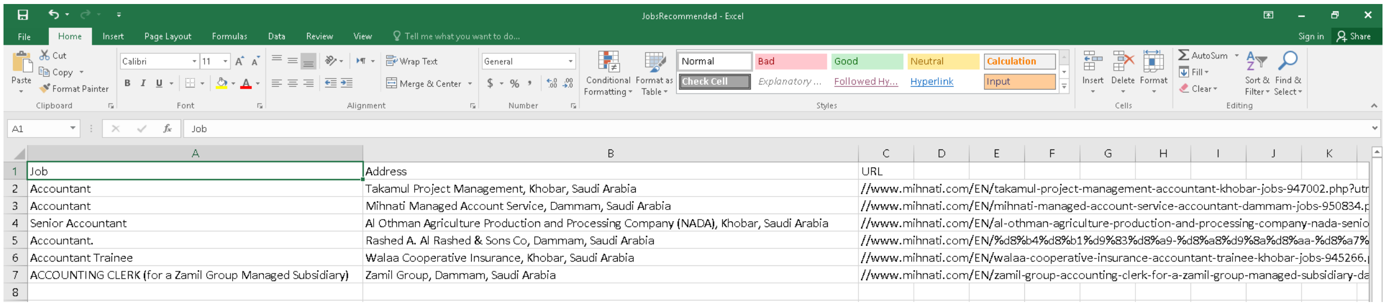
Markers are one of the most common ways to visualize locations on a map. In this module, the geocoded addresses of the recommended jobs are visualized on a map using markers. Fig. 8 shows on a map the locations of the six recommended jobs shown in Fig. 6.

To implement this module, Google Maps JavaScript API is used.



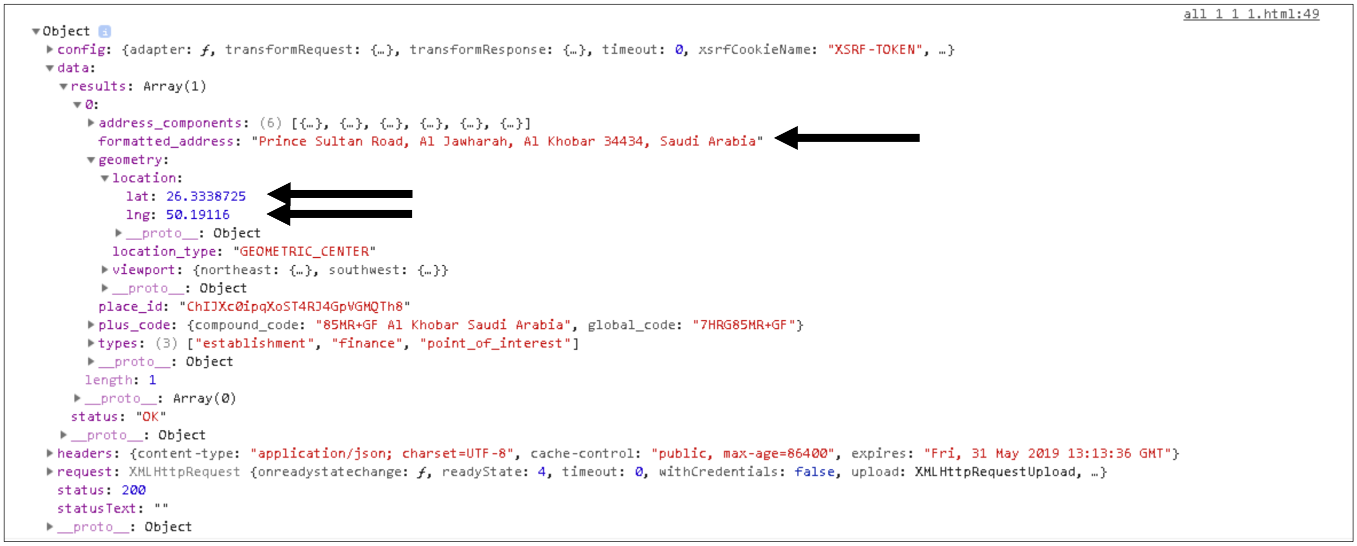
	A	B	C	D	E	F	G	H	I	J	K
1	Job	Address	URL								
2	BCM Consultant	MushawarUKltd, Riyadh, Saudi Arabia	//www.mihnati.com/EN/mushawarukltd-bcm-consultant-riyadh-jobs-940188.php?utm_sou								
3	Security Operations Center Analyst	Addar Group, Multiple Cities, Saudi Arabia	//www.mihnati.com/EN/addar-group-security-operations-center-analyst-riyadh-jeddah-jobs								
4	ممثلات مبيعات	kedenn, Riyadh, Saudi Arabia	//www.mihnati.com/EN/kedenn-riyadh-jobs-939634.php?utm_source=jobSearch&utm_me								
5	مبرمج - تطبيقات المنصات	شركة وطنية عربية, Riyadh, Saudi Arabia	//www.mihnati.com/EN/riyadh-jobs-939640.php?utm_source=jobSearch&utm_medium=w								
6	مسئول مبيعات	QTC, Dammam, Saudi Arabia	//www.mihnati.com/EN/qtc-dammam-jobs-939716.php?utm_source=jobSearch&utm_med								
7	Production Manager	BEDOON ESSM TRADING CO., Dammam, Saudi Arabia	//www.mihnati.com/EN/bedoon-essm-trading-co-production-manager-dammam-jobs-9395								
8	Business Development Manager	Mihnati Managed Account Service, Multiple Cities, Saudi Arabia	//www.mihnati.com/EN/mihnati-managed-account-service-business-development-manager								
9	Receptionist	Nando's, Riyadh Al Khabra, Saudi Arabia	//www.mihnati.com/EN/nandos-receptionist-riyadh-al-khabra-jobs-939708.php?utm_sourc								
10	Financial Manager	خفيف الغذائية, Riyadh, Saudi Arabia	//www.mihnati.com/EN/financial-manager-riyadh-jobs-939720.php?utm_source=jobSearch								
11	Computer Engineer	Cleveland State University, Multiple Cities, Saudi Arabia	//www.mihnati.com/EN/cleveland-state-university-computer-engineer-tabuk-riyadh-madina								
12	IT Sales Manager - Software Solution Sales	MGS, Riyadh, Saudi Arabia	//www.mihnati.com/EN/mgs-it-sales-manager-software-solution-sales-riyadh-jobs-939916.p								
13	Nurse	Innovative Technology, Dammam, Saudi Arabia	//www.mihnati.com/EN/innovative-technology-nurse-dammam-jobs-939928.php?utm_sou								
14	Material Requirements Planning, Planner	Middle East Propulsion Company Ltd., Riyadh, Saudi Arabia	//www.mihnati.com/EN/middle-east-propulsion-company-ltd-material-requirements-planni								
15	Photo Copier And Printer Technician	The Saudi Technology Developement and Investment Company (TAQN	//www.mihnati.com/EN/the-saudi-technology-developement-and-investment-company-taq								
16	Administrative Assistant	The Saudi Technology Developement and Investment Company (TAQN	//www.mihnati.com/EN/the-saudi-technology-developement-and-investment-company-taq								
17	Aircraft Performance Engineer	flyadeal, Jeddah, Saudi Arabia	//www.mihnati.com/EN/flyadeal-aircraft-performance-engineer-jeddah-jobs-939698.php?u								
18	IT Aircraft and Flight Operations Systems Project Manag	flyadeal, Jeddah, Saudi Arabia	//www.mihnati.com/EN/flyadeal-it-aircraft-and-flight-operations-systems-project-manager-								
19	Dot Net Developer	MushawarUKltd, Riyadh, Saudi Arabia	//www.mihnati.com/EN/mushawarukltd-dot-net-developer-riyadh-jobs-938930.php?utm_s								
20	Aviation Refueling / Safety Manager	MMU Aviation CO, Jeddah, Saudi Arabia	//www.mihnati.com/EN/mmu-aviation-co-aviation-refueling-safety-manager-jeddah-jobs-93								
21	Customer Service Executive	GWR Consulting, Multiple Cities, United Arab Emirates	//www.mihnati.com/EN/gwr-consulting-customer-service-executive-abu-dhabi-sharjah-duba								
22	سائق اوبر	UBER, Riyadh, Saudi Arabia	//www.mihnati.com/EN/uber-%d8%b3%d8%a7%d8%a6%d9%82-%d8%a7%d9%88%d8%a8								
23	سائق اوبر	UBER, Jeddah, Saudi Arabia	//www.mihnati.com/EN/uber-%d8%b3%d8%a7%d8%a6%d9%82-%d8%a7%d9%88%d8%a8								
24	مسؤول علاقات حكومية	مؤسسة بندر الحمود للمزادات, Multiple Cities, Saudi Arabia	//www.mihnati.com/EN/%d9%85%d8%a4%d8%b3%d8%b3%d8%a9-%d8%a8%d9%86%d8								
25	Biomedical Technician	Care & Innovation for Maintenance, Riyadh, Saudi Arabia	//www.mihnati.com/EN/care-innovation-for-maintenance-biomedical-technician-riyadh-job								
26	Dispatcher / Coordinator	Afdal Al Mazaya, Riyadh, Saudi Arabia	//www.mihnati.com/EN/afdal-al-mazaya-dispatcher-coordinator-riyadh-jobs-929036.php?u								
27	Delivery Associate / Courier	BNE, Multiple Cities, Saudi Arabia	//www.mihnati.com/EN/bne-delivery-associate-courier-riyadh-jeddah-dammam-jobs-92901								

Fig. 5. Jobs Grabbed from Mihnati.com.



	A	B	C	D	E	F	G	H	I	J	K
1	Job	Address	URL								
2	Accountant	Takamul Project Management, Khobar, Saudi Arabia	//www.mihnati.com/EN/takamul-project-management-accountant-khobar-jobs-947002.php?utm								
3	Accountant	Mihnati Managed Account Service, Dammam, Saudi Arabia	//www.mihnati.com/EN/mihnati-managed-account-service-accountant-dammam-jobs-950834.p								
4	Senior Accountant	Al Othman Agriculture Production and Processing Company (NADA), Khobar, Saudi Arabia	//www.mihnati.com/EN/al-othman-agriculture-production-and-processing-company-nada-senio								
5	Accountant.	Rashed A. Al Rashed & Sons Co, Dammam, Saudi Arabia	//www.mihnati.com/EN/%d8%b4%d8%b1%d9%83%d8%a9-%d8%a8%d9%8a%d8%aa-%d8%a7%								
6	Accountant Trainee	Walaa Cooperative Insurance, Khobar, Saudi Arabia	//www.mihnati.com/EN/walaa-cooperative-insurance-accountant-trainee-khobar-jobs-945266.p								
7	ACCOUNTING CLERK (for a Zamil Group Managed Subsidiary)	Zamil Group, Dammam, Saudi Arabia	//www.mihnati.com/EN/zamil-group-accounting-clerk-for-a-zamil-group-managed-subsidiary-da								
8											

Fig. 6. Recommended Jobs.



```
▼Object
 ▶config: {adapter: ƒ, transformRequest: {…}, transformResponse: {…}, timeout: 0, xsrfCookieName: "XSRF-TOKEN", …}
 ▼data:
  ▼results: Array(1)
   ▼0:
    ▶address_components: (6) [{…}, {…}, {…}, {…}, {…}, {…}]
     formatted_address: "Prince Sultan Road, Al Jawharah, Al Khobar 34434, Saudi Arabia"
    ▼geometry:
     ▼location:
        lat: 26.3338725
        lng: 50.19116
      ▶__proto__: Object
      location_type: "GEOMETRIC_CENTER"
     ▶viewport: {northeast: {…}, southwest: {…}}
     ▶__proto__: Object
     place_id: "ChIJXc0ipqXoST4RJ4GpVGMQTh8"
    ▶plus_code: {compound_code: "85MR+GF Al Khobar Saudi Arabia", global_code: "7HRG85MR+GF"}
    ▶types: (3) ["establishment", "finance", "point_of_interest"]
    ▶__proto__: Object
    length: 1
   ▶__proto__: Array(0)
  status: "OK"
  ▶__proto__: Object
 ▶headers: {content-type: "application/json; charset=UTF-8", cache-control: "public, max-age=86400", expires: "Fri, 31 May 2019 13:13:36 GMT"}
 ▶request: XMLHttpRequest {onreadystatechange: ƒ, readyState: 4, timeout: 0, withCredentials: false, upload: XMLHttpRequestUpload, …}
  status: 200
  statusText: ""
 ▶__proto__: Object
```

Fig. 7. Geocoded Address of a Selected Recommended Job, Arrows Shows the Formatted Address, Latitude and Longitude of a given Address.

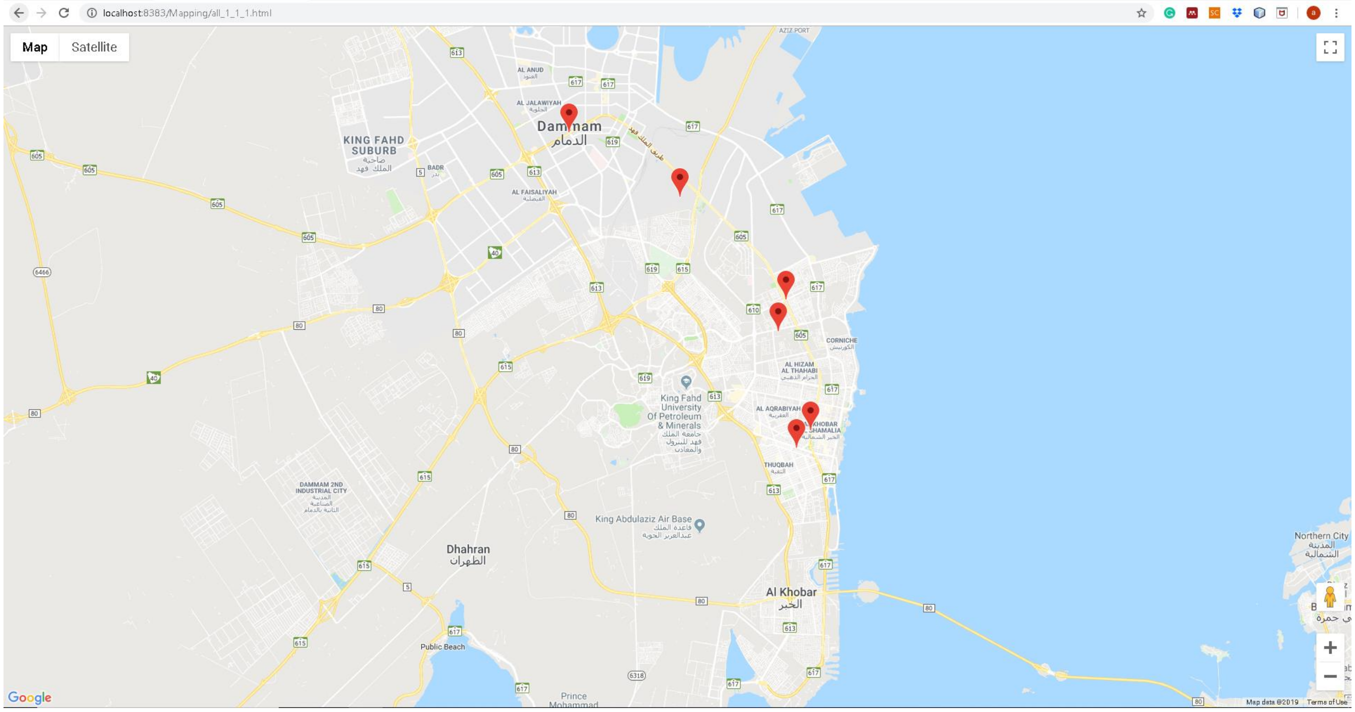


Fig. 8. The Recommended Jobs are Shown on a Map.

IV. A Comparison between the Map-based Job Recommender Prototype and other Saudi Job Systems

As an initial evaluation for the proposed model, the prototype that has been implemented based on that model is compared with JADARAH. The most commonly used system by Saudi job seekers. JADARAH was developed on 2011 by the ministry of civil service to help Saudi job seekers find jobs within Saudi government sectors. The comparison is made in terms personalization, map-based support, and type of jobs offered.

For personalization criterion, both systems are recommender systems. That is, the jobs are personalized based on job seeker's qualifications and preferences. However, in our prototype, the location is highly considered during the recommendation process. We believe that, in Saudi Arabia, job location is essential when looking for a new job. This is due to the large geographical area of the country and the long distances among its several regions. Unlike our system, JADARAH does not pay great attention to job seeker location. That is, a person who lives in Qassim which is located in the central region may be got a job in Abha which is located in the southern region.

The proposed system is map-based as the recommended jobs are displayed on a map. Unlike our system, JADARAH displays the recommended job in textual form.

In terms of types of jobs, our system considers jobs from private companies as well as governmental sectors. On the other hand, JADARAH is only for governmental jobs.

It is worth mentioning that there are other job requirement systems in Saudi Arabia such as bayt.com, wadhefa.com, and tanqeeb.com, however, these systems were not included in the comparison as these systems do not provide personalized search, i.e. not recommender systems.

V. Conclusion

Although there are many job recommendation systems, these systems do not offer mapping support. In line with the digital transformation objectives of the new Saudi vision 2030, and in order to improve job search in general and in Saudi Arabia in particular, this work proposed a personalized and map-based job search model. The model was theoretically based on the existing recommender systems in the literature and used content-based recommendation with integration of mapping feature for location-based search which has never been used in the previous systems. The cosine similarity was used for the content-based recommendation with a minimum threshold of 0.8 for job title search and further narrowed down by job location comparison. As a proof of concept, a prototype was implemented based on the proposed model. The proposed system provides better features compared with JADARAH, the most common recruitment system used by Saudi job seekers. The proposed system in this research work will help Saudi job seekers finding the desired job in an efficient way using maps. In this work, the proposed system is meant by Saudi job seekers as it is based on Mihanti.com as a main source for jobs. In the future, we plan make more general and grab jobs from several international job's boards.

Acknowledgment

This research is funded by the Deanship of Scientific Research (SRD), Qassim University, Saudi Arabia. The associated project number of the research is coc_2018_1_14_S_5143.

References

[1] R. Rafter, K. Bradley, and B. Smyth, "Personalised Retrieval for Online Recruitment Services," 2000.

[2] D. H. Lee and P. Brusilovsky, "Fighting Information Overflow with Personalized Comprehensive Information Access: A Proactive Job Recommender," in Third International Conference on Autonomic and Autonomous Systems (ICAS'07), 2007, pp. 21–21.

[3] J. O. Daramola, O. O. Oladipupo, and A. G. Musa, "A fuzzy expert system (FES) tool for online personnel recruitments," Int. J. Bus. Inf. Syst., vol. 6, no. 4, p. 444, 2010.

[4] A. Singh, C. Rose, K. Visweswariah, V. Chenthamarakshan, and N. Kambhatla, "PROSPECT: A system for screening candidates for recruitment," in Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10, 2010, p. 659.

[5] M. Hutterer, "Enhancing a Job Recommender with Implicit User Feedback," vol. 2011, 2011.

[6] W. Hong, S. Zheng, H. Wang, and J. Shi, "A Job Recommender System Based on User Clustering," J. Comput., vol. 8, no. 8, Aug. 2013.

[7] S. Guo, F. Alamudun, and T. Hammond, "RésuMatcher: A personalized résumé-job matching system," Expert Syst. Appl., vol. 60, pp. 169–182, Oct. 2016.

[8] W. Shalaby et al., "Help Me Find a Job: A Graph-based Approach for Job Recommendation at Scale," Dec. 2017.

[9] Z. Siting, H. Wenxing, Z. Ning, and Y. Fan, "Job recommender systems: A survey," in 2012 7th International Conference on Computer Science & Education (ICCSE), 2012, pp. 920–924.

[10] S. T. Al-Otaibi and M. Ykhlef, "A survey of job recommender systems," Int. J. Phys. Sci., vol. 7, no. 29, pp. 5127–5142, Jul. 2012.

[11] M. Diaby, E. Viennet, and T. Launay, "Toward the next generation of recruitment tools: An online social network-based job recommender system," in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2013, 2013, pp. 821–828.

[12] H. Wenxing, C. Yiwei, Q. Jianwei, and H. Yin, "iHR+: A mobile reciprocal job recommender system," in 2015 10th International Conference on Computer Science & Education (ICCSE), 2015, pp. 492–495.

[13] M. Zhao, F. Javed, F. Jacob, and M. McNair, "SKILL: A system for skill identification and normalization," in Proceedings of the National Conference on Artificial Intelligence, 2015, vol. 5, pp. 4012–4017.

[14] S. Maheshwary and H. Misra, "Matching Resumes to Jobs via Deep Siamese Network," in WWW '18 Companion Proceedings of the The Web Conference 2018, 2018, pp. 87–88.

[15] K. Haruna, M. Akmar Ismail, D. Damiasih, J. Sutopo, and T. Herawan, "A collaborative approach for research paper recommender system," PLoS One, vol. 12, no. 10, p. e0184516, Oct. 2017.

[16] Y. Balachander and T.-S. Moh, "Ontology Based Similarity for Information Technology Skills," in 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018, pp. 302–305.

[17] Y. Wang, M. Wang, and W. Xu, "A Sentiment-Enhanced Hybrid Recommender System for Movie Recommendation: A Big Data Analytics Framework," 2018.

A Comparison Review based on Classifiers and Regression Models for the Investigation of Flash Floods

Talha Ahmed Khan[1]
Usman Institute of Technology
Pakistan British Malaysian Institute
Universiti Kuala Lumpur
Malaysia

Muhammad Alam[2]
Institute of Business Management, Pakistan
Universiti Kuala Lumpur
Malaysia

Kushsairy Kadir[3]
British Malaysian Institute
Universiti Kuala Lumpur, Malaysia

Zeeshan Shahid[4]
Institute of Business Management, Pakistan

M.S Mazliham[5]
Universiti Kuala Lumpur, Malaysia

***Abstract*—Several regions of the world have been affected by one of the natural disasters named as flash floods. Many villagers who live near stream or dam, they suffer a lot in terms of property, cattle and human lives loss. Conventional early warning systems are not up to the mark for the early warning announcements. Diversified approaches have been carried out for the identification of flash floods with less false alarm rate. Forecasting approaches includes some errors and ambiguity due to the incompetent processing algorithms and measurement readings. Process variables like stream flow, water level, water color, precipitation velocity, wind speed, wave's pattern and cloud to ground (CG) flashes have been measured for the robust identification of flash floods. A vibrant competent algorithm would be required for the investigation of flash floods with less false alarm rate. In this research paper classifiers have been applied on the collected data set so that any researcher could easily know that which classifier is competent and can be further enhanced by combining it with other algorithms. A novel comprehensive parametric comparison has been performed to investigate the classification accuracy for the robust classification of false alarms. For the better accuracy more than one process variables have been measured but still contained some false alarm. Appropriate combination of sensor was integrated to increase the accuracy in results as multi-modal sensing device has been designed to collect the data. Linear discriminant analysis, logistic regression, quadratic support vector machine, k-nearest neighbor and Ensemble bagged tree have been applied to the collected data set for the data classification. Results have been obtained in the MATLAB and discussed in detail in the research paper. The worst accuracy of the classification (62%) has been achieved by the coarse k-NN classifier that means coarse k-NN produced 38% false negative rate that is not acceptable in the case of forecasting. Ensemble bagged trees produced best classification results as it achieved 99 % accuracy and 1% error rate. Furthermore, according to the comprehensive parametric comparison of regression models Quadratic SVM found to be the worst with mean square error of 0.5551 and time elapsed 13.159 seconds. On the other hand, Exponential Gaussian process regression performed better than the other existing approaches with the minimum root mean squared error of 0.0002 and prediction speed of 35000 observations per second.**

***Keywords*—Flash floods; classification; SVM; k-NN; logistic regression; quadratic SVN; ensemble bagged trees; exponential GPR**

I. Introduction

Low cost effective solution has been designed using Android phone and Arduino. Echolocation strategy has been applied to measure the water level [1]. Bunch of sensors were deployed from the upper stream to the village. Sensors nodes were designed to observe the data. Supervisor control data acquisition (SCADA) based system was designed to forecast the floods on real time basis [2]. Torrential downpour can be considered as the main cause of flash floods. Heavy precipitation for the short time known as torrential downpour. Two meteorological radars have been used to observe the torrential downpour in Czech Republic [3]. Ultrasonic sensor based observations have been analyzed to determine the dam level [4]. Data from 2007 to 2010 were collected from three stations of Selangor to design a novel model for the prediction of flash floods using ANN. Feed forward back propagation with the tangent sigmoid function was proposed to estimate the floods. The process variables like humidity, rainfall, and temperature were taken as input and the rainfall data from the stations were set as the targets. The proposed model best results [4]. Another research elaborated a novel method by measuring the magnetic field lines by tesla meter or magnetometer to predict the flash floods. Research proved by showing the results that during the flash floods magnetic field line reduced abruptly therefore magnetic field lines that are radiated from the center of the earth can be regarded as the significant yardstick to measure the flash floods [5]. A practical early warning system must detect the flash floods in appropriate timings as there is no use of the system if the flood is detected after the starting of the event. Keeping in the mind the time constraint a flood prediction model using hybrid approach NNARX (Neural Network Autoregressive with Exogenous Input) with EKF (extended Kalman Filter) was developed. 120 samples of the data set were tested and results showed that hybrid approach worked better [6]. In a previous method a novel solution was suggested by measuring the carbon dioxide levels in the environment and soil flux as the newly discovered phenomena proved that plants take less water due to the increased amount of carbon dioxide levels. Multi-layer perceptron was applied on the collected data set to reduce the false alarms in determining the flash floods [7]. IT based flash flood monitoring was performed for the immediate emergency rescue of the affected people in Jeddah [8]. Majority casualties happen due to the lack of the data and information regarding the propagation of the flash flood. Therefore, an urban flash flood monitoring was performed to know the actual and predicted flow of the flash floods for the evacuation announcement [9]. Kalman filtering, fuzzy logic, clustering, Neural network autoregressive model with exogenous input (NNARX), Particle swarm optimization (PSO) and Support vector machine have been applied for the prediction and estimation of flash floods [10]. Fuzzy logic based a disaster management device has been designed for the announcement of exit routes during the hazard [11]. Ensemble learning model has also been designed for the better generalization model of classification [12]. The false positive rate (FP) is the date values which have been estimated wrong due to the presence of error. The negative false rate (FN) is the data that is defined falsely as negative [13]. False alarm ratio relies on the relation of the complexity and anisotropy of the sea-floor Williams et al. [14] [15]. Gaussian process regression model can be acknowledged as the competent approach for solving non-linear regression issues. It performs regression in a simple way parameterization and Bayesian. It also removes ambiguity and uncertainty in the prediction of event [16]. Data driven approaches are usually capable to determine the complex and non-linear data to be transformed for the prediction of the event [17]. Support vector machine was developed in 1990s and became popular among the classification model to its better learning generalization [18]. Signal attenuation and distortion in Television satellites due to the rainfall was observed and flooding was mapped. It has been analyzed that Ku band frequencies varied due to rain fall and climate change. Simulated maps of flash floods were compared with the existing mapping methods to validate the approach [22].

Table I demonstrates the tabular chart for the comparison of AI based algorithms. Root mean square error has been a yardstick to estimate the performance of the algorithms. The results of actual run time data using existing approach of MLP-PSO show probability of 95.15%. The proposed algorithm of ANN-PSO has performed the investigation of flash flood with 0.0047 error probability and enhanced accuracy [20].

TABLE. I. COMPARATIVE ANALYSIS OF VARIOUS APPROACHED FOR THE FLASH FLOOD INVESTIGATION [20]

Performance Indices	ANN	SVM	ANFIS	NNARX	ANN-PSO
RMSE	0.194	0.390	0.116	0.090	0.0047
Best Fit	73	64	78	80.10	98.7
Results	Satisfactory	Unsatisfactory	Satisfactory	Satisfactory	Satisfactory
Hourly Data	6 hrs	6 hrs	3 hrs	3 hrs	3 hrs
Accuracy	73	64	78	80.16	98.99
Precision	Medium	Low	High	High	High
Reliability	Medium	Low	High	High	High
Power Utilixation	Limited	Limited	Limited	Limited	Limited

II. PROBLEM STATEMENT

It has been highlighted in the literature review that sensors and transducers produce false alarms. Errors are generated usually in instrumentation and measurement [10]. Sometime prediction of flash floods can be wrong due to the incompetent decision algorithms and poor sensitivity of sensors. Due to the increased number of false alarms a competent and vigorous classifier and regression model was required for the discrimination of true positive vale and false positive value. Sensors data values may contain false alarms and missed values [21].

III. MATERIALS AND METHODS

Multi-modal sensing device was developed to collect the data from any sea shore. The data has been collected from the sea shore of Kund Malir, Pakistan. Selection of transducers was not an easy task as appropriate combination of competent sensors was needed. According to the literature review almost all the parameters have been used for the flash floods investigation.

A. *Multi-Resolution Sensing Device for Data Collection*

Fig. 1 shows that a device has been developed for the investigation of flash flood as a hazard monitoring device. Multi-modal sensing device comprised of the following sensors: (a) pressure (b) temperature (c) water level (d) gas sensor for detecting CO_2 and (e) ultrasonic sensor. Selection of sensors was very complex task as bunch of appropriate sensors must be used for the accurate and precise results without any false alarm rate [21].

B. *Fundamental Flow Diagram*

In Fig. 2, Ultrasonic sensor, Passive infrared sensor, MQ2 sensor, humidity sensor, pressure and temperature sensors have been used to measure the data near the sea shore of Kund Malir, Pakistan. Data labeling was performed. The data may contain random, missed and repetitive values it must be filtered or normalized before the processing. Therefore, robust classification and regression model were needed.

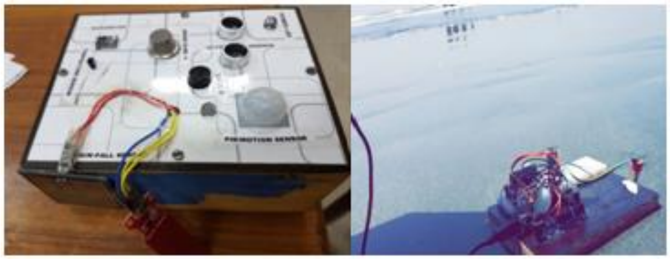


Fig. 1. Multi-Resolution Data Collection Device.

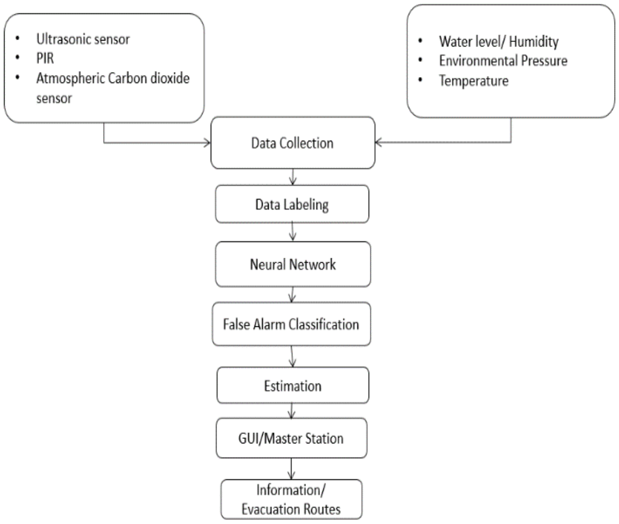


Fig. 2. Main Flow Diagram.

C. *Collected Data Set*

Table II shows that the data set has been collected and all the values have been saved in parallel at the same time. All the sensors have been correlated to each other. More than ten thousand number of instances were recorded and observed for the data processing having total six attributes. Six attributes are the different sensor data values that have been represented in the Table I.

D. *Threshold*

Table III represents the threshold and range related to the sensor values. All of the sensors must exceed the threshold in order to activate the trigger function. Moreover, all the sensors have been interfaced to each other. Sensors produced random results due to the high wind speed. Change of venue for data collection requires modification in threshold of sensors depending on the meteorology of location.

TABLE. II. COLLECTED DATA SET FROM THE MULTI-MODAL SENSING DEVICE [21]

PIR	Distance (mm)	Rainfall	CO$_2$ (ppm)	Temperature (°C)	Pressure (hectopascal)
0	300	0	0	30.13	100270.6
0	37.42	458	539	29.95	100268.4
0	37.18	459	533	29.71	100277.5
0	3661.34	453	418	29.6	100279.8
0	3654.66	453	397	29.46	100277.5
0	3691.18	450	356	37	100280.1
0	3691.18	450	356	29.4	100280.1
1	4.51	504	355	28	110424.2
1	4.39	494	412	23	110424.2
1	4.96	485	399	23	110424.2
1	4.84	479	382	23	100283.5
1	5.06	476	349	23	110424.2
1	4.66	483	400	23	110424.2
1	4.66	478	387	23	110424.2
1	32.45	551	392	23	110424.2
0	22.45	512	387	23	110424.2
0	29.64	518	388	23	110424.2
1	3619.3	529	389	23	110424.2
1	78.39	498	386	23	110424.2
0	0	491	385	23	110424.2
0	30.13	517	397	23	110424.2
0	36.96	501	395	23	110424.2
0	31.92	496	392	23	110424.2
0	32.95	497	390	23	110424.2
0	41.18	503	387	23	110424.2
0	135.14	494	385	23	110424.2
0	3846.76	496	385	23	110424.2
0	6.47	519	393	56	110424.2
0	104.96	506	389	23	110424.2
0	103.76	498	387	23	110424.2
0	31.2	488	385	25	110424.2
0	104.63	486	386	25	110424.2
0	34.51	535	389	25	110424.2
1	6.72	530	1014	25	110424.2
1	20.91	527	1012	25	110424.2
0	101.85	511	1010	25	110424.2
0	3749.04	499	347	25	110424.2
0	36.96	493	303	25	110424.2
0	9.3	492	324	25	110424.2
0	0	493	388	25	110424.2
0	110.74	492	363	25	110424.2
0	105.54	495	341	25	110424.2
0	23.86	496	392	25	110424.2
0	99.76	498	1012	25	110424.2
0	3279.56	502	1012	25	110424.2
0	75.99	510	1012	25	110424.2
0	3259.63	536	1010	25	110424.2
0	37.8	534	1010	25	110424.2
0	3279.9	522	1010	25	110424.2
0	35.02	517	1011	25	110424.2
0	3223.39	516	1011	25	110424.2
0	30.96	543	1010	25	110424.2
0	3244.81	557	1009	25	110424.2
0	32.29	571	1009	25	110424.2
0	3233.67	553	1010	25	110424.2
0	35.19	530	1010	23	110424.2
0	35.79	521	1010	23	110424.2
0	3300.48	522	1010	23	110424.2
0	29.72	549	1009	23	110424.2

TABLE. III. THRESHOLD VALUES FOR SENSORS

	MINIMUM LIMIT	MAXIMUM LIMIT
PIR	1	
DISTANCE	0	50
RAINFALL	>300	
CO2	> 600	
TEMPERATURE	0	50
PRESSURE	>5000	
ALTIMETER	>1000	

IV. Extensive Parametric Comparison of Data Classifiers

Table IV demonstrated that Comprehensive parametric comparison has been performed to investigate the classification accuracy for the robust classification of false alarm in predicting flash floods. Linear discriminant analysis, logistic regression, quadratic support vector machine, k-nearest neighbor and Ensemble bagged tree have been applied to the collected data set for the data classification. Initially seventy-five percent of the data were used as a training and other 25 percent data was saved for the testing purpose. Both of the data files training and testing were converted into the variable so that it may utilized in the MATLAB as all the simulations have been performed in the MATLAB. MATLAB based simulations produced the confusion matrix and all the parametric results which have been presented in the table. The worst accuracy of the classification (62%) has been achieved by the coarse k-NN classifier that means coarse k-NN produced 38% false negative rate that is not acceptable in the case of forecasting. Ensemble bagged trees produced best classification results as it achieved 99 % accuracy and 1% error rate.

A. Linear Discriminant Analysis

Fig. 3 explains that data set was trained for the linear discriminant classification and it produced confusion matrix and other results. Confusion matrix showed that LDA achieved 89% true positive rate and 11 false negative rate with 97% accuracy. Prediction speed and training time was found to be 12000 observation/seconds and 1.5238 seconds respectively. This classification model was up to the mark but accuracy can be further improved. Data can be regularized in discriminant analysis classifier for the robust classification model.

B. Logistic Regression Classification Model

Fig. 4 represents that Logistic regression classification model was developed in the MATLAB using the collected data set. The classification model achieved 96.4% accuracy. Prediction time and training time was found to be 15000 observation/second and 3.9633 seconds. It took almost double time to classify the faulty data compared to the linear discriminant analysis with slight less accuracy.

C. Quadratic Support Vector Machine Classification Model

Fig. 5 illustrates that Quadratic support vector machine classifier has been applied to the collected data set form the sea shore of Kund Malir. The model is based on predict SVM model in which predictors have been defined in the matrix and then comparison would be performed between the observed and predicted. The trained Quadratic SVM model may be compact or full. The trained model has been exported for the testing purpose. yout = predict(QSVMModel,X). Quadratic support vector machine classifier model achieved 93% true positive rate and 7% False negative rate in the confusion matrix. 96.8% accuracy achieved by Q-SVM in 0.9237 seconds of training time. Prediction speed was found to be 40000 observations per second.

TABLE. IV. PARAMETRIC COMPARISON OF LDA, LR, QSVM, K-NN AND ENSEMBLE

Classification Models	True Positive rate (%)	False Negative rate (%)	Accuracy (%)	Prediction speed per second	Training Time (s)
Linear Discriminant	89	11	97.0	12000	1.5238
Logistic Regression	93	7	96.4	15000	3.9633
Quadratic SVM	93	7	96.8	40000	0.9327
Fine k-NN	96	4	98.6	9600	1.5307
Medium k-NN	91	9	96.6	19000	0.927
Coarse k-NN	62	38	89.3	16000	0.90243
Ensemble Bagged Trees	99	1	99.4	2800	6.4101

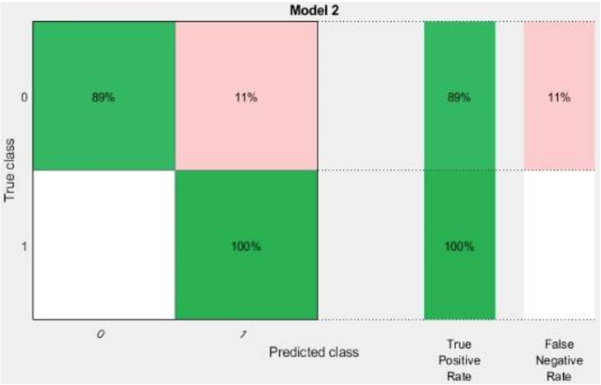


Fig. 3. Linear Discriminant Analysis Confusion Matrix.

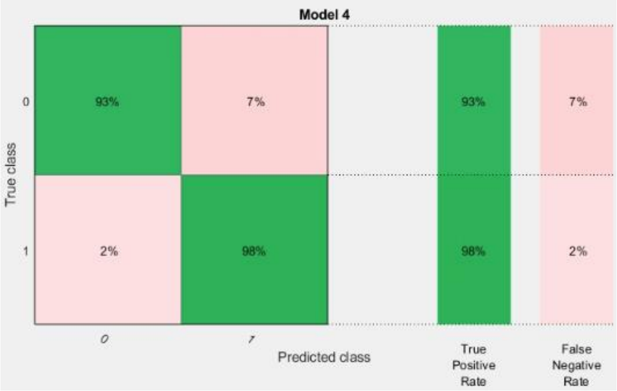


Fig. 4. Logistic Regression Confusion Matrix.

D. Fine k-Nearest Neighbor Classification Model

Fig. 6 shows the confusion matrix for the Fine k-nearest neighbor algorithm. Fine k-nearest neighbor classification model has been developed by measuring the standardized Euclidian distance.

$$E=(u_u−V_v)B^{−1}(u_u−V_v) \quad (1)$$

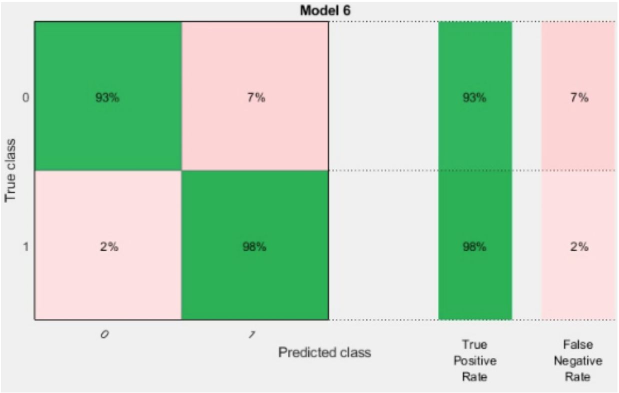


Fig. 5. Q-SVM Confusion Matrix.

Where, U and V are the data matrix. B can be considered as the diagonal matrix and S can be acknowledged as scaling factor. Model was trained in the MATLAB using collected data set. 96% true positive rate and 4% of false negative rate was achieved in this classification model. The fine k-NN classification model achieved 98.6% accuracy with training time of 1.5307 second and prediction speed of 9600 observations per second.

E. Medium k-Nearest Neighbor Classification Model

Fig. 7 explains that medium k-NN classifiers have been used widely as a bench mark for learning. k-NN classifier has the capability to be modified easily as well. X can be considered as the numeric data for the training. Moreover, Chebychev distance, Euclidean distance, city block distance and Minkwoski distance can be measured for determining the *kd*-Tree. Medium k-nearest neighbor classification model produced 19% true positive rate and 9% false negative rate. 96.6% accuracy was achieved with training time of 0.927 second and 19000 observations per second.

F. Coarse k-Nearest Neighbor Classification Model

In Fig. 8, the worst accuracy of the classification (62%) has been achieved by the coarse k-NN classifier that means coarse k-NN produced 38% false negative rate that is not acceptable in the case of forecasting.

G. Ensemble Bagged Trees Classification Model

Fig. 9 represented the confusion matrix of Ensemble Bagged Trees. The Ensemble bagged trees has been applied to the data set for the better decision. The motive of the ensemble is to construct a better learning model which can produce better classification performance on the applied data set. `Yfit = ensemblebaggedtree(B,X)` returns a vector of forecasted responses for the predictor data in the table or matrix `X`, based on the ensemble of bagged decision trees `B`. `Yfit` is a cell array of character vectors for classification and a numeric array for regression. By default, `predict` takes a democratic (non-weighted) average vote from all trees in the ensemble. X can be considered as a numeric Matrix. Ensemble Bagged Tree learning model for classification achieved best performance accuracy of 99.4% with prediction speed of 2800 observations/second and training time is 6.4101 seconds. 99% True positive rate and 1% false negative rate was achieved by the Ensemble Bagged Trees.

H. Graphical Illustration of Model

Fig. 10 displays the Graphical analysis of ensemble bagged tree prediction model has been represented in figure no. 9. Blue color shows the output value of "zero" and orange color represents "one". It can be easily observed from the results that Ensemble bagged tree performed better than the other existing classifiers.

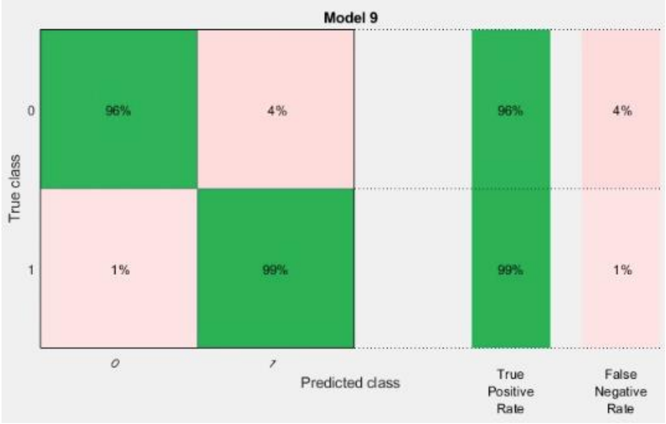


Fig. 6. Fine k-Nearest Neighbor Confusion Matrix.

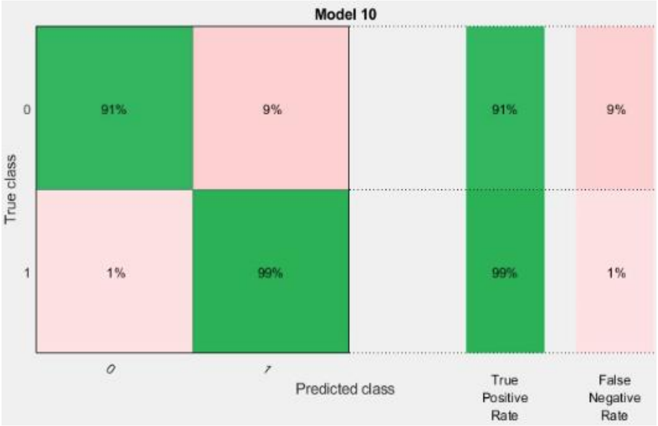


Fig. 7. Medium k-NN Confusion Matrix.

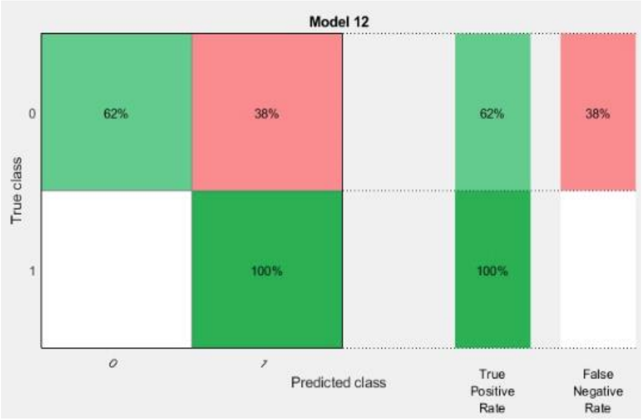


Fig. 8. Coarse k-NN Confusion Matrix.

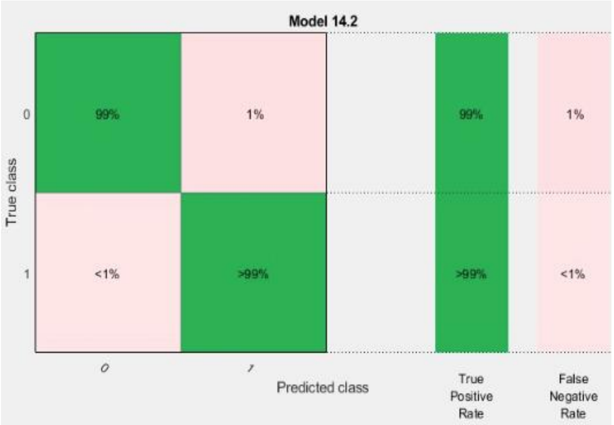


Fig. 9. Ensemble Bagged Trees Confusion Matrix.

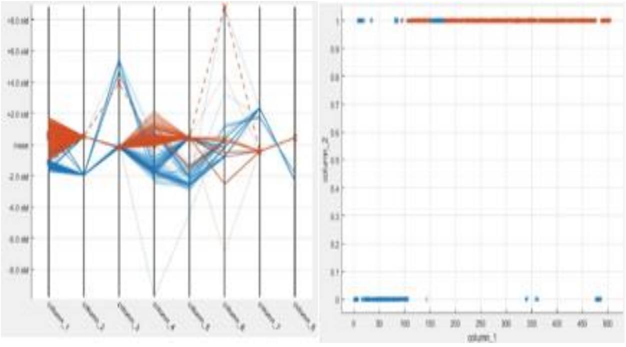


Fig. 10. Graphical Analysis of Prediction Model.

V. Exhaustive Parametric Comparison of Regression Models

Table V shows the comprehensive parametric comparison for finding out the best performer regression model. LR, interactions linear, robust linear, step wise linear, linear support vector machine, Quadratic SVM, Gaussian SVM, Rational quadratic GPR, Exponential GPR and ensemble bagged trees regression models have been developed for the prediction of flash floods. These regression models have been applied on the collected data set. According to the comprehensive parametric comparison of regression models Quadratic SVM found to be the worst with mean square error of 0.5551 and time elapsed 13.159 seconds. On the other hand, Exponential Gaussian process regression performed better than the other existing approaches with the minimum root mean squared error of 0.0002 and prediction speed of 35000 observations per second.

A. *Graphical Illustration of Linear Regression, Interactions Linear, Robust Linear and Step Wise linear*

Fig. 11 demonstrates the graphical illustration of model 1 (Linear Regression), Model 2 (Interactions Linear), Model 3 (Robust Linear) and Model 4 (step wise linear). Moreover, Blue color represents the true data and yellow color depicts the predicted data. The graphs have been plotted between number of records and flood response (hurricane response).

TABLE. V. Parametric Comparison of Regression Approaches

Predictive Models	RMSE	R^2	MSE	MAE	Prediction speed per second	Training Time (s)
Linear Regression	0.1900	0.82	0.0361	0.113	11000	4.5756
Interactions Linear	0.1087	0.94	0.0118	0.044	39000	0.8332
Robust Linear	0.267	0.65	0.0712	0.071	80000	0.898
Step Wise Linear	0.1118	0.94	0.0125	0.046	70000	3.779
Linear SVM	0.2316	0.73	0.0536	0.127	47000	11
Quadratic SVM	0.5551	-0.53	0.0308	0.322	73000	13.159
Gaussian SVM	0.0721	0.97	0.0052	0.045	140000	0.818
Rational Quadratic GPR	0.0006	1	3×10^{-7}	9×10^{-5}	22000	1.1555
Exponential GPR	0.0002	1	4×10^{-8}	3×10^{-5}	35000	2.1752
Ensemble Bagged Trees	0.0750	0.97	0.0056	0.0222	34000	2..3892

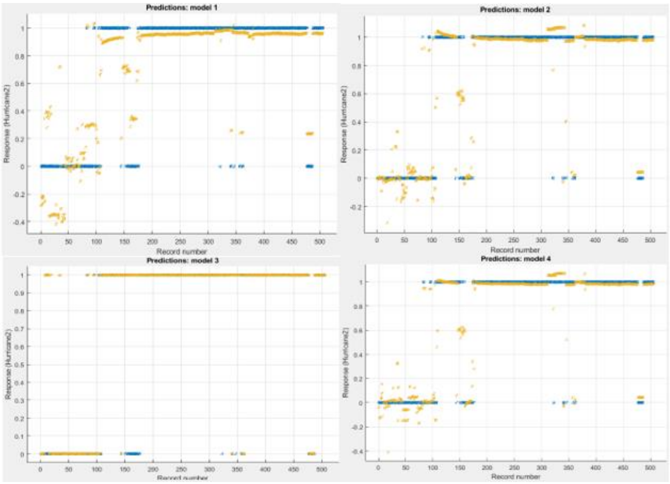


Fig. 11. Graphical Illustration of Prediction Model 1, Model 2, Model 3, Model 4.

B. *Graphical Illustration of Linear SVM, Quadratic SVM, Gaussian SVM and Rational Quadratic GPR*

Fig. 12 demonstrated the graphical illustration of model 5 (Linear SVM), Model 6 (Quadratic SVM), Model 7 (Gaussian SVM) and Model 8 (Quadratic GPR). Moreover, Blue color represents the true data and yellow color depicts the predicted data. The graphs have been plotted between number of records and flood response (hurricane response).

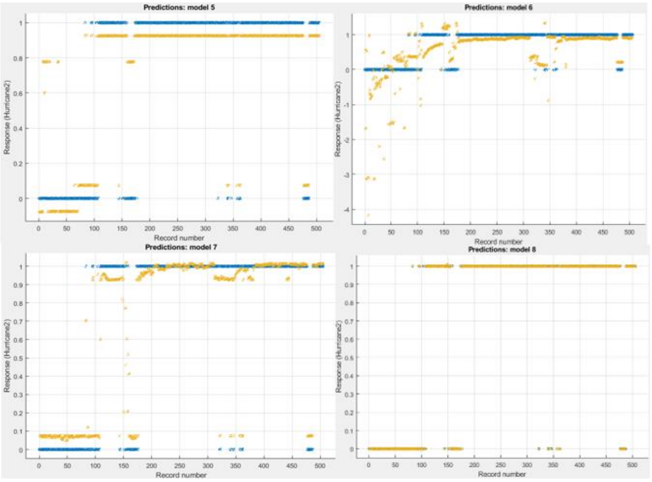


Fig. 12. Graphical Illustration of Prediction Model 5, Model 6, Model 7, Model 8.

C. Graphical Illustration of Exponential GPR and Ensembled Bagged Trees

Fig. 13 portrayed the graphical analysis of model 9 (Exponential GPR) and model 10 (Ensemble Bagged Trees). Moreover, Blue color represents the true data and yellow color depicts the predicted data. The graphs have been plotted between number of records and flood response (hurricane response). (Gaussian Processes). For set S, mean function μ : S 7→ R and any covariance function (also called kernel) k : S×S 7→ R, there exists a GP f(x) on S,

s.t. E[f(x)] = μ(x), Cov(f(xs), f(xt)) = k(xs, xt), x, xs, xt ∈ S (2)

It denotes f ∼ GP(μ, k). For a regression problem y = f(x) + ε, by Gaussian process method the unknown function f is assumed to follow a GP(μ, k). Given n pairs of observations (x1, y1), . . . ,(xn, yn),

we have

y = f(X) + ε (3)

where y = [y1, y2, . . . , yn] T are the outputs, X = [x1, x2, . . . , xn] T are the inputs, and ε = [ε1, ε2, . . . , εn] T are independent identically distributed Gaussian noise with mean 0 and variance σ 2 n [19].

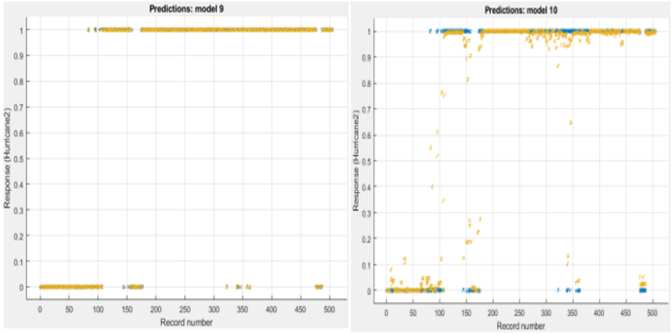


Fig. 13. Graphical Analysis of Prediction Model 9 and Model 10.

VI. Conclusion and Future Work

In this research paper various classification model and regression model have been investigated to determine the competent approach for the identification of flash floods with less false alarm rates. Parametric comparison has been shown so that researchers may have an idea about the technical and practical implications related to the flash flood research. Actually the collected data or observed data may contain inconsistent, missed or random values. Therefore, vigorous classification and predictive model is required for the accurate and precise identification of the flash floods. Moreover, Bayesian regularization and scaled conjugate gradient were applied to the data set. Results showed that scaled conjugate gradient performed better. Comprehensive parametric comparison of classifiers and regression models has been performed and compare to find out the better classification and regression model. It can be concluded that the worst accuracy of the classification (62%) has been achieved by the coarse k-NN classifier that means coarse k-NN produced 38% false negative rate that is not acceptable in the case of forecasting. Ensemble bagged trees produced best classification results as it achieved 99 % accuracy and 1% error rate. According to the comprehensive parametric comparison of regression models Quadratic SVM found to be the worst with mean square error of 0.5551 and time elapsed 13.159 seconds. On the other hand, Exponential Gaussian process regression performed better than the other existing approaches with the minimum root mean squared error of 0.0002 and prediction speed of 35000 observations per second.

Acknowledgment

Prof. Dr. Mohammad Alam’s (Universiti Kuala Lumpur, Institute of Business Managament) has suggested me how to develop multi-modal sensing gadget for the collection of data. Authors are extremely thankful to Dr. Muhammad Alam as he recommended some competent classifier algorithms as his expertise are in Artificial Intelligence. Dr. Kushsairy Kadir (UniKL-BMI) and Dr. Zeeshan Shahid (IoBM) were found to be true mentor as they always mentored and improved proposed research. Prof. Dr. Mazliham Mohd Su’ud provided the latest research methodology and research synthesis approaches which helped us to design a novel technique for the prior investigation of flash floods.

References

[1] O. Intharasombat and P. Khoenkaw, "A low-cost flash flood monitoring system," 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE), Chiang Mai, 2015, pp. 476-479.

[2] K. Achawakorn, K. Raksa and N. Kongkalai, "Flash flood warning system using SCADA system: Laboratory level," 2014 International Electrical Engineering Congress (iEECON), Chonburi, 2014, pp. 1-4.

[3] P. Rapant et al., "Early warning of flash floods based on the weather radar," Proceedings of the 2015 16th International Carpathian Control Conference (ICCC), Szilvasvarad, 2015, pp. 426-430.

[4] S. Muthukumar, W. S. Marry, S. Ajithkumar, M. Arivumathi and V. Sowndharya, "Network based flash flood alert system," 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, 2018, pp. 239-241.

[5] T. Khan, K. Kadir, M. Alcm, Z. Fchiihid and M. S. Mazliham, "Geomagnetic field measurement at earth surface: Flash flood forecasting using tesla meter," 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T), Kuala Lumpur, 2017, pp. 1-4.

[6] F. A. Ruslan, A. M. Samad and R. Adnan, "Modelling of flood prediction system using hybrid NNARX and Extended Kalman Filter," 2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA), Batu Ferringhi, 2017, pp. 149-152.

[7] T. A. Khan, M. Alam, K. Kadir, Z. Shahid and S. M Mazliham, "A Novel Approach for the Investigation of Flash Floods using Soil Flux and CO2: An Implementation of MLP with Less False Alarm Rate," 2018 2nd International Conference on Smart Sensors and Application (ICSSA), Kuching, 2018, pp. 130-134.

[8] M. Hijji, S. Amin, R. Iqbal and W. Harrop, "A Critical Evaluation of the Rational Need for an IT Management System for Flash Flood Events in Jeddah, Saudi Arabia," 2013 Sixth International Conference on Developments in eSystems Engineering, Abu Dhabi, 2013, pp. 209-214.

[9] M. Mousa, X. Zhang and C. Claudel, "Flash Flood Detection in Urban Cities Using Ultrasonic and Infrared Sensors," in IEEE Sensors Journal, vol. 16, no. 19, pp. 7204-7216, Oct.1, 2016. doi:10.1109/JSEN.2016. 2592359.

[10] T. A. Khan, M. Alam, Z. Shahid and M. M. Suud, "Prior investigation for flash floods and hurricanes, concise capsulization of hydrological technologies and instrumentation: A survey," 2017 IEEE 3rd International Conference on Engineering Technologies and Social Sciences (ICETSS), Bangkok, 2017, pp. 1-6.

[11] N. Bhardwaj, N. Aggarwal, N. Ahlawat and C. Rana, "Controls and intelligence behind “NISTARA-2”—A disaster management machine (DMM)," 2014 Innovative Applications of Computational Intelligence on Power, Energy and Controls with their impact on Humanity (CIPECH), Ghaziabad, 2014, pp. 34-37.

[12] Kotsiantis S.B., Tsekouras G.E., Pintelas P.E. (2005) Bagging Model Trees for Classification Problems. In: Bozanis P., Houstis E.N. (eds) Advances in Informatics. PCI 2005. Lecture Notes in Computer Science, vol 3746. Springer, Berlin, Heidelberg.

[13] L. S. Solanki, S. Singh, and D. Singh, "An ANN approach for false alarm detection in microwave breast cancer detection," in 2016 IEEE Congress on Evolutionary Computation (CEC), 2016, pp. 1370-1374.

[14] O Daniell, Y Petillot, and S Reed., “Unsupervised seafloor classification for automatic target recognition”. Proc. International Conf. Remote Sens., (October), 2012.

[15] DP Williams and E Fakiris. Exploiting environmental information for improved underwater target classification in sonar imagery. IEEE Trans. Geosci. Remote Sens., 52(10):6284–6297, 2013.

[16] C. E. Rasmussen, Evaluation of Gaussian processes and other methods for non-linear regression, University of Toronto, 1999.

[17] Suresh, P. V. S., Venkateswara Rao, P. and Deshmukh, S. G., “A Genetic Algorithmic Approach for Optimization of Surface Roughness Prediction Model,” International Journal of Machine Tools and Manufacture, Vol. 42, 2002, pp. 675-680.

[18] Yujun Yang, Jianping Li and Yimei Yang, "The research of the fast SVM classifier method," 2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, 2015, pp. 121-124.

[19] C. E. Rasmussen, C. K. Williams, Gaussian processes for machine learning, Vol. 1, MIT press Cambridge, 2006.

[20] Talha Khan, Muhammad Alam, Faraz Shaikh, Sheroz Khan, Kushsairy Kadir, Zeeshan Shahid, M.S Mazliham & Yahya,” Flash floods prediction using real time data: An implementation of ANN-PSO with less false alarm”, 2019 IEEE International Instrumentation & Measurement Technology Conference 20-23 May 2019, Grand Millenium Auckland, New Zealand.

[21] T. Khan, M. Alam, and M. Mazliham, “Artificial Intelligence Based Multi-modal Sensing for Flash Flood Investigation”, jictra, pp. 40-47, Jun. 2018.

[22] F. Mercier, N. Akrour, L. Barthès, C. Mallet and R. Hallali, "Fine-scale evaluation of rainfall from TV-sats: A new method for water monitoring and flash flood prevention," in URSI Radio Science Bulletin, vol. 2017, no. 360, pp. 80-88, March 2017.

Authentication Modeling with Five Generic Processes

Sabah Al-Fedaghi[1], MennatAllah Bayoumi[2]
Computer Engineering Department
Kuwait University, Kuwait

***Abstract*—Conceptual modeling is an essential tool in many fields of study, including security specification in information technology systems. As a model, it restricts access to resources and identifies possible threats to the system. We claim that current modeling languages (e.g., Unified Modeling Language, Business Process Model and Notation) lack the notion of *genericity*, which refers to a limited set of elementary processes. This paper proposes five generic processes for modeling the structural behavior of a system: creating, releasing, transferring, receiving, and processing. The paper demonstrates these processes within the context of public key infrastructure, biometric, and multifactor authentication. The results indicate that the proposed generic processes are sufficient to represent these authentication schemes.**

Keywords—Security; authentication; conceptual modeling; diagrammatic representation; generic processes

I. Introduction

Security is a necessary feature in information technology (IT) systems. Security specification requires identifying risks, access requirements, and recovery strategies, and comprises well-developed security mechanism processes [1]. Early-stage development of security specification assists in lowering the possibility of security breaches.

Authorization and authentication both play vital roles in the configuration of security mechanisms. Authorization is the process of allowing users to access system objects based on their identities. Authentication confirms that the user is who he or she claims to be.

Conceptual modeling is a description of reality using a modeling language to create a more-or-less formalized schema [2]. A conceptual model in the security field restricts access to the resources and identifies possible threats to the system. In modeling, *notations* (diagrams, symbols, or abbreviated expressions) are required to specify technical facts and related concepts of systems. They are necessary to articulate complex ideas succinctly and precisely [3]. For a notation to convey accurate communication, it must effectively represent the different aspects of a system and be well understood among project participants. The historic roots of modeling notations in software engineering can be traced back to structured analysis and design, which are based on data flow diagrams [3].

A. Security Modeling

Many languages and mechanisms, such as Business Process Model and Notation (BPMN) [4], secure Tropos [5], misuse cases [6], and mal-activity diagrams [7], are used in the field of security modeling. For space consideration, we focus here on the Unified Modeling Language (UML) and BPMN.

The UML [8] has been utilized as a graphical notation to construct and visualize security aspects in object-oriented systems. It is currently utilized as a primary notation for security and authentication because it provides a spectrum of notations representing the various aspects of a system. The use of the UML for conceptual modeling requires special care to not confuse software features with aspects of the real world being modeled [9].

BPMN was designed to be used by people without much training in software development. "UML diagrams look technical, and in practice, they are much harder for businesspeople to understand than BPMN diagrams" [10]. BPMN includes a rich set of model constructs for business process modeling.

This paper is about conceptual modeling. It is part of a research project that applies a new modeling language, the thing machine (TM), to modeling computer attacks [11]. The paper concentrates on using the TM to model authentication. The thesis promoted in our research works is that modeling in the abovementioned languages lacks *genericity*, a notion for representing systems that forms the base for process modeling. This has caused conceptual vagueness that obstructs the differentiation of objects. A specific goal of the paper is to substantiate the viability of the TM by applying it to modeling authentication.

B. Modeling Authentication

In the twenty-first century, few matters are more pressing than those related to identity authentication. Authentication is a mechanism used to make sure that those obtaining session access are who they say they are. To access online systems and services, we all face the challenge of proving our identities [12].

In the real world, thousands have found themselves blocked from opening bank accounts, making payments, or travelling because of an unfortunate name similarity to those individuals or entities on a sanctions list. Hundreds of thousands have been victims of identity fraud, often only learning of the crime when they apply for credit and find their credit rating has been compromised by fraudulent loans obtained in their names [12].

In this paper, we focus on individual and entity authentication for digital interactions. We concentrate on authentication in the context of usability of IT systems in terms of who is using the system, what they are using it for, and the environment in which they are using it (ISO standard 9241 Part 11). The ISO 9241 standard for identity authentication is made up of three components: what you are (e.g., biometric information), what you have (e.g., having a token), and what you know (e.g., PINs, passwords).

C. Examples of Modeling Authentication

Fig. 1 shows a typical authentication process—in this case, a single sign-on (SSO) that allows a user to access multiple applications with one set of login credentials. This SSO is modeled using a UML activity diagram. With an SSO, a client accesses multiple resources connected to a local area network [13]. Partner companies act as identity providers and control usernames and other information used to identify and authenticate users for web applications. Each partner provides Google with the URL of its SSO service, as well as the public key that Google will use to verify Security Assertion Markup Language (SAML, a protocol that refers to what is transmitted regarding identity information between parties) responses. When a user attempts to use a hosted Google application, Google generates an SAML authentication request and sends a redirect request back to the user's browser that points to the specific identity provider. The SAML authentication request contains the encoded URL of the Google application that the user is trying to reach. This authentication process continues as shown in part in Fig. 1 [13]. Nevertheless, in general, "activity diagrams have always been poorly integrated, lacked expressiveness, and did not have an adequate semantics in UML" [14]. With further development of the UML, "several new concepts and notations have been introduced, e.g., exceptions, collection values, streams, loops, and so on" [14].

Plavsic and Secerov [15] model the classic login procedure using different kinds of UML diagrams: deployment diagrams, use case diagrams, interaction overview diagrams, sequence diagrams (see Fig. 2), and class diagrams. The code shown in Fig. 2 was generated from the class diagrams. According to Plavsic and Secerov [15], the use case diagram counts as a starting point; however, use cases do not describe system structure or details of behavior. Therefore, Plavsic and Secerov use other diagrams to follow messages that are exchanged between objects and realize system functionality.

Lee [3] gives an example wherein field officers are required to provide authentication before they can use a system called FRIEND. Authentication is modeled as an authenticate-use case. Later, two more use cases are introduced: AuthenticateWithPassword, which enables field officers to login without any specific hardware, and AuthenticateWithCard, which enables field officers to log in using smart cards. The two use cases are represented as specializations of the Authenticate use case (see Fig. 3).

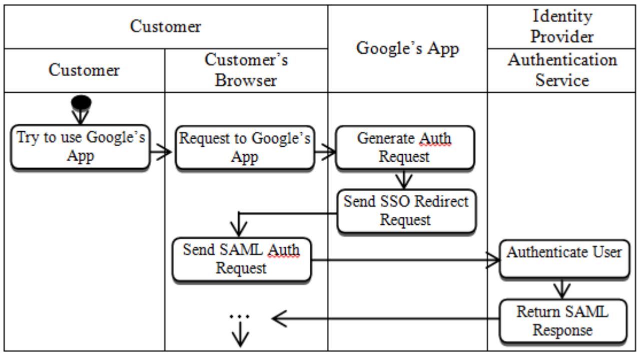


Fig. 1. An Example of a UML Activity Diagram for SSO to Google Apps. (Partially Redrawn from [13]).

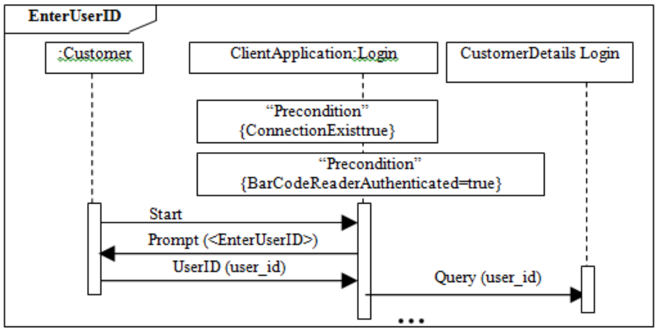


Fig. 2. EnteruserId Sequence Diagram. (Partially Redrawn from [15]).

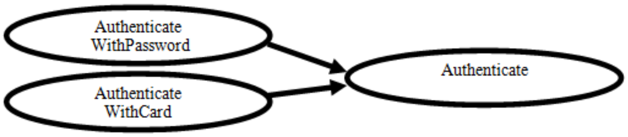


Fig. 3. The Authenticate use Case is a High-Level use Case Describing, in General Terms, the Process of Authentication. AuthenticatewithPassword and AuthenticatewithCard are Two Specializations of Authenticate. (Partially Redrawn from [3]).

In the next section, we will briefly review the TM with a new contribution related to the notion of genericity. In Section 3, we give an example. In Section 4, we apply the TM to model authentication.

II. Thing Machine with Five Generic Processes

We claim that a modeling methodology is based on five generic (a notion to be discussed later) processes: creating, releasing, transferring, receiving, and processing (changing). These elementary processes form a complex abstract machine called a Thing Machine, as shown in Fig. 4. Fig. 5 shows a TM formulated to align with the classical input–process–output model.

The machines constitute a mosaic or network of machines. Additionally, the TM model embraces *memory* and *triggering* (represented as dashed arrows, relations among the processes' stages (machines). A TM manifests structure and behavior simultaneously. Only five elementary processes are used because they represent genericity in operation, the way the three states of water (liquid, vapor, and solid) represent three generic concepts. These elementary processes have been called different names.

- Create: generate, produce, manufacture, give birth, initiate, assemble, emerge, appear (in a system), etc. Process: change, modify, adjust, amend, etc.
- Receive: obtain, accept, collect, take, get, etc.
- Release: allow, relieve, discharge, let, free, etc.
- Transfer: transport, transmit, carry, communicate, etc.

The TM model has been applied to many real systems, such as phone communication [16], physical security [17], vehicle tracking [18], unmanned aerial vehicles [19], and programming [20].

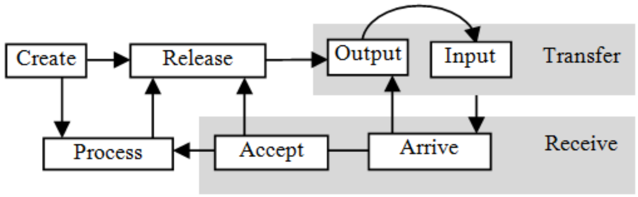


Fig. 4. Thinging Machine.

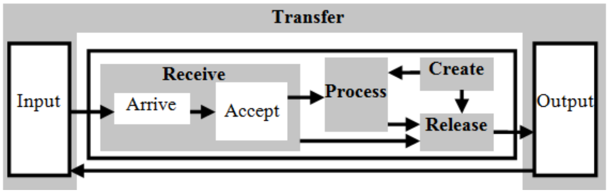


Fig. 5. Another form of Description of a TM.

A. Philosophical Foundation of Things and Thinging

The TM model is constructed on the philosophical foundations of Heidegger's notions of a *thing* and *thinging* [21]. Heidegger's philosophy gives an alternative analysis of "(1) eliciting knowledge of routine activities, (2) capturing knowledge from domain experts and (3) representing organizational reality in authentic ways" [22]. Additional information about the philosophical foundations of the TM can be found in Al-Fedaghi [23-24].

Briefly, in a TM, a thing is defined as that which can be created, processed, released, transferred, and received. It encounters humans through its givenness (Heidegger's term). "In contrast to object-orientation, which represents things as quantifiable objects to be controlled and dominated, Heidegger's definition of a thing encompasses a particular concrete existence along with its interconnectedness to the world" [25]. According to Heidegger [21], thinging expresses how a "thing things", which he explained as gathering or tying together its constituent parts.

A TM operates by creating, processing, receiving, releasing, and transferring things. For example, a *tree* is a machine "through which flows of sunlight, water, carbon dioxide, minerals in the soil, etc. flow. Through a series of operations, the machine transforms those flows of matter, those other machines that pass through it into various sorts of cells" [26]. In the TM approach, a thing is not just an entity; it is also a machine that handles other things.

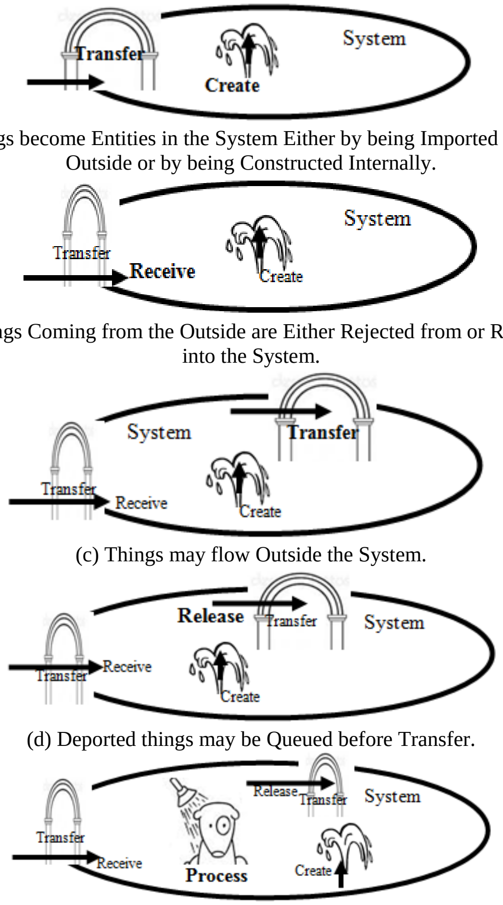
B. Genericity

The TM's five processes are categorical. Members of each category have the following features:

- They focus on essential properties and ignore variations in the created category—for example, newness regardless of who, what, how, etc.
- They capture the blueprint aspect: e.g., creation is a "popping up" phenomenon wherein a thing either "emerges into" the system or as a result of existing things being processed to trigger the creation of other things.
- Things have attributes similar to *objects*—a created thing comes to "life", and a processed (changed) thing remains the same thing.
- Things have actions similar to *subjects* (machines)—creating a thing is "bringing it to life" and processing a thing changes it in some way.

A sketch of the proof of the necessity for the five generic processes can be outlined as shown in Fig. 6. Thus, informal justification for the five TM stages can be specified as follows:

- Things become entities in the system either by being imported from the outside (transfer/input) or by being internally constructed (creation). See Fig. 6(a).
- Things coming from the outside are either rejected from or received (receive) into the system. See Fig. 6(b).
- Things may flow outside the system (transfer/output). See Fig. 6(c).
- Deported things may be queued before transfer (release). See Fig. 6(d).
- Things inside the system may be processed (process). See Fig. 6(e).



(a) Things become Entities in the System Either by being Imported from the Outside or by being Constructed Internally.

(b) Things Coming from the Outside are Either Rejected from or Received into the System.

(c) Things may flow Outside the System.

(d) Deported things may be Queued before Transfer.

(e) Things Inside the System may be Processed.

Fig. 6. Informal Sketch of the Generic Processes.

This is what we mean by *generic* processes. Even though they are used differently according to the setting, members of each generic process seem to be synonymous with respect to *things*. In language, such a phenomenon appears in the case of the adjectives *big*, *great*, and *large*, which are seemingly synonymous words but are likely to be used in different ways in different settings [27]. Processes recognized as being of the same kind of "meaning" in the above sense are said to possess a generic property. Generic processes are conduits through which various types of processes flow.

III. Thing Machine Modeling Example

Guizzardi and Wagner [2] give an example of a service queue system in which customers arrive at random times at a service desk. They have to wait in a queue when the service desk is busy. Otherwise, when the service desk is not busy, they are immediately served by the clerk. Whenever a service is completed, the next customer from the queue (if any) is served [28].

Fig. 7 shows the TM model of the example. The customer arrives (circle 1) to get into the queue (Q). We assume a circular queue structure stored in Q(0:n - 1) with mod n operation; *rear* points to the last item and *front* is one position counterclockwise from the first item in Q. As typically described, the queue has a *rear*, which, upon the arrival of the customer (2), is retrieved/released (3) and incremented (4). Hence:

- If Q is full (the maximum capacity of the queue when *(rear+1)mod n =front*), the system blocks any newly arriving customers.
- The new rear value is stored (6).

Accordingly, the customer is assigned a position (given a number) in the queue and joins the other customers waiting in the queue (8).

Whenever the service agent is *not busy* (9):

- The first customer in the queue is released to the service area (10).
- The arrival of the customer to the service area changes its state to *busy* (11).
- The customer is then processed (12).
- The customer is released (13), which triggers the not busy state (14).
- The customer leaves the service area (15).

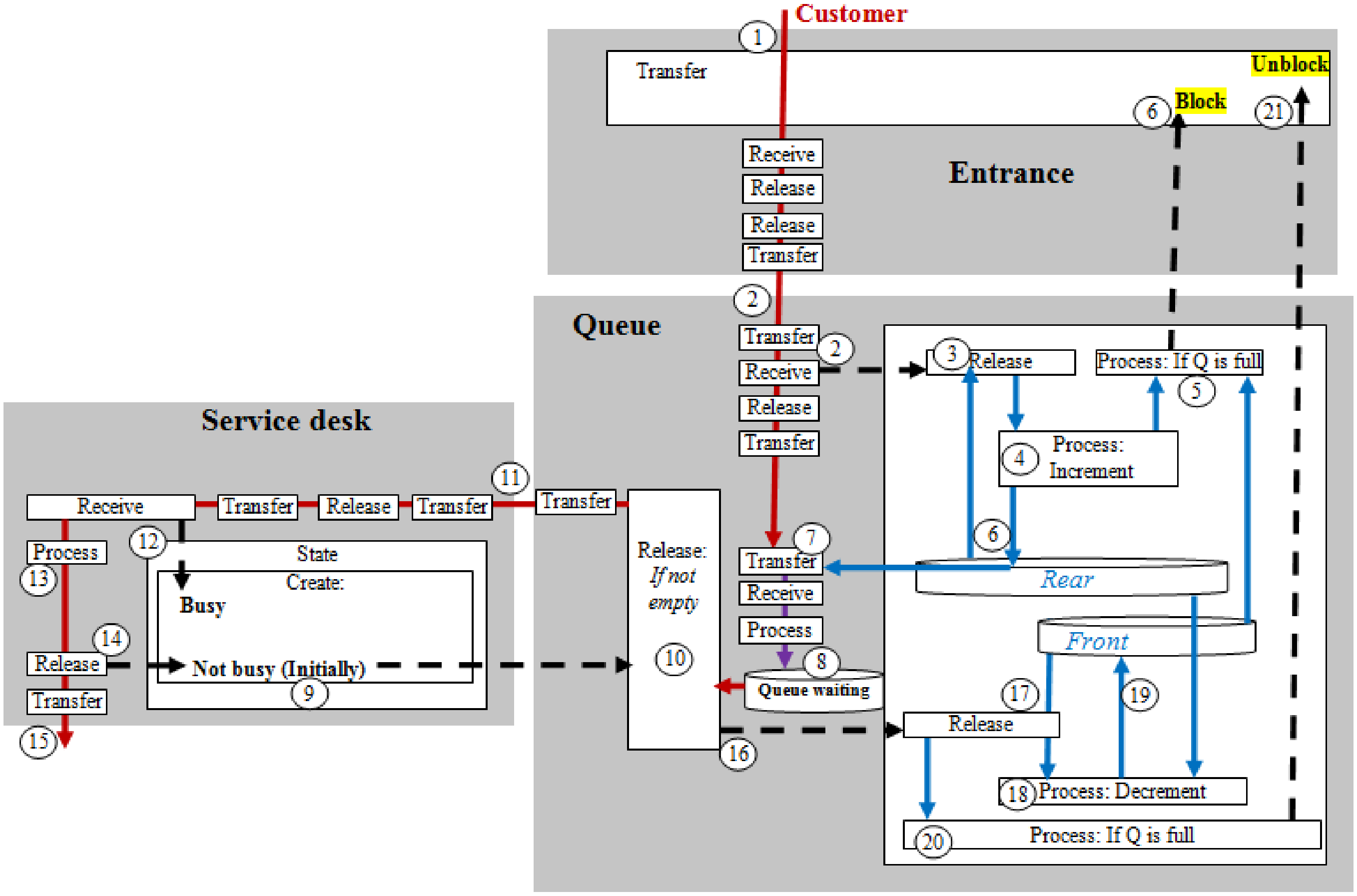


Fig. 7. Static TM Description of the Example.

Triggering the *not busy* state results in taking a new customer from the queue, as mentioned previously (10), and also updates the queue data (16). Thus,

- The front value is retrieved (17) and decremented (18), and the new value is stored (19).
- The original *front* value (before decrementing it) is checked (20), and if the Q was full, the blockage of new customers from entering the queue is lifted (21).

Initially, we assume that the entrance is not blocked, the queue is empty, and the service is not busy.

The dynamic behavior of the system can be developed based on events. An event in a TM is treated as a thing/machine—that is, it can be created, processed, released, transferred, and received. For example, the event a customer moves from the queue to the service desk is represented as shown in Fig. 8. It has two submachines: time and region where the event takes place. An event also denotes a change. All stages in the static description of Fig. 7 indicate elementary changes; however, we are typically interested in larger events that include several stages, as demonstrated in the event a customer moves from the queue to the service. Accordingly, we identify the following events in this example (see Fig. 9):

Event 1 (E_1): The service is open.
Event 2 (E_2): The service is closed (blocked).
Event 3 (E_3): A customer joins the queue.
Event 4 (E_4): Top is retrieved and incremented, and the new value is stored.
Event 5 (E_5): The queue is full (i.e., new value = max).
Event 6 (E_6): The queue is not full.
Event 7 (E_7): A customer joins the queue.
Event 8 (E_8): The service agent is not busy.
Event 9 (E_9): A customer moves from the queue to the service.
Event 10 (E_{10}): The service becomes busy.
Event 11 (E_{11}): The customer leaves the service.
Event 12 (E_{12}): Top is retrieved and decremented, and the new value is stored.
Event 13 (E_{13}): Top becomes less than max.

Fig. 10 shows the behavior of the system in terms of the chronology of its events.

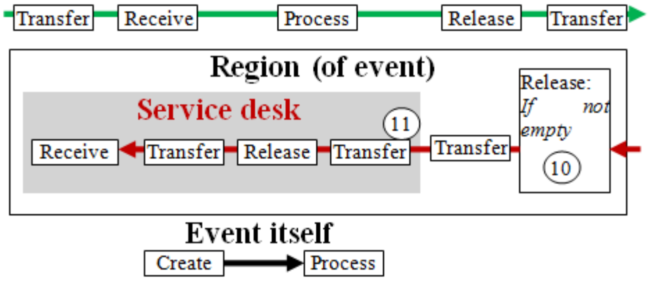


Fig. 8. Event with Region and Time Submachines.

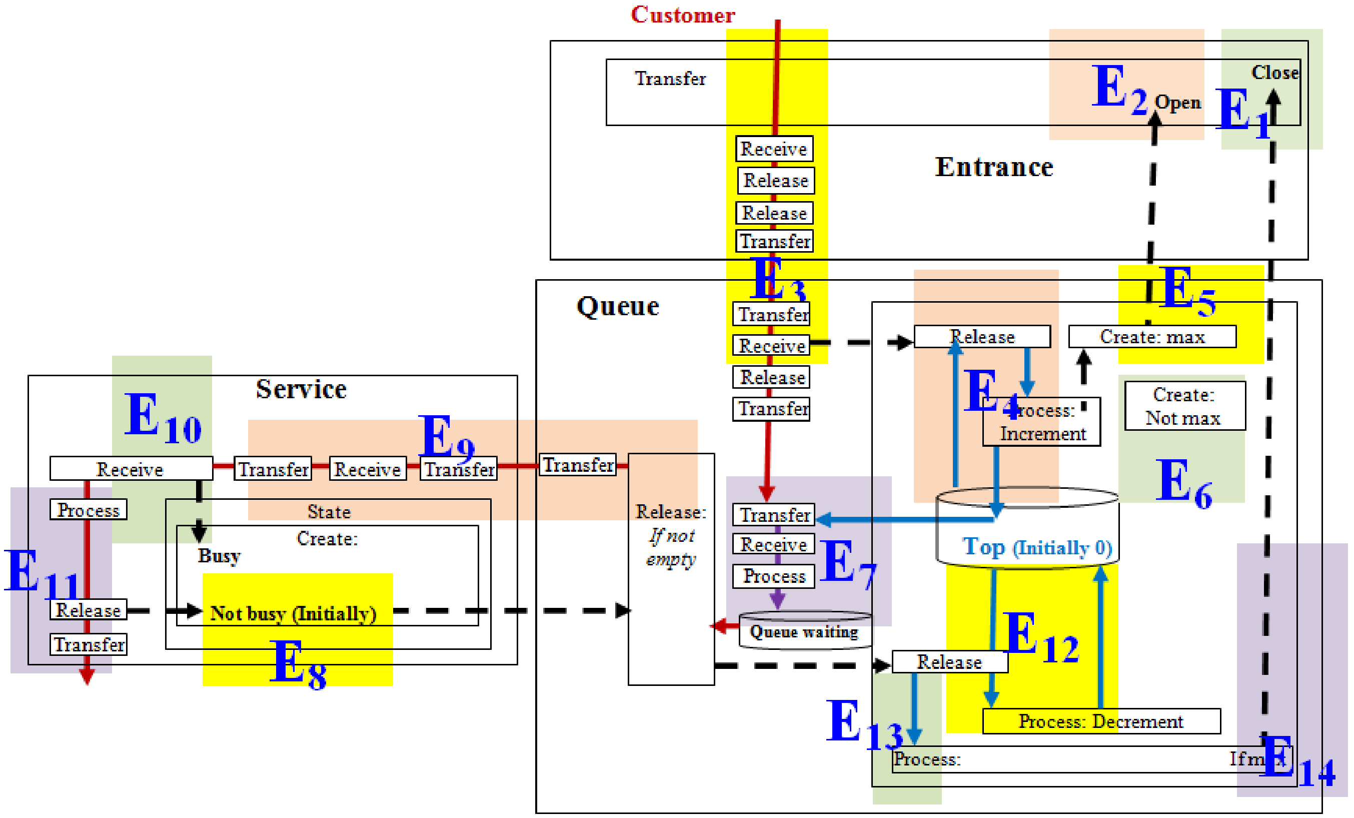


Fig. 9. Identifying the Events in the Static Description of the Example.

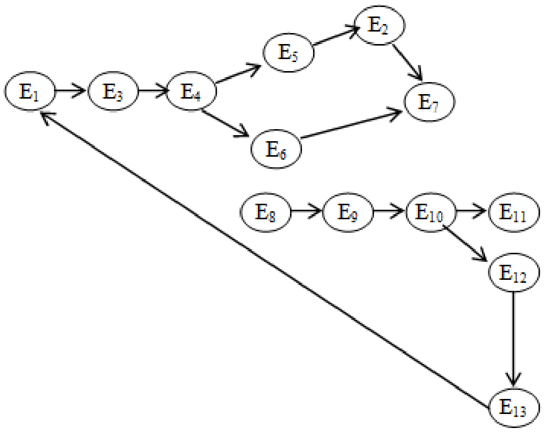


Fig. 10. Chronology of Events.

IV. Case Study: Modeling Authentication

To apply a TM to modeling authentication, we adopt a security case study that involves insider attackers as presented by Nostro et al. [29]. This case study is interesting because it adopts a modeling approach using UML diagrammatic and textual use cases in line with the level of modeling applied in this paper. Additionally, UML use cases give us an opportunity to contrast use case diagrams with TM diagrams.

The case study includes the taxonomy of users physically or logically involved within the system and investigates their roles as potential insiders. The users are system administrator (SA), system expert, unknown user, domain expert, human sensor, and operator. Nostro et al. [29] explore only the SA and system expert, and we, in this paper, focus on the SA performing a *software update*. Fig. 11 shows the use case related to the SA; the darkened part indicates our region of emphasis. Fig. 12 shows the textual description of the use case.

Based on such a use case model that "guides the whole process," Nostro et al. [29] identify and assess insider threats and develop countermeasures that are oriented toward prevention, deterrence, or detection. They also use an ad hoc attack execution graph called ADVISE (see Fig. 13).

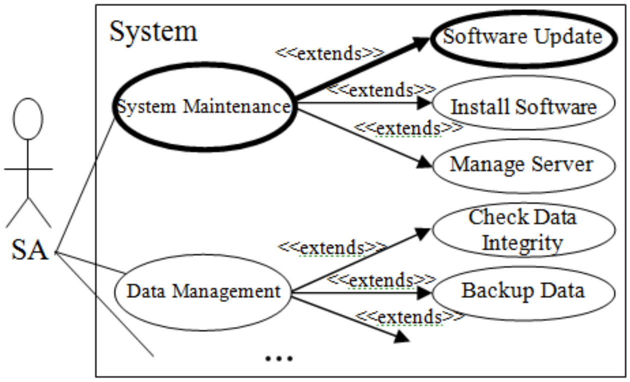
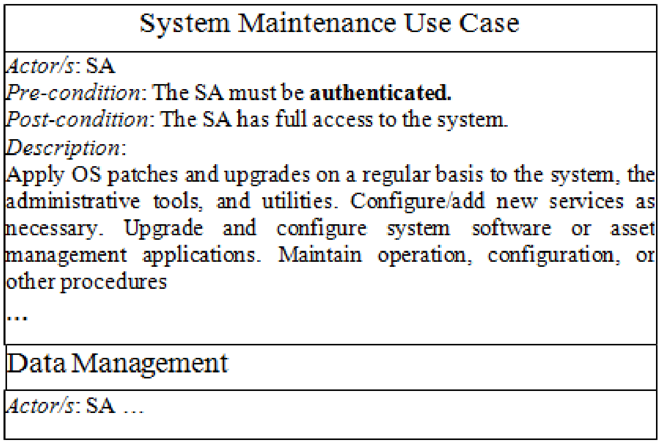


Fig. 11. UML use Case Diagram Involving the SA. (Partially redrawn from [29]).



System Maintenance Use Case

Actor/s: SA
Pre-condition: The SA must be **authenticated.**
Post-condition: The SA has full access to the system.
Description:
Apply OS patches and upgrades on a regular basis to the system, the administrative tools, and utilities. Configure/add new services as necessary. Upgrade and configure system software or asset management applications. Maintain operation, configuration, or other procedures
…

Data Management

Actor/s: SA …

Fig. 12. Description of UML use Case Diagram—SA. (Partially Taken from [29]).

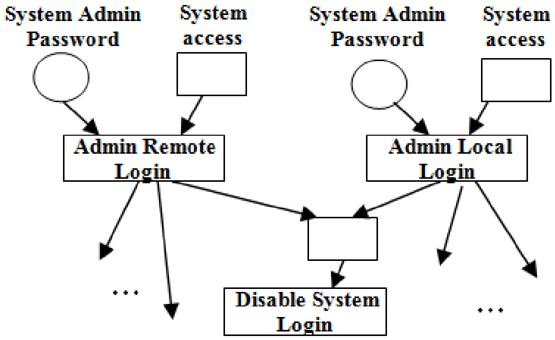


Fig. 13. Sample Attack Execution Graph. (Partially Redrawn from [29]).

We claim in this paper that the TM model presents a systematic alternative (one kind of notion) in modeling security. Without loss of generality, we will focus on the authentication part of Nostro et al. [29] to demonstrate the viability of the TM model.

V. Modeling Authentication

Authentication plays an important role in the security of computing, hence the existence of several authentication techniques. An authentication process attempts to verify a user's identity prior to the user's access to any resources in order to protect the system against various attack types. Once authenticated, the user is permitted to connect with cloud servers to request services [30-33]. Without loss of generality and due to space limitations, we will apply the TM model to only three authentication methods: public key infrastructure (PKI) authentication, biometric authentication, and multifactor authentication. As discussed in the case study in Section IV, the authentication of the SA is a precondition of all four use cases (system maintenance, data management, profile management, and crisis management, as represented in Fig. 11). The login session allows the SA to begin requesting services from the system. However, no requests from any of these four use cases will be serviced until the SA is authenticated by the system.

The first SA role to be investigated is the system maintenance case. This case is an umbrella to three subcases involving software updates, installing software, and managing servers.

A. *Public Key Infrastructure Authentication*

Fig. 14 shows the TM representation of SA roles under the PKI framework system, whereas Fig. 15 shows the corresponding dynamic system, assuming the SA is already certified. The figure comprises two main machines: the SA and the system (highlighted in yellow).

- The SA logs into his or her account (Circle 1 in Fig. 14).
- Assuming correct credentials, the system creates (2) a session.
- The SA issues a request (3) for system maintenance, such as a software update.
- Upon receiving the request, the system performs the authentication process (4) [34] as follows:
  - The system generates random data (5) using the SA's public key and sends it to him or her (6).
  - The SA processes (7) the random data using his or her private key (8) and sends its encrypted version to the system (9).
  - The system uses the SA's public key (10) to decrypt (11) the incoming encrypted data, producing decrypted data (12).
  - The decrypted data are compared (13) to the original random data; if they are equivalent, a system maintenance session is opened for the SA (14).

A selected set of events are described as follows (see Fig. 15):

Event 1 (E_1): The SA logs into his or her account, and the system creates a session accordingly.
Event 2 (E_2): The SA issues a request to maintain the system.
Event 3 (E_3): The system starts the authentication process by generating random data and sending it to the SA.
Event 4 (E_4): The SA processes the random data using his or her private key and sends the encrypted data to the system.
Event 5 (E_5): The system uses the SA's public key to decrypt the incoming encrypted data, producing a decrypted dataset.
Event 6 (E_6): The original random data are transferred to the comparison module.
Event 7 (E_7): The decrypted data are compared to the original random data.
Event 8 (E_8): If the data are equivalent, a system maintenance session is opened for the SA.

Fig. 16 shows the chronology of these events that model the behavior of the PKI-based authentication system.

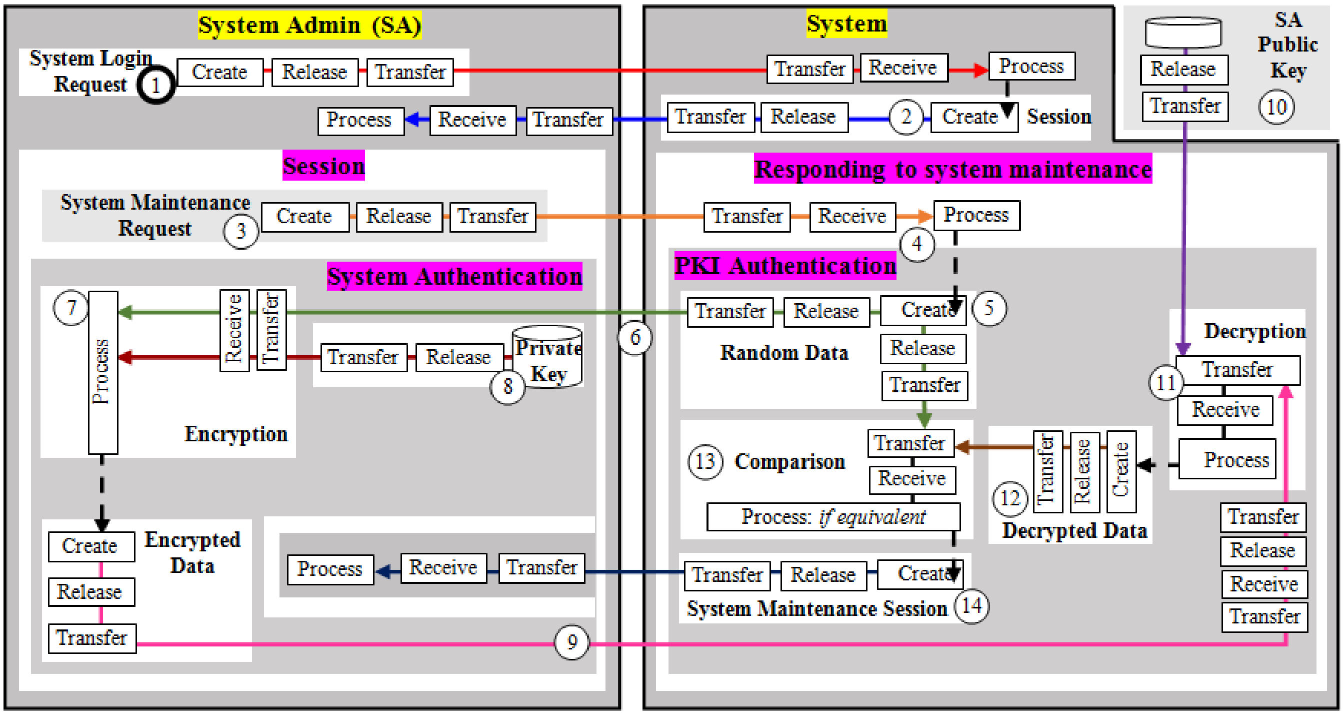


Fig. 14. TM Representation of UML use Case Involving the SA in PKI Authentication.

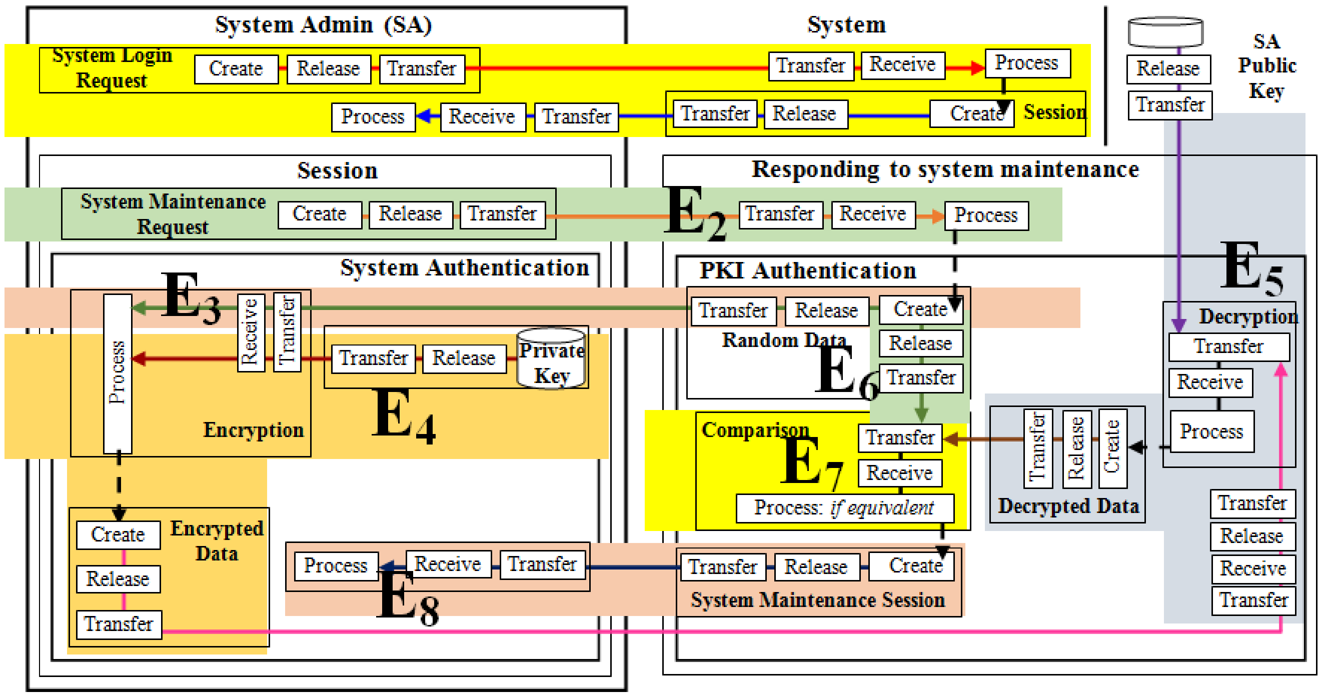


Fig. 15. Meaningful Events During PKI Authentication.

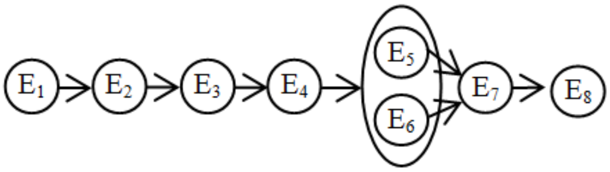


Fig. 16. Control of the PKI Event Sequence.

B. Biometric Authentication

Fig. 17 and 18 show the static and dynamic TM representations of the SA's roles under a physical biometric authentication system. A typical physical biometric system carries out authentication in two stages—the enrollment stage and the verification stage.

Fig. 17 comprises two main machines: the SA and the system (highlighted in yellow).

- Initially, in the enrollment stage, the SA requests (1) the biometric trait desired, such as a face or fingerprint.
- In response, the system requests (2) the SA to present his or her chosen biometric trait.
- The SA then presents (3) the trait to the scanning hardware.
- The system then extracts (4) the scanned trait for encryption and storage (5).
- To initiate an interaction with the system, the SA logs into his or her account (6). With the correct credentials, the system creates (7) a session.
- The SA issues a request (8) to maintain the system (e.g., software update).
- The system starts the authentication process (9) (verification stage) [35].
    - The system requests (10) the SA to present his or her chosen biometric trait.
    - The SA then presents (11) the trait to the scanning hardware.
    - The system then extracts (12) the scanned trait for comparison purposes.
    - The originally encrypted trait is decrypted (13) and compared with the trait extracted from the scanning hardware (14). If they are equivalent, a system maintenance session is opened to the SA (15).

Fig. 18 shows the dynamic description of the model. A selected set of events is described as follows:

Event 1 (E_1): The SA requests the biometric trait desired for the enrollment stage, and the system requests the SA to present the chosen biometric trait.
Event 2 (E_2): The SA presents the trait to the scanning hardware for extraction.
Event 3 (E_3): The extracted data are then encrypted and stored.
Event 4 (E_4): The SA logs into his or her account, and the system creates a session accordingly.
Event 5 (E_5): The SA issues a request for system maintenance.
Event 6 (E_6): The system starts the authentication process by requesting the SA to present the chosen biometric trait.
Event 7 (E_7): The SA presents the trait to the scanning hardware for extraction.
Event 8 (E_8): The system decrypts the originally encrypted trait.
Event 9 (E_9): The extracted trait is compared to the decrypted data.
Event 10 (E_{10}): If the data are equivalent, a system maintenance session is opened to the SA.

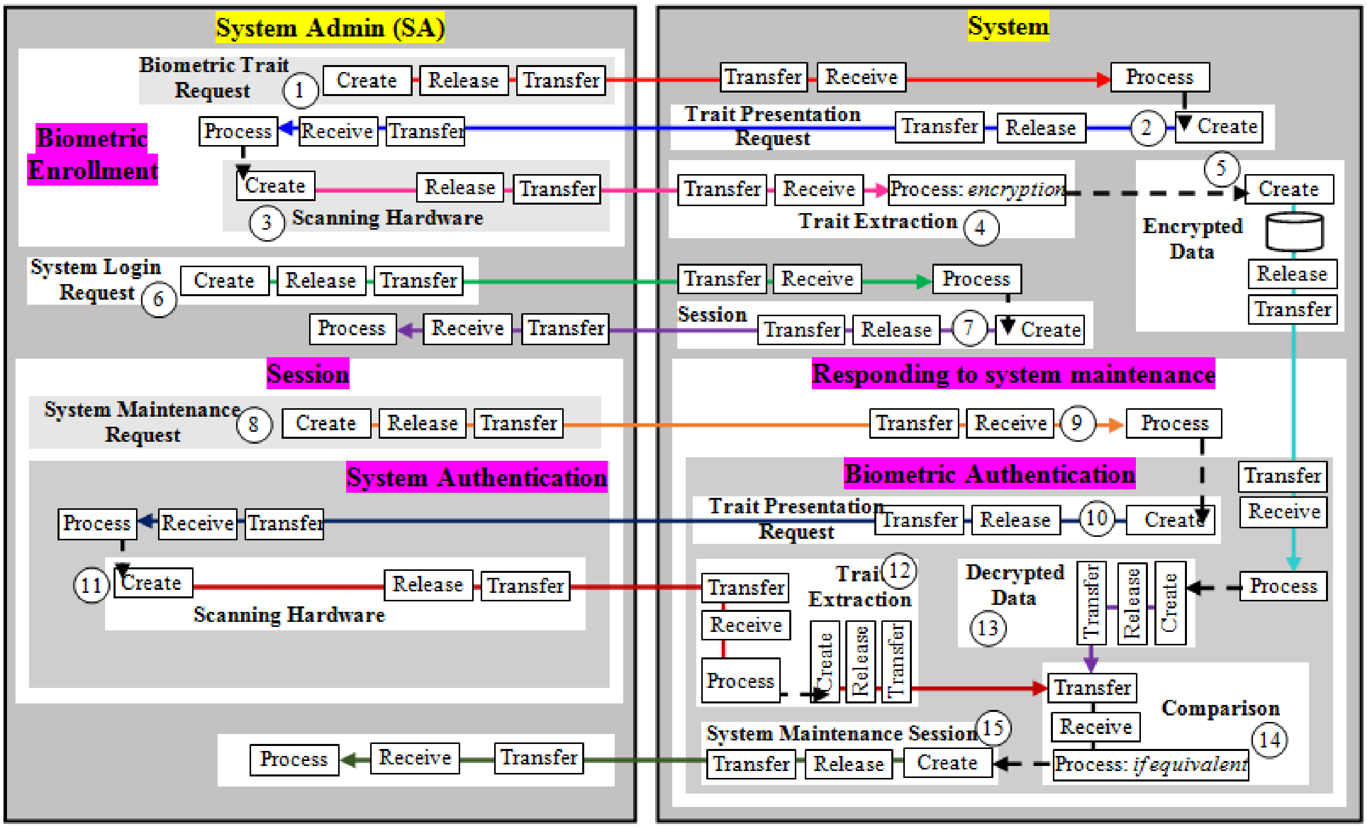


Fig. 17. TM Representation of UML use Case Involving the SA in Physical Biometric Authentication.

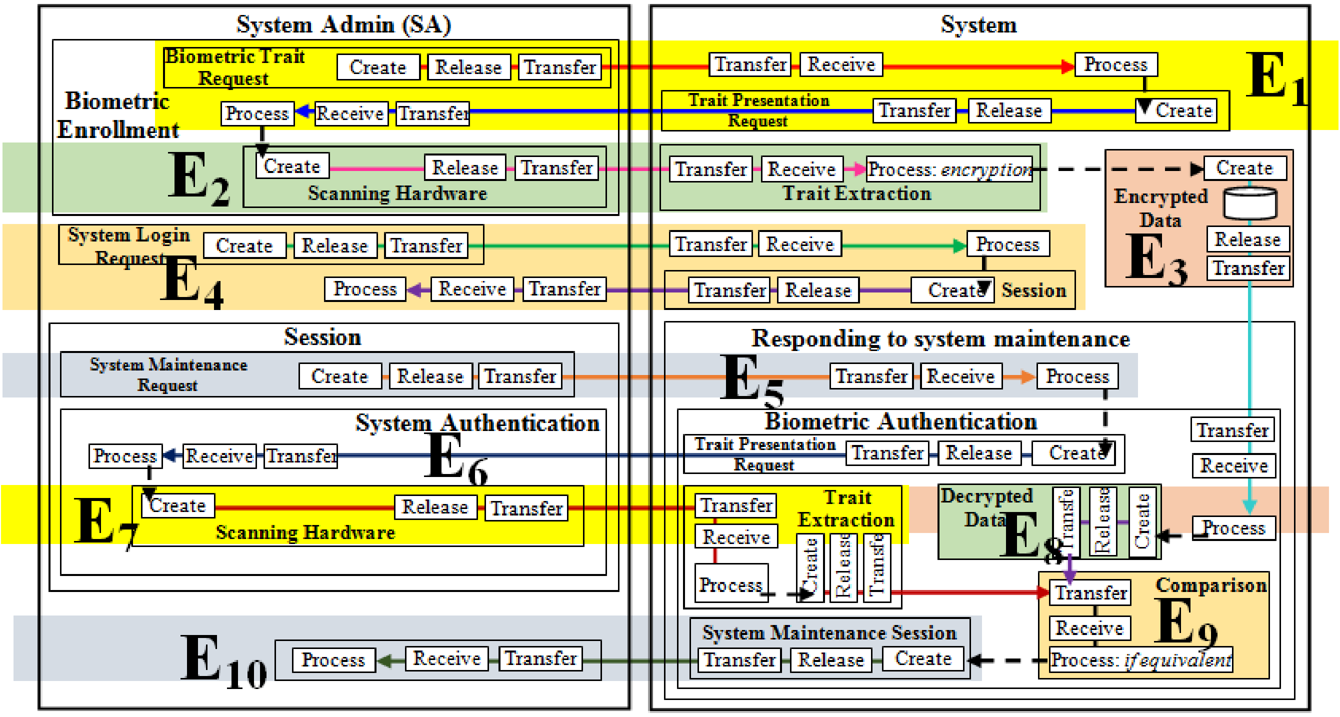


Fig. 18. Meaningful Events During Physical Biometric Authentication.

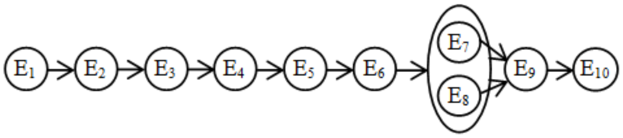


Fig. 19. Control of the Physical Biometric Event Sequence.

Fig. 19 shows the chronology of events modeling the behavior of the biometric authentication system.

C. Multifactor Authentication

The diagrams for multifactor authentication are not shown for space considerations. A typical multifactor system carries out authentication in two or more stages—the login stage and other verification stage(s) involving other types of authentication.

D. Multifactor Authentication

The diagrams for multifactor authentication are not shown for space considerations. A typical multifactor system carries out authentication in two or more stages—the login stage and other verification stage(s) involving other types of authentication.

This paper assumes the common choice of randomly generated one-time passwords (OTPs) with two-factor authentication. The TM model comprises two main machines: the SA and the system (highlighted in yellow).

- To initiate an interaction with the system, the SA logs into his or her account. With the correct credentials, the system creates a session.
- The SA issues a request to maintain the system (e.g., software update).
- The system starts the authentication process [36].
  - The system identifies the SA's registered phone number and uses it to generate an OTP.
  - This password is embedded in an SMS and transferred to the system's phone.
  - The system sends the SA an SMS containing the OTP.
  - The SA then inputs the requested OTP in the displayed form.
  - The system extracts the entered OTP for comparison.
  - The OTP entered is compared to the one initially sent. If they are the same, a system maintenance session is opened to the SA.

VI. Conclusion

In this paper, we presented the thesis that five generic processes—creating, releasing, transferring, receiving, and processing—have the expressive power to model key public infrastructure, biometric, and multifactor authentications. Expressiveness refers to things said in a description in a language [2]. The interesting aspect of the TM is the question of whether TM's five generic processes express all things required in conceptual modeling in software engineering. Indicators including modeling authentication in this paper point to the viability of this hypothesis. Further research should pursue this line of thinking.

References

[1] G. Kotonya and I. Sommerville, Requirements Engineering: Processes and Techniques. Hoboken: John Wiley & Sons, 1998.

[2] S. Patig, "Measuring Expressiveness in Conceptual Modeling," in Advanced Information Systems Engineering, A. Persson and J. Stirna, Eds. Berlin: Springer, 2004, pp. 127–141 [CAiSE 2004, Lecture Notes in Computer Science, vol. 3084].

[3] R. Y. Lee, "Chapter 4: Modeling with UML," in Object-Oriented Software Engineering with UML: A Hands-On Approach. Hauppauge, NY: Nova Science Publishers, Inc., January 2019.

[4] O. Altuhhova, R. Matulevičius, and N. Ahmed, "Towards definition of secure business process," in Lecture Notes in Business Information Research. Berlin: Springer, 2012, pp. 1–15 [CAiSE 2012 Workshop on Information Systems Security Engineering, 2012].

[5] P. Bresciani, P. Giorgini, F. Giunchiglia, J. Mylopoulos, and A. Perini, "TROPOS: An agent-oriented software development methodology," J. Auton. Agents Multi-Agent Syst, vol. 8, no. 3, pp. 203–236, May 2004.

[6] I. Soomro and N. Ahmed, "Towards security risk-oriented misuse cases," in Business Process Management Workshops. Berlin: Springer, vol. 132, 2013, pp. 689–700.

[7] G. Sindre, Mal-Activity Diagrams for Capturing Attacks on Business Processes. In: Sawyer P., Paech B., Heymans P. (eds) Requirements Engineering: Foundation for Software Quality. REFSQ 2007. Lecture Notes in Computer Science, vol 4542. Springer, Berlin, Heidelberg, pp. pp 355-366, 2007.

[8] Object Management Group, OMG Unified Modeling Language Superstructure. Version 2.2, http://www.omg.org.

[9] J. Evermann, "Thinking ontologically: Conceptual versus design models in UML," in Ontologies and Business Analysis, M. Rosemann and P. Green, Eds., Location: Idea Group Publishing, 2005.

[10] M. Brambilla and P. Fraternali, "Chapter 11: Tools for model-driven development of interactive applications," In View on ScienceDirect Interaction Flow Modeling Language, M. Brambilla, P. Fraternali, and M. Kaufmann (eds.), Elsevier Science, pp. 335-358, 2015.

[11] S. Al-Fedaghi and M. Bayoumi, "Computer attacks as machines of things that flow," 2018 International Conference on Security and Management, Las Vegas, NV, July 30–August 2, 2018.

[12] H. Morris, To Be, To Have, To Know: Smart Ledgers & Identity Authentication, Z/Yen Group, February 2019. https://www.zyen.com/media/documents/To_Be_To_Have_To_Know_Smart_Ledgers__Identity_Authentication.pdf

[13] The Unified Modeling Language, Single Sign-On for Google Apps UML Activity Diagram Example, accessed 3/8/2019. https://www.uml-diagrams.org/google-sign-on-uml-activity-diagram-example.html

[14] H. Storrle and J. H. Hausmann, "Towards a formal semantics of UML 2.0 activities," Software Engineering 2005, vol. P-64 of Lecture Notes on Informatics, Bonn, Germany, pp. 117-128, 2005.

[15] V. Plavsic and E. Secerov, "Modeling of login procedure for wireless application with interaction overview diagrams," Comput. Sci. Inf. Syst. vol. 5, no. 1, pp. 87–108, June 2008.

[16] S. Al-Fedaghi and G. Aldamkhi, "Conceptual modeling of an IP phone communication system: A case study," 18th Annual Wireless Telecommunications Symposium, New York, NY, April 9–12, 2019.

[17] S. Al-Fedaghi and O. Alsumait, "Toward a conceptual foundation for physical security: Case study of an IT department," Int. J. Saf. Secur. Eng., vol. 9, no. 2, pp. 137–156, 2019.

[18] S. Al-Fedaghi and Y. Atiyah, "Modeling with thinging for intelligent monitoring system," IEEE 89th Vehicular Technology Conference: VTC2019-Spring Kuala Lumpur, Malaysia, April 28–May 1, 2019.

[19] S. Al-Fedaghi and J. Al-Fadhli, "Modeling an unmanned aerial vehicle as a thinging machine," 5th International Conference on Control, Automation and Robotics, Beijing, China, April 19–22, 2019.

[20] S. Al-Fedaghi and E. Haidar, "Programming is diagramming is programming," 3rd International Conference on Computer, Software and Modeling, Barcelona, Spain, July 14–16, 2019.

[21] M. Heidegger, "The thing," in Poetry, Language, Thought, A. Hofstadter, Trans. New York: Harper & Row, 1975, pp. 161–184.

[22] K. Riemer, R. B. Johnston, D. Hovorka, and M. Indulska, "Challenging the philosophical foundations of modeling organizational reality: The case of process modeling," International Conf. on Information Systems, Milan, Italy, 2013. http://aisel.aisnet.org/icis2013/proceedings/ BreakthroughIdeas/4/.

[23] S. Al-Fedaghi, "Five generic processes for behaviour description in software engineering," Int. J. Comp. Sci. Inf. Secur., vol. 17, no. 7, July 2019.

[24] S. Al-Fedaghi, "Toward maximum grip process modeling in software engineering," Int. J. Comput. Sci. Inf. Secur., vol. 17, no. 6, June 2019.

[25] L. W. Howe, "Heidegger's discussion of 'the Thing': A theme for deep ecology," Between Species, vol. 9, no. 2, art. 11, 1993. doi:10.15368/bts.1993v9n2.9.

[26] L. R. Bryant, "Towards a machine-oriented aesthetics: On the power of art," paper presented at The Matter of Contradiction Conference, Limousin, France, 2012.

[27] P. Byrd, Generic Meaning, accessed 5/8/2019. http://www2.gsu.edu/~eslhpb/grammar/lecture_5/generic.html

[28] G. Guizzardi and G. Wagner, "Tutorial: Conceptual simulation modeling with onto-UML," Proceedings of the 2012 Winter Simulation Conference, Berlin, Germany, December 9–12, 2012.

[29] N. Nostro, A. Ceccarelli, A. Bondavalli, and F. Brancati, "Insider threat assessment: A model-based methodology," Op. Syst. Rev., vol. 48, no. 2, pp. 3–12, December 2014.

[30] S. M. Dejamfar and S. Najafzadeh, "Authentication techniques in cloud computing: A review," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 7, no. 1, pp. 95–99, January 2017.

[31] A. Banerjee and M. Hasan, Token-Based Authentication Techniques on Open Source Cloud Platforms, Systems and Telematics, Vol. 16, No. 47, pp. 9-29, October-December, 2018.

[32] M. Qasaimeh, R. Turab, R. S. Al-Qassas, Authentication techniques in smart grid: a systematic review, TELKOMNIKA, Vol.17, No.3, pp.1584-1594, June 2019.

[33] A. Agarkar and H. Agrawal, A review and vision on authentication and privacy preservation schemes in smart grid network, Security and Privacy, Vol. 2, No. 2, pp. 1-18,March/April 2019.

[34] M. Furuhed (2018). Public key infrastructure (PKI) explained in 4 minutes, Nexusgroup.com, accessed 5/8/2019. https://www. nexusgroup.com/blog/crash-course-pki.

[35] W. Yang, S. Wang, J. Hu, G. Zheng, and C. Valli, "Security and accuracy of fingerprint-based biometrics: A review," Symmetry, vol. 11, no. 2, art. 141, January 2019. https://www.mdpi.com/2073-8994/11/2/141.

[36] K. Garska (2018). Two-Factor Authentication (2FA) Explained: Email and SMS OTPs, Identity Automation Site, September 27, 2018. https://blog.identityautomation.com/two-factor-authentication-2fa-explained-email-and-sms-otps.

Internal Threat Defense using Network Access Control and Intrusion Prevention System

Andhika Surya Putra[1], Nico Surantha[2]
Computer Science Department, BINUS Graduate Program–Master of Computer Science
Bina Nusantara University, Jakarta, Indonesia 11480

***Abstract*—This study aims to create a network security system that can mitigate attacks carried out by internal users and to reduce attacks from internal networks. Further, a network security system is expected to be able to overcome the difficulty of mitigating attacks carried out by internal users and to improve network security. The method used is to integrate the ability of Network Access Control (NAC) and the Intrusion Prevention System (IPS) that have been designed and implemented in this study, then an analysis is performed to compare the results of tests that have been carried out using only the NAC with the results using integration of NAC capabilities and IPS. The results obtained from the tests that have been carried out, namely, the security system by using the integration of NAC and IPS capabilities is better than using only the NAC.**

Keywords—Attack; integration; Intrusion Prevention System (IPS); mitigation; Network Access Control (NAC); network security

I. Introduction

For decades, technology plays an important role in most activities. Most organizations use technology to support their business processes. Nowadays, internet is used for almost all activities especially business activities. Thus, network infrastructure plays a vital function in an organization. Most organizations are connected to the internet to make all information easily accessed from anywhere and anytime. Network can also be considered as a major risk for an organization. Today's advancement of IT technology bring to the surface the issue of security. Thus, it is important to secure the network infrastructure [1] [2]. In the operation of network can be compromised by any vulnerability in their functionality to attack the networks. Some mechanisms are widely used to secure the network, namely Intrusion Detection System (IDS) that has the ability to detect malicious and unauthorized activities and Intrusion Prevention System (IPS) that has the ability to make an action for detected intrusion [3] [4] [5]. The purpose of using IDPS is to monitor and protect attacks from intruder who want to enter the system, and then give a report to the network administrator if there are attacks that occur in the network environment [6] [7]. So, using IDPS can help to detect and carry out security against intrusions that occur on the network.

Attack threats can be caused by either outsiders or insiders in an organization. Insider attacks are malicious attacks carried out on networks or computer systems with authorized/official system access [8] [9]. Insider attack is one of the most difficult threats to be detected because (IDS) is built to defend against outside attacks [10]. Generally, IPS is placed in the edge of a network, it is done so to avoid incoming intrusion flows from the outside [11]. Thus, concerns that attacks can still arise from inside intruders to network before reaching IPS still exists in the network. Therefore, a network security is needed from the lowest level of the internal network as well.

Network Access Control (NAC) is an approach designed to increase network security by controlling the access and the resources for legitimate users [12]. NAC not only allows network access requested by the user, but also provide specific access based on the user's identity [13]. One of the threats to enterprise networks is the personal devices of employees and guests that do not have anti-virus, patches or host intrusion prevention system in place. An NAC solution can protect a network from such end devices and detect and rectify these problems [14]. NAC function has certain weaknesses, in particular it is unable to detect and stop users that have legitimate network access form carrying out intentional or unintentional attacks from within. Example of intentional attack is when an internal user has a desire to destroy the internal system due to personal problems, whereas unintentional attacks can happened through downloading files or applications that contain malware or viruses. This condition can happen because NAC does not have the ability to detect attacks like IPS.

Based on the weaknesses of the NAC, there is a need to improve network access security from within. In this research, a solution is proposed to improve network security from internal sides of the network by integrating NAC and IPS capabilities. The benefits obtained from this solution can minimize the threat of attacks on the network.

II. Review of Related Literature

A. NAC

NAC systems combine endpoint security solutions to grant access control and enforce security rules or policies to every device connected to the network. The NAC policy is able to identify endpoints that are connected to the network. This policy is carried out to restrict access of devices that do not comply with predetermined network access rules [15]. NAC also provides security and control for those who have access to networks and resources within the network. Basically, NAC performs posture, quarantine, and remediation checks involved in requests for network access by users. If the user does not have the appropriate posture in his computer such as the latest OS/security patch or the most updated antivirus, then the user will not be allowed to enter the network, but the user will be quarantined by being separated into different networks or VLANs until the user performs a remediation process to meet the requirements needed for entry into the network [16] [15].

Some reasons for using the NAC solution are: to identify and authenticate users and endpoints, to limit user access to the network, to limit access based on the endpoint security posture, and to remediate an endpoint if the endpoint does not have a posture that complies with the provisions [17]. Another reason for implementing the NAC is due to the threat that comes from using your own device (BYOD) approach. With many users using their own devices to work and use them for work purposes, NAC is increasingly needed because many security threats might occur due to devices that do not have enterprise-level device security standards such as patch OS and antivirus.

Comparative studies of existing NAC systems has concluded that the NAC solutions from Cisco, Trustwave, and Forescout can be implemented in accordance with the existing network infrastructure so that it can produce maximum profits where NAC can limit the access of devices and users that are defined based on existing roles and ensure network access obtained according to what is needed [17]. The main benefit of NAC systems is to prevent potentially malicious or infected devices from entering the network in order to keep the network clean [18]. So that network security can be increased from the user level by using NAC.

B. *IDS and IPS*

There are a couple of widely used mechanisms to secure the network, namely Intrusion Detection System (IDS) and Intrusion Prevention System (IPS). Intrusion detection is the process of monitoring events on a network or computer system and analyzing them for possible threat incidents and violations of standard computer security practices, usage policies, or security policies [3]. IDS is a hardware component or software that automates the intrusion detection process. It monitors events that occur on network and computer systems and responds to alert with an indication of potential network security policy violations [19] [20]. IPS is a network device or software that identify and block network threats by assessing each and every packet based on the network protocols in the network layer, tracking each session. Intrusion Prevention System is a defense mechanisms designed to detect malicious packets within network traffic and stop intrusions, blocking the aberrant traffic automatically before it does any [3] [21]. IPS is an improvement from IDS because it does not only have the ability to detect intrusion, but also can take action against intrusion or potential malicious network activity [22] [23].

There are several approaches that have been carried out by previous researchers. The objective to be achieved in their research is to evaluate and analyze the performance of NGIPS in securing networks through penetration tests using HTTP ports, so that the inspection and protection performance of NGIPS is known. The benefit of this research is that it can be a point of reference in improving network security using the NGIPS method and to obtain optimal mechanism for implementing NGIPS. Based on the results of these penetration tests, it proves that NGIPS can save attacks that exploit vulnerabilities from HTTP ports [24]. By using IPS, attacks that cannot be detected by a firewall and NAC can be detected properly and therefore increase network security.

III. Methodology

A. *System Design*

As summarized in Table I, IPS used in this study is products from Cisco, Cisco FirePower 8250 with OS 6.2.3 series. The IPS has been connected between a firewall device and core switch. Using the NAC system from Cisco requires a Cisco NAC device called the Cisco Identity Service Engine (ISE). Cisco ISE is a Cisco appliance used for NAC systems. The ISE will be linked and integrated with existing network infrastructure devices such as switches and radius servers to authenticate. Physically, the Cisco ISE will be connected to the server-farm switch. This is done so that Cisco ISE can be integrated with all segments in the network infrastructure. The hardware of Cisco ISE used is appliance with SNS 3495 type and the OS version of Cisco ISE used is ISE 2.3. The NAC will be able to communicate with IPS to carry out the expected integration in accordance with the objectives of this study. The access switch that directly connected to the user’s PC uses Cisco Catalyst 2960X with 15.2(2)E7 IOS version. The computer used as an attacker and the target is HP ProOne 600 using windows 10.

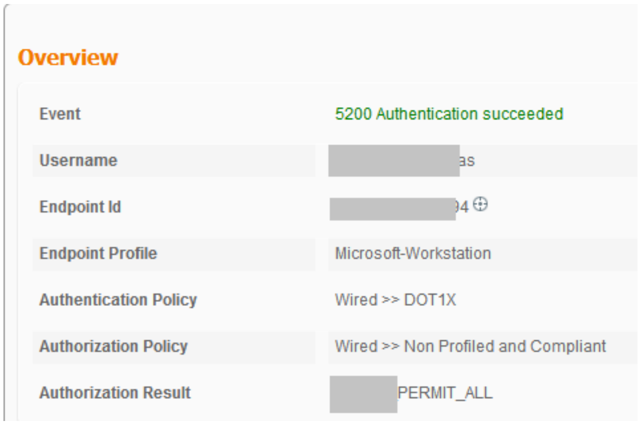
Fig. 1 below proposes a new topology using Cisco FirePower and Cisco ISE connected to the network. In this study, the device was integrated with existing infrastructure. The integration carried out in this study is physical and logical connection where the Cisco ISE NAC and IPS Cisco FirePower must be able to connect with existing infrastructure devices, change server farm connections from core switch through IPS, configure to integrate between the Cisco ISE NAC, IPS Cisco FirePower and existing infrastructure devices, as well as making policies and rules on the Cisco ISE NAC and IPS Cisco FirePower to achieve the objectives in this study which are to create a network security system that can mitigate internal users who carry out attacks and to reduce attacks from internal networks by using NAC and IPS system integration.

B. *Implementation and Testing*

In this study, tests were carried out to prove the solution given to address the existing problems. These tests were carried out by using a system and infrastructure design that has been integrated with the NAC and IPS systems. The tests took place by trying to connect an internal user PC to the internal network with the Windows 10 operating system which acted as an attacker, placing the target server connected to the firewall using workstations with vulnerable OS installed. Then, worked on the IPS and NAC configuration so that the two systems can communicate and integrate in order to achieve the objectives of this study.

TABLE. I. System Specifications

Device	Vendor	OS Version
PC	HP ProOne 600	Windows 10
Access Switch	Cisco Catalyst 2960X	15.2(2)E7
NAC	Cisco ISE SNS 3495	2.3
IPS	Cisco FirePower 8250	6.2.3



Overview

Event	5200 Authentication succeeded
Username	[redacted]as
Endpoint Id	[redacted]4 ⊕
Endpoint Profile	Microsoft-Workstation
Authentication Policy	Wired >> DOT1X
Authorization Policy	Wired >> Non Profiled and Compliant
Authorization Result	[redacted]PERMIT_ALL

Fig. 1. User Compliant Status on NAC.

In this test, the IPS is located in the middle of a network that is configured inline so that the IPS can immediately make decisions on packages that have been inspected. The package is analyzed by the IPS based on its signature. If the package contains crime or vulnerability, the IPS will immediately prevent it by blocking the malicious package. Then, the IPS provides information on the source of the attack to the NAC so that the NAC can immediately prevent and quarantine the computer that is the source of the attack so that it cannot launch attacks again on the network. In this test, there were two types of users, a compliant user which is an official network access condition with certain requirements and a noncompliant user which is an unofficial network access condition because the user does not comply with the specified requirements. Screening tests were conducted by using several legitimate traffic samples such as HTTP and SSH as well as malicious traffic such as sql injection and os bash injection.

IV. Result and Discussion

A. *Compliant user with Legitimate Traffic*

In the test, the computer used by a compliant user tried to access the network by physically connecting the cable from the PC to the access switch. The PC user is considered compliant by the NAC because it complies with the specified requirements such as a join domain and has anti-virus as shown in Fig. 2. Then the user tried to access legitimate HTTP traffic to the server with the vulnerable OS used in this study successfully as shown in Fig. 3. Afterward, the user tried to access legitimate SSH traffic to the server with the vulnerable OS used in this study successfully as shown in Fig. 4. Comparing the result of legitimate HTTP and SSH traffic by using only the NAC with the one using integration of the NAC and the IPS, the result is similar that the user can access any legitimate HTTP and SSH traffic.

B. *Compliant user with Malicious Traffic*

Based on tests, the computer used by a user tried to access malicious traffic with sql injection attack. The attack used "hi' or 1=1--" command on login field in web browser that aimed to trick sql server to bypass the login on the server. With only using the NAC, this attempt succeeds to bypass the website login with the sql injection command as shown in Fig. 5. This can be done because the NAC cannot detect the sql injection attack as the NAC only knows that the attacker is a compliant or authorized user. But by using integration of the NAC and the IPS, the sql injection attempt was detected and blocked by the IPS. The IPS instructed the NAC to quarantine the infected user immediately, so the user cannot do any more attack because the NAC is blocking the attacker connection via the access switch as shown in Fig. 6.

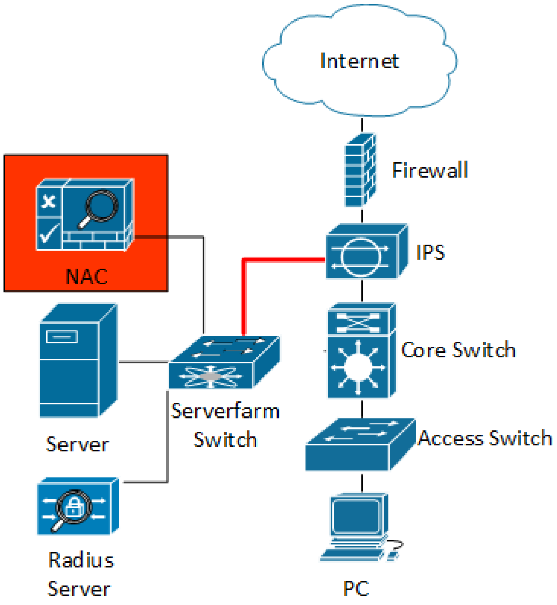
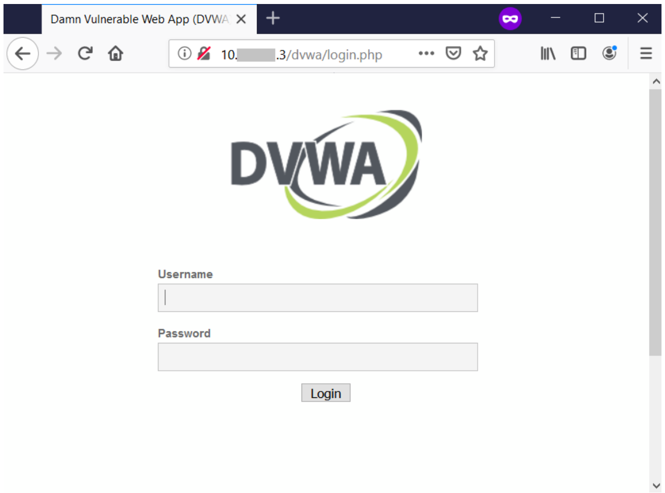


Fig. 2. Proposed Network Topology.



Damn Vulnerable Web App (DVWA

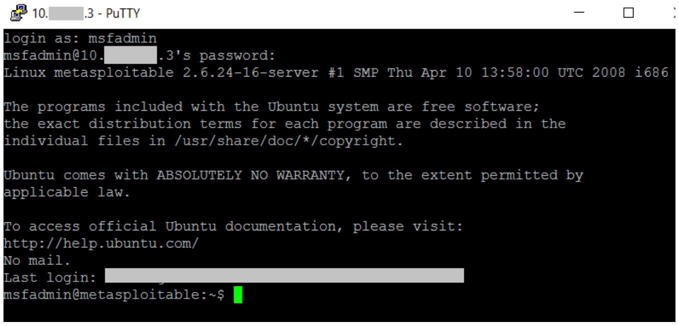
10.[redacted].3/dvwa/login.php

Username

Password

Login

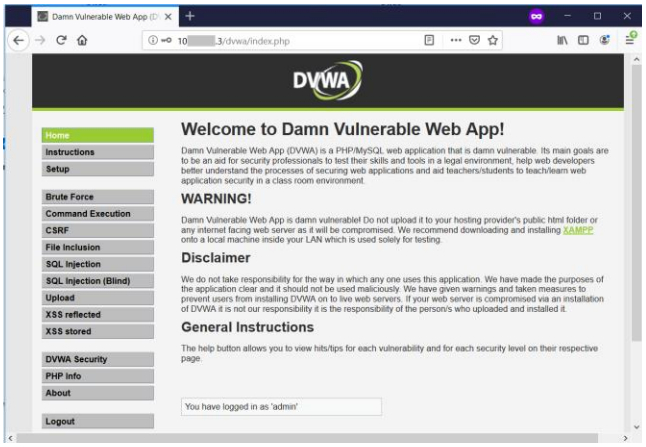
Fig. 3. Compliant user Legitimate HTTP Traffic.



10.[redacted].3 - PuTTY

```
login as: msfadmin
msfadmin@10.[redacted].3's password:
Linux metasploitable 2.6.24-16-server #1 SMP Thu Apr 10 13:58:00 UTC 2008 i686

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

To access official Ubuntu documentation, please visit:
http://help.ubuntu.com/
No mail.
Last login: [redacted]
msfadmin@metasploitable:~$
```

Fig. 4. Compliant user Legitimate SSH Traffic.



Damn Vulnerable Web App (D ×

10.[redacted].3/dvwa/index.php

Welcome to Damn Vulnerable Web App!

Damn Vulnerable Web App (DVWA) is a PHP/MySQL web application that is damn vulnerable. Its main goals are to be an aid for security professionals to test their skills and tools in a legal environment, help web developers better understand the processes of securing web applications and aid teachers/students to teach/learn web application security in a class room environment.

WARNING!

Damn Vulnerable Web App is damn vulnerable! Do not upload it to your hosting provider's public html folder or any internet facing web server as it will be compromised. We recommend downloading and installing XAMPP onto a local machine inside your LAN which is used solely for testing.

Disclaimer

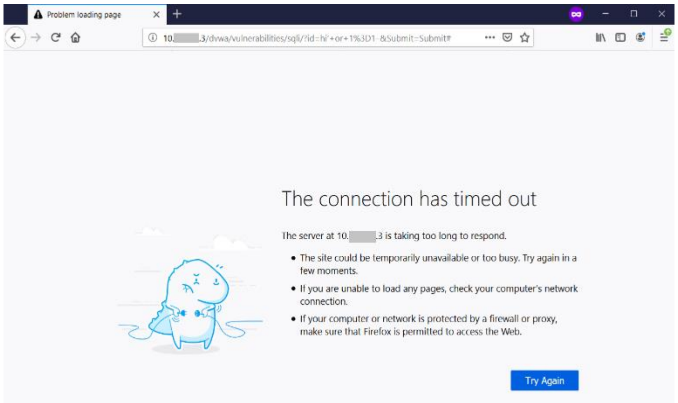
We do not take responsibility for the way in which any one uses this application. We have made the purposes of the application clear and it should not be used maliciously. We have given warnings and taken measures to prevent users from installing DVWA on to live web servers. If your web server is compromised via an installation of DVWA it is not our responsibility it is the responsibility of the person/s who uploaded and installed it.

General Instructions

The help button allows you to view hits/tips for each vulnerability and for each security level on their respective page.

You have logged in as 'admin'

Fig. 5. Successful Login using Sql Injection Attack.



Problem loading page ×

10.[redacted].3/dvwa/vulnerabilities/sqli/?id=hi'+or+1%3D1-&Submit=Submit#

The connection has timed out

The server at 10.[redacted].3 is taking too long to respond.

- The site could be temporarily unavailable or too busy. Try again in a few moments.
- If you are unable to load any pages, check your computer's network connection.
- If your computer or network is protected by a firewall or proxy, make sure that Firefox is permitted to access the Web.

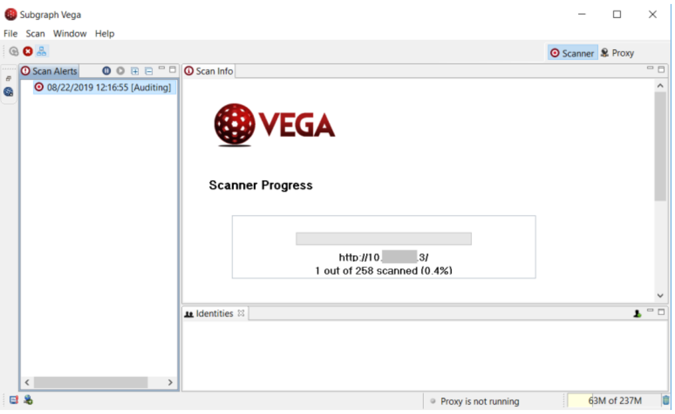
Fig. 6. Blocked Attacker Network Access by NAC.

The second malicious attempt used is through os bash injection. Os bash injection attacks were carried out on the target server by using the vega vulnerability scanner tool on the attacker's computer. Run the vega tools then run a scan to test the os bash injection attack as shown in Fig. 7. With only using the NAC, the attempts still succeed to launch attacks to the targeted vulnerable server. But by using integration of the NAC and the IPS, the os bash injection attempt was detected and blocked by the IPS. The IPS instructed the NAC to quarantine the infected user immediately, so the user cannot do any more attack because the NAC is blocking the attacker connection via the access switch as shown in Fig. 8.

C. Noncompliant User

A user is deemed noncompliant by the NAC because the user does not pass the required NAC system, the user cannot get any network access and mitigated by the NAC by denying the network access for the user as shown in Fig. 9. Therefore, user cannot go through with either malicious traffic or even legitimate traffic as they have no access to the network.

Table II summarized the results based on the tests that have been completed for this study. It shows significantly different results in treating compliant users who commit malicious traffic on the network only with the NAC with the one using the proposed solution of integration between the NAC and the IPS. The expected test results in this study can be achieved by using the proposed solution. The proposed solution shows that by integrating capabilities of the NAC and the IPS can mitigate attacks from internal users and can reduce attacks from internal networks by 40% based on the test scenarios performed. Therefore, integration of the NAC and the IPS can increase network security compared to the use of NAC alone.



08/22/2019 12:16:55 [Auditing]

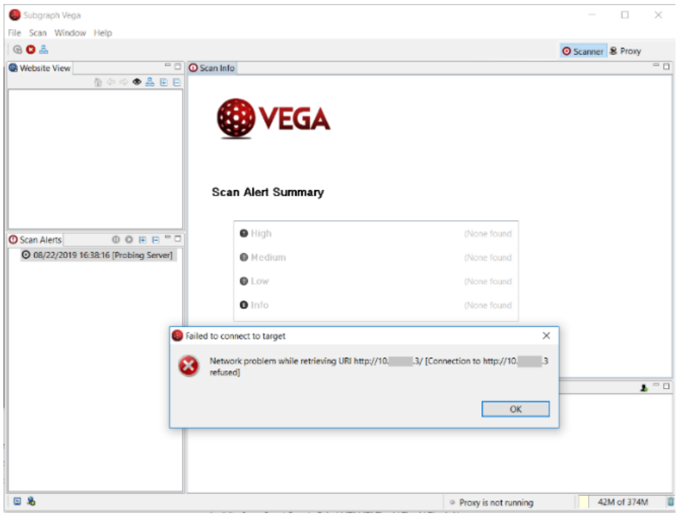
Scanner Progress

http://10.[redacted].3/
1 out of 258 scanned (0.4%)

Proxy is not running

63M of 237M

Fig. 7. OS Bash Injection Attack Attempt.



08/22/2019 16:38:16 [Probing Server]

Scan Alert Summary

High	(None found
Medium	(None found
Low	(None found
Info	(None found

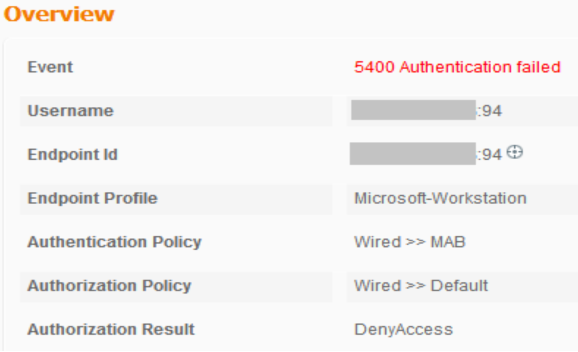
Failed to connect to target

Network problem while retrieving URI http://10.[redacted].3/ [Connection to http://10.[redacted].3 refused]

Proxy is not running

42M of 374M

Fig. 8. Blocked Attacker Network Access by NAC.



Overview

Event	5400 Authentication failed
Username	[redacted]:94
Endpoint Id	[redacted]:94
Endpoint Profile	Microsoft-Workstation
Authentication Policy	Wired >> MAB
Authorization Policy	Wired >> Default
Authorization Result	DenyAccess

Fig. 9. Denied Network Access of Noncom Pliant user.

TABLE. II. RESULTS COMPARISON

No	User	Traffic	Target	Expectation	Result	
					NAC	**NAC & IPS (Proposed solution)**
1	Compliant	HTTP	Vulnerable OS	Allow	Allowed	**Allowed**
2	Compliant	SSH	Vulnerable OS	Allow	Allowed	**Allowed**
3	Compliant	SQL Injection Attack	Vulnerable OS	Block	Allowed	**Blocked**
4	Compliant	OS Bash Injection Attack	Vulnerable OS	Block	Allowed	**Blocked**
5	Noncompliant	HTTP	Vulnerable OS	Block	Blocked	**Blocked**

V. CONCLUSION AND FUTURE WORK

Based on the test results performed in this study, the proposed solution is that by integrating the NAC system with the IPS can mitigate attacks from internal users on internal networks and attacks from internal networks. That network security with the integration of the NAC systems with the IPS can be increased as compared to the use of the NAC alone. However, this study still has many limitations, particularly on the types of attacks tested. There are so many different types of attacks on the internet. Therefore, in the future it is recommended to increase the types of attacks carried out in similar tests and do more detail experiment to compare the application on the internal network with the application on the external network in order to achieve a more comprehensive result.

REFERENCES

[1] W. Bul'ajoul, A. James and M. Pannu, "Improving network intrusion detection system performance through quality of service configuration and parallel technology," Journal of Computer and System Sciences, vol. 81, no. 6, p. 981–999, 2015.

[2] W. Bul'ajoul, A. James and S. Shaikh, "A New Architecture for Network Intrusion," IEEE Access, vol. 7, pp. 18558-18573, 2019.

[3] H. A. Razzak, A. Karim, S. S. Handa and M. V. Ramana Murthy, "A methodical approach to implement intrusion detection system in hybrid network," International Journal of Engineering Science and Computing, vol. 7, no. 3, pp. 4817-4820, 2017.

[4] G. Ahmed, M. N. A. Khan and M. Shamraiz, "A linux-based IDPS," Computer Fraud & Security, pp. 13-18, 2015.

[5] [5] S. P. Anilbhai and C. Parekh, "Intrusion detection and prevention system for IoT," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 2, no. 6, pp. 771-776, 2017.

[6] S. Khadafi, B. D. Meilani and S. Arifin, "Sistem keamanan open cloud computing menggunakan ids (intrusion detection system) dan ips (intrusion prevention system)," Jurnal IPTEK, vol. 21, no. 2, pp. 67-76, 2017.

[7] F. Arsin, M. Yamin and L. Surimi, "Implementasi security system menggunakan metode IDPS (intrusion detection and prevention system) dengan layanan realtime notification," semanTIK, vol. 3, no. 2, pp. 39-48, 2017.

[8] A. Borkar, A. Donode and A. Kumari, "A survey on intrusion detection system (IDS) and internal intrusion detection and protection system (IIDPS)," in International Conference on Inventive Computing and Informatics, 2017.

[9] M. Warren, "Modern IP theft and the insider threat," Computer Fraud & Security, no. 6, pp. 5-10, 2015.

[10] F. Y. Leu, K. L. Tsai, Y. T. Hsiao and C. T. Yang, "An internal intrusion detection and protection system by using data mining and forensic techniques," IEEE Systems Journal, pp. 1-12, 2015.

[11] R. S. Silva and E. L. C. Macedo, "A cooperative approach for a global intrusion detection system for internet service providers," Cyber Security in Networking Conference, vol. 1, pp. 1-8, 2017.

[12] J. F. Matthews, "Challenges to implementing network access control," SANS Institute InfoSec Reading Room, p. 2, 2017.

[13] M. Roopesh, G. Reethika, B. V. Srinath and A. Sarumathi, "Network access control," International Journal on Computer Science and Engineering (IJCSE). Vol. 9, pp. 338-343, 2017.

[14] M. S. Inamdar and A. Tekeoglu, "Security analysis of open source network access control in virtual networks," International Conference on Advanced Information Networking and Applications Workshops, vol. 32, pp. 475-480, 2018.

[15] M. A. Muhammad and A. Ayesh, "A behaviour profiling based technique for network access control systems," International Journal of Cyber-Security and Digital Forensics (IJCSDF), vol. 8, no. 1, pp. 23-30, 2019.

[16] A. Sood, "Network access control," Rivier Academic Journal, vol. 3, pp. 1-12, 2007.

[17] T. J. Dildy, "Network access control-has it evolved enough for enterprises?," ISACA Journal Vol. 4, pp. 1-5, 2016.

[18] K. O. Detken, M. Jahnke, C. Kleiner and M. Rohde, "Combining network access control (nac) and siem functionality based on open source," in IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Bucharest, 2017.

[19] R. R. Chaudhari and S. P. Patil, "Intrusion detection system: classification, techniques and datasets to implement," International Research Journal of Engineering and Technology, vol. 4, no. 2, pp. 1860-1866, 2017.

[20] V. Mahajan and S. K. Peddoju, "Deployment of intrusion detection system in cloud: a performance-based study," IEEE Computer Society, pp. 1103-1108, 2017.

[21] R. Jamar, A. Sogani, S. Mudgal, Y. Bhadra and P. Churi, "E-shield: detection and prevention of website," IEEE International Conference On Recent Trends in Electronics Information & Communication Technology, vol. 2, pp. 706-710, 2017.

[22] B. Y. Choi and D. G. Allison, "Intrusion prevention and detection in small to medium-sized enterprises," in SAIS, 2017.

[23] P. Rengaraju, V. R. Ramanan and C.-H. Lung, "Detection and prevention of DoS attacks in software-defined cloud networks," IEEE Conference on Dependable and Secure Computing, pp. 217-223, 2017.

[24] G. Duppa and N. Surantha, "Evaluation of network security based on next generation intrusion prevention system," Telkomnika, vol. 17, no. 1, pp. 39-48, 2019.

CBRm: Case based Reasoning Approach for Imputation of Medium Gaps

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

***Abstract*—This paper presents a new algorithm called CBRm for univariate time series imputation of medium-gaps inspired by the algorithm called Case Based Reasoning Imputation (CBRi) for short-gaps. The performance of the proposed algorithm is analyzed in meteorological time series corresponding to maximum temperatures; also it was compared with several similar techniques. Although the algorithm failed to overcome in some cases to other proposals regarding precision, the results achieved are encouraging considering that some weaknesses of other proposals with which it was compared were outperformed.**

Keywords—Case Based Reasoning; CBR; CBRm; univariate time series imputation; medium-gaps

I. Introduction

Time series data exist in nearly every scientific field, where data are measured, recorded and monitored, so it is understandable that missing values may occur [1]. The imputation or completeness of missing values in time series is a very important task, since if it is not performed it is very complicated or impossible to be able to successfully carry out a prediction or forecasting process.

In the research field of imputation, univariate time series are a special challenge, most of the standard algorithms rely on inter-attribute correlations to estimate values for the missing data [2]. In the univariate case no additional attributes can be employed directly, so effective univariate algorithms instead need to make use of the time series characteristics.

In time series, different gaps sizes of NA values can be found: 1 or 2 consecutive NAs (short-gaps), from 3 to 10 consecutive NAs (medium-gaps) and more than 10 consecutive NAs (big-gaps) [3]. In this paper, a new algorithm for univariate time series imputation of medium-gaps is proposed, which is based on Case Based Reasoning (CBR) in such a way that the historical data of the time series can be used to improve the estimation of NA values. This algorithm is called CBRm and is implemented very similarly to CBRi “unpublished” [4] algorithm.

CBRm uses the same case base that was implemented for CBRi “unpublished” [4], this case base was built from maximum daily temperatures of 9 years (2007-01-01 to 2015-12-31) recorded at the Punta de Coles weather station located in the Moquegua region - Peru. The fundamental difference respect to CBRi lies in the operation of both techniques. Fig. 1 shows in summary the CBRi imputation process. As it’s appreciated, this operation for medium-gaps can introduce bias to the left of the gap, this because CBRi was designed to impute time series for short-gaps, between 1 and 2 consecutive NAs. Something similar happens with the LANN and LANN+ algorithms that were also designed for short-gaps.

The CBRm imputation process is shown in Fig. 2. As can be seen when a value between prior and next is calculated, it is not assigned immediately after prior, but is assigned to the center of the NA series by doubling in the case that the total of NAs is an even number.

Additionally, this work also presents the results achieved by the algorithms called Local Average Nearest Neighbors LANN [3] and LANN+ [3] in medium-gaps imputation processes. So, a small adaptation for these algorithms was done, specifically in the part corresponding to the determination of the prior and next values.

The present work has been organized as follows: in the second section, a brief description of the work related to univariate time series imputation is shown. The third section shows the theoretical bases necessary for a better understanding of the content of the work. The fourth section describes the proposed algorithm and its implementation. The fifth section describes the results achieved, which are compared with different univariate time series imputation techniques. The sixth section shows the conclusions reached in the present work and finally in the seventh section, it is indicated, the works that can be carried out based on the results of the work presented.

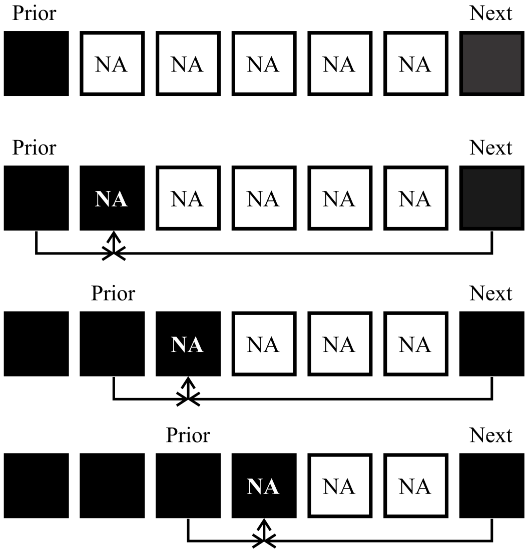


Fig. 1. CBRi Imputation Process.

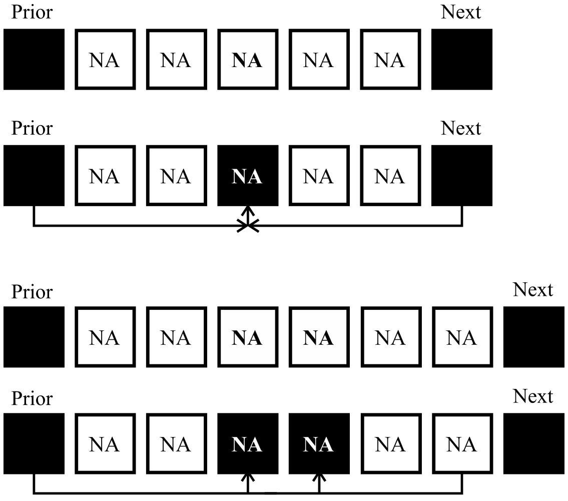


Fig. 2. CBRm Imputation Process.

II. Related Work

This section shows the results of the review of different techniques or algorithms for univariate time series imputation, from the oldest to the newest.

The first techniques for univariate time series imputation were quite simple and consisted of using arithmetic mean, median, mode, interpolation and Last Observation Carried Forward (LOCF) [5].

Last Observed Carried Forward (LOCF) is a technique for filling a NA value with the closest non-NA value prior to it [6]. Each individual NA value is replaced by the last observed value of that variable.

Baseline Observation Carried Forward (BOCF) [7] is similar to the LOCF; it replaces NA values with the non-missing baseline observation of the time series.

Hot-deck [8] [9], an NA value is replaced with an observed value that is closer in terms of distance. The Hot-deck algorithm randomly selects a value from a set of non-NA values and replaces the NA value. For comparative analysis, in this work, the hot-deck algorithm implemented in VIM R package is used.

Missing Value Imputation by Weighted Moving Average [10], is a set of algorithms that use the average or mean of the non-NA elements around an NA value. For an NA value at position *i* of a time series and assuming a window size of k=2, the observations *i-1, i+1* and *i+1, i+2* are used to calculate the mean.

There are three algorithms for univariate time series imputation in this category such as: Simple Moving Average (SMA) [10], Linear Weighted Moving Average (LWMA) [10] and Exponential Weighted Moving Average (EWMA) [10].

Simple Moving Average (SMA) [10] [11]: This algorithm for calculating the mean use all observations in the window which are equally weighted.

Linear Weighted Moving Average (LWMA) [10] [11]: In this algorithm weights decrease in arithmetical progression. The observations directly next to an NA value in position *i*, have weight *1/2*, the observations one further away (*i-2,i+2*) have weight *1/3*, the next (*i-3,i+3*) have weight *1/4*, and so on.

Exponential Weighted Moving Average (EWMA) [1] [10] [11]: it is an approach that allows imputing NA values by calculating the exponentially weighted moving average. Initially, the value of the window for the moving average is established, and then the average is calculated from the same number of observations on each side of the central missing value or NA value. The observations directly next to a central value *i*, have weight $(1/2)^1$, the observations one further away (*i-2,i+2*) have weight $(1/2)^2$, the next (*i-3,i+3*) have weight $(1/2)^3$, and so on. In this work, the algorithms SMA, LWMA and EWMA are implemented for comparative analysis using the imputeTS R package.

The Kalman filter [12], also known as LQE (linear quadratic estimation), is an algorithm that uses a series of measurements observed over time, which contains statistical noise and other inaccuracies, and produces estimates of unknown variables that tend to be more accurate than those based on a single measurement. Kalman filter integrated with ARIMA produces very good results in regression processes. In this work imputeTS package in R is used for implementing Kalman ARIMA imputation, imputeTS implements auto.arima [11] for better results.

LANN and LANN+ [3] are two fairly simple algorithms based on moving averages that show very good results in the short-gaps imputation process. As mentioned earlier, these techniques were adapted for the respective evaluation in medium-gaps. This adaptation only consisted of modifying the way in which these algorithms obtained the prior and next values.

For a comparative analysis of the results achieved by the imputation algorithm (CBRm) proposed in the present work, two well-known multivariate imputation algorithms were also implemented, such as KNN (K-Nearest Neighbor) [13] and MICE (Multiple Imputation by Chained Equations) [14] [15], these algorithms were implemented using the R VIM package for KNN and the mice package for MICE. In section V of this work, the achieved results can be seen.

III. Theoretical Background

A. Time Series

A time series is a sequence of data, observations or values, measured at certain time periods and sorted chronologically. The data can be spaced at equal intervals or uneven. For the analysis of the time series, different methods are used that help to interpret them and that allow extracting representative information about the underlying relationships between the data of the series.

One of the most common uses of time series is its analysis for prediction and forecasting. Time series are studied in different areas such as statistics, signal processing, econometrics, etc. Some features or characteristics of time series are: trends, cycles of seasonality and non-seasonality, pulses and steps, and outliers.

B. Missing Data

Depending on what causes missing data, the gaps will have a certain distribution. Understanding this distribution may be helpful in two ways [16]. First, this knowledge can be used to select the most appropriate imputation algorithm to complete the NA values. Secondly, this knowledge can help design an imputation model, which allows the elimination of the NA values from a set of test data. This model will help generate the NA values where the true values are known. Therefore, the quality of the model can be tested through different regression metrics such as RMSE, MAPE, etc.

Mechanisms of missing data can be classified into three categories: Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR). The process of completing NA-gaps in time series is sometimes complicated, since the underlying mechanisms are unknown [16]. The diagnosis of MAR and NMAR requires a manual analysis of data patterns and the application of domain knowledge, while MCAR can be tested with the t-test or Little's test [17].

C. Univariate Time Series

This term refers to a time series that consists of single observations recorded sequentially over successive time periods. Although a univariate time series is usually considered as one column of observations, time is in fact an implicit variable [16].

D. Univariate Imputation Methods

Techniques capable of doing imputation for univariate time series can be roughly divided into three categories [16]:

- Univariate algorithms. These algorithms work with univariate inputs and commonly do not employ the time series features. Some of them are: mean, mode, median, random simple, last observed carried forward, etc.
- Univariate time series algorithms. Most of these algorithms are developed in section II, and some of them are: Missing Value Imputation by Weighted Moving Average [3] (SMA, LWMA and EWMA), Kalman, ARIMA, ARIMA-Kalman, Local Average of Nearest Neighbors [3] (LANN y LANN+), and Case Based Reasoning Imputation (CBRi) among others no cited in this work.
- Multivariate algorithms on lagged data. Commonly, multivariate algorithms cannot be used for univariate time series. However, using lags and leads it is possible to apply multivariate time series algorithms to a univariate time series and thus take advantage of features offered by multivariate algorithms.

E. Case Based Reasoning (CBR)

CBR is a nature inspired problem solving methodology [18]. It uses a solution that worked for a problem to solve a similar new problem, it's called reasoning by remembering.

The first principle of the CBR approach is: similar problems have alike solutions i.e. to solve a new problem [18], the existing problems and their solutions from the case base are retrieved and re-used.

The second principle is that the type of problems which an agent faces tends to repeat [18]. Thus, there is similarity between past and current problems or current and future problems. Therefore, it is worth to remember and reuse. This leads to construction of the case base which contains completely resolved problems and their respective solutions.

The complete Case Based Reasoning process is shown in Fig. 3.

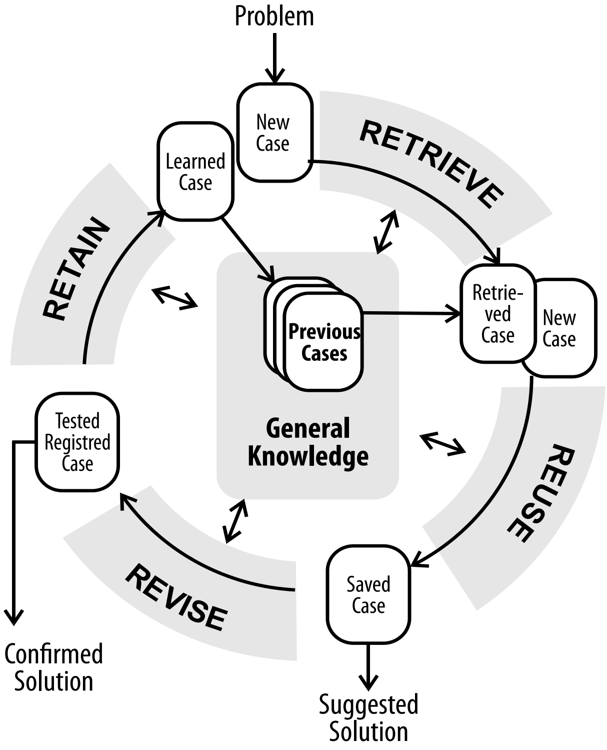


Fig. 3. CBR Process.

IV. CBRM

CBRm is inspired by the CBRi "unpublished" [4] algorithm that was designed for short-gaps imputation processes and that when applied to medium-gaps imputation processes can present problems of bias towards the prior value. Taking this assessment into consideration, CBRm begins the imputation from the middle of the series of consecutive NAs as shown in Fig. 2.

Fig. 4 shows the proposed CBR system within which implements the CBRm algorithm. Implementation process for CBRm is quite similar to CBRi "unpublished" [4], below are the required steps to implement it.

A. Time Series Selection

A time series of maximum daily temperatures corresponding to 9 years was chosen, from 2017-01-01 to 2015-12-31. These data correspond to the Punta de Coles weather station in Moquegua region (Peru) and were retrieved from the SENAMHI institutional repository at the following web link: https://www.senamhi.gob.pe/?&p = download-hydrometeorological-data.

B. *Case Base Implementation*

An algorithm was implemented to build the case base. The case base matrix consists of something similar to what is shown in Table I.

The algorithm in Javascript language to build the case base is shown in Table II “unpublished” [4]. This algorithm aims to create the matrix or case base (Q). It receives as arguments the empty Q matrix and a temperature vector, and returns as a result the matrix of cases Q.

TABLE. I. CASE BASE FROM 9-YEAR TIME SERIES

	..	**17.2**	**17.3**	**17.4**	**17.5**	**17.6**	**17.7**	**17.8**	**17.9**	**18.0**	**18.1**	**18.2**	**18.3**	**18.4**	**…**
…		16.8*16.4*17.2*17.4*17.5*17		16.6*17*16.6*17.2		17.4*17.4		17.4*17.6*19.4		16.6*18.6		18*18.2*17.6		17*18.8	
17.2															
17.3		15.8*17*17.2*18*17.6*17.1		17.6*17.1		17.2	19.2	17.6*19.4		18.2*17.6		17.8*17.2*17.6		18	
17.4		18.8						17.6							
17.5				18.2*17.4*17.8*17.8*17.8		18*17.8*17.2		16.8*19*19*17.8*19.1		17.8*17*17.8*18.2		18*18.8*18.4		16.6*18	
17.6												18			
17.7		15*17.4*17.8		18*17.9		17.4*18.8*17.4*17.8*17.5		18.6*17.4*17.6*17.6*17.6*17.8*17.2	17.8	17.4*17.6		18*18.2*18.4*18*18.8*18		17.4*18.8	
17.8					18.6										
17.9		16.8*18.2*17.2		17.2*18.2*17.8	17.8	17.8		17.6*18.2*17.8		18.2*18.2		17.8*19.2*18.4*18.4*18.4*18.2*18.4*18.1		18.2*18.2*18.8*18.7	
18.0														18.2	
18.1		18*17.4		17*17.8		17.4*18.2*18.4*17.2		16.4*18*18.6*18*18*18.2*18.4		17.8*17.6*18.8*17.8*17.8*18.2*18.2		18*18		17.8*18.6*19.4*19.4*18.8*18*	
18.2															
18.3		18.8		18.8*18.6		17.4		18*17.8		18.6*19.2*18.2*18.2		17.6*18*19.8*18.6*18.2		17.6*18.2	
…															

TABLE. II. ALGORITHM TO BUILD THE CASE BASE (Q)

```
function fillMatrix(Q,temv)
{    nQ=Q.length;
     for(i=0; i<nQ; i++)
     {    prior=temv[i];
          for(j=0; j<nQ; j++)
          {     next=temv[j];
                res=look4cases(prior,next);
                if(res!="")
                      Q[i][j]=res;
          }
     }
     return Q;
}
```

C. *CBRm Implementation*

According to Fig. 4, four blocks of code can be seen in the CBRm algorithm, and their detail can be seen in the code shown in Table III. The CBRm algorithm receives as inputs the time series with NA values and an array with the positions of each NA value.

As it shows in Fig. 4, for the first block of code that corresponds to the determination of the prior and next values that are required by the getMoreSimilar() function to extract the most similar case from the case base; these values are determined through the code between line 4 and line 18 using for this task the array of positions of the NA values.

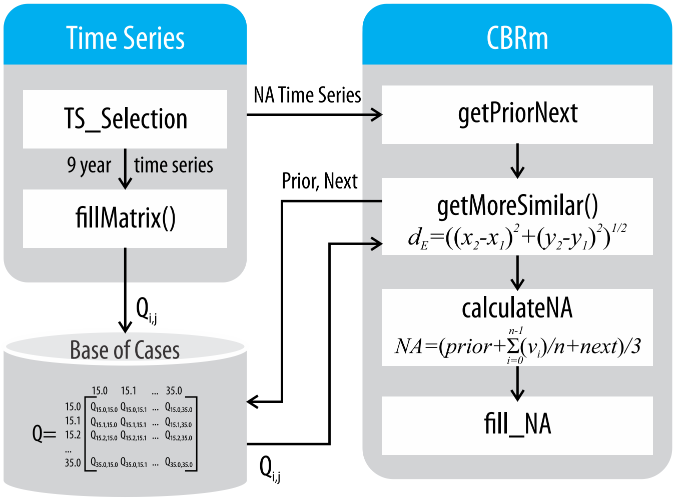

Time Series
TS_Selection
9 year
time series
fillMatrix()
$Q_{i,j}$
Base of Cases
$Q=\begin{array}{c|cccc} & 15.0 & 15.1 & \dots & 35.0 \\ 15.0 & Q_{15.0,15.0} & Q_{15.0,15.1} & \dots & Q_{15.0,35.0} \\ 15.1 & Q_{15.1,15.0} & Q_{15.1,15.1} & \dots & Q_{15.1,35.0} \\ 15.2 & Q_{15.2,15.0} & Q_{15.2,15.1} & \dots & Q_{15.2,35.0} \\ \dots & & & & \\ 35.0 & Q_{35.0,15.0} & Q_{35.0,15.1} & \dots & Q_{35.0,35.0} \end{array}$
NA Time Series
Prior, Next
$Q_{i,j}$
CBRm
getPriorNext
getMoreSimilar()
$d_E=((x_2-x_1)^2+(y_2-y_1)^2)^{1/2}$
calculateNA
$NA=(prior+\sum_{i=0}^{n-1}(v_i)/n+next)/3$
fill_NA

Fig. 4. CBR System.

TABLE. III. CBRm ALGORITHM

```
1.  function CBRm(tsna,pos)
2.  {     npos=pos.length;
3.        while(npos>0)
4.        {     nna=0;
5.              ini1=pos[0];
6.              fin1=pos[0];
7.              pini=0;
8.              pfin=pini;
9.              prior=parseFloat(tsna[pos[0]-1]);
10.             nav=tsna[ini1];
11.             while(nav=="NA")
12.             {     nna++;
13.                   fin1++;
14.                   pfin++;
15.                   nav=tsna[fin1];
16.             }
17.             next=parseFloat(nav);
18.             fin1--;
19.             data=getMoreSimilar(prior,next);
20.             dat=data.split("*");
21.             ndat=dat.length;
22.             s=0;
23.             for(k=0;k<ndat;k++)
24.                   s+=parseFloat(dat[k]);
25.             NA=(prior+(s/ndat)+next)/3;
26.             sNA=NA.toFixed(1);
27.             rna=nna%2;
28.             pna=Math.floor((ini1+fin1)/2);
29.             del=Math.floor((pini+pfin)/2);
30.             if(rna==0)
31.             {     m1=pna;
32.                   m2=pna+1;
33.                   tsna[m1]=smed;
34.                   tsna[m2]=smed;
35.                   pos.splice(del-1,2);
36.             }
37.             else
38.             {     tsna[pna]=sNA;
39.                   pos.splice(del,1);
40.             }
41.             npos=pos.length;
42.       }
43.       return tsna;
44. }
```

In the second block of code (line 19) the getMoreSimilar () function is called, this function implements a similarity search in the base of cases (Q) using the prior and next values determined in the previous code block, returning a string containing the values that will be used in the next code block. The getMoreSimilar() function implements Euclidean Distance according to equation (1) to determine the similarity between two points.

$$d_E = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (1)$$

In the third block of code between lines 20 and 26, the string returned by the getMoreSimilar () function is used and the NA value is calculated according to equation (2).

$$NA = \frac{(prior+\frac{\sum_{i=0}^{n-1}(Vi)}{n}+next)}{3} \qquad (2)$$

In the last block of code from line 27 to line 40, the NA value is filled with the value calculated according to the imputation process shown in Fig. 2. The process is repeated until the time series does not contain NAs values.

V. Results and Discussion

In this section, the performance of the proposed algorithm CBRm is compared with different techniques described in Section II, the comparative results are shown below.

According to Table IV, for a 15-days maximum temperature time series with 73.33% of NA values, the best techniques were LWMA and EWMA in the first place (RMSE 0.6941); second is LANN+ (RMSE 0.7077); and thirdly very close to the previous one is CBRm (RMSE 0.7083). For a percentage of 60% of NAs, the best technique was LANN+ (RMSE 0.6616); secondly LANN (RMSE 0.7187); and thirdly CBRi (RMSE 0.7461). For a percentage of 46.67% of NAs, the best technique was CBRi (RMSE 0.4140); second is LANN (RMSE 0.4423); and finally, thirdly EWMA (RMSE 0.4780).

It is important to highlight that for the first two cases, ARIMA Kalman produced no results.

A graphical comparison of CBRm RMSE with other techniques can be seen in Fig. 5.

TABLE. IV. COMPARISON WITH OTHER UNIVARIATE IMPUTATION TECHNIQUES (15 DAYS)

Technique	RMSE (NAs 73.33%)	RMSE (NAs 60%)	RMSE (NAs 46.67%)
CBRm	0.7083	0.8164	0.6152
CBRi	0.8575	0.7461	0.4140
LANN	0.8050	0.7187	0.4423
LANN+	0.7077	0.6616	0.6175
Hotdeck	0.9534	0.9189	1.0823
SMA	0.7323	0.8432	0.4928
LWMA	0.6941	0.8096	0.5209
EWMA	0.6941	0.7958	0.4780
ARIMA Kalman	NA	NA	0.5976

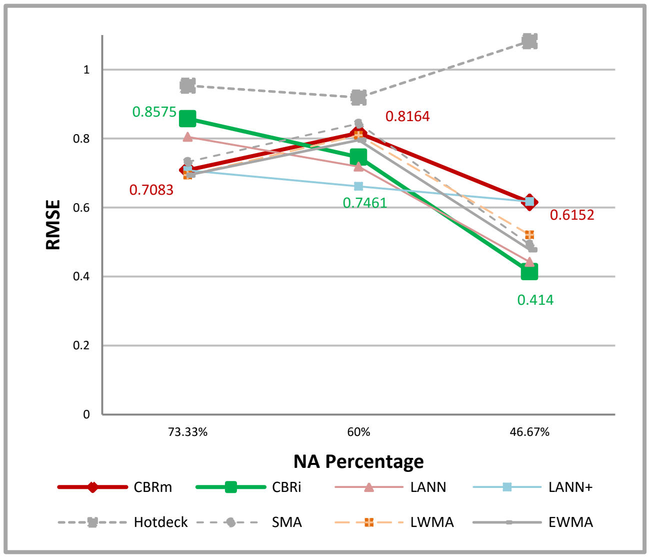


Fig. 5. Comparison with other Techniques (15 Days).

According to Table V, for 90-day time series with a percentage of 80% of NA values, the best technique was CBRm (RMSE 0.6844); second is LWMA (RMSE 0.7673); and thirdly EWMA (RMSE 0.7682). For a percentage of 65.55% of NAs, the best technique was SMA (RMSE 0.7035), followed by LWMA (RMSE 0.7083), and thirdly LANN+ (RMSE 0.7339). For a percentage of NAs of 54.44%, the best technique was LWMA (RMSE 0.8106), followed by SMA (RMSE 0.8403), and thirdly EWMA (RMSE 0.8535)

A graphical comparison of CBRm RMSE with other techniques can be seen in Fig. 6.

Also, CBRm was compared with two multivariate imputation techniques such as MICE and KNN. For this task, the data from the closest weather station to Punta de Coles, the Ilo station was used. In Table VI, the results are shown.

TABLE. V. COMPARISON WITH OTHER UNIVARIATE IMPUTATION TECHNIQUES (90 DAYS)

Technique	RMSE (NAs 80%)	RMSE (NAs 65.55%)	RMSE (NAs 54.44%)
CBRm	**0.6844**	0.8050	0.8968
CBRi	0.8086	0.8112	0.8905
LANN	0.8422	0.8198	0.9053
LANN+	0.8276	**0.7339**	0.8608
Hotdeck	1.4337	1.6323	1.4996
SMA (k=1)	0.8324	**0.7035**	**0.8403**
LWMA (k=4)	0.7673	0.7083	**0.8106**
EWMA (k=4)	**0.7682**	0.7456	**0.8535**
ARIMA Kalman	5.4275	6.7383	2.6836

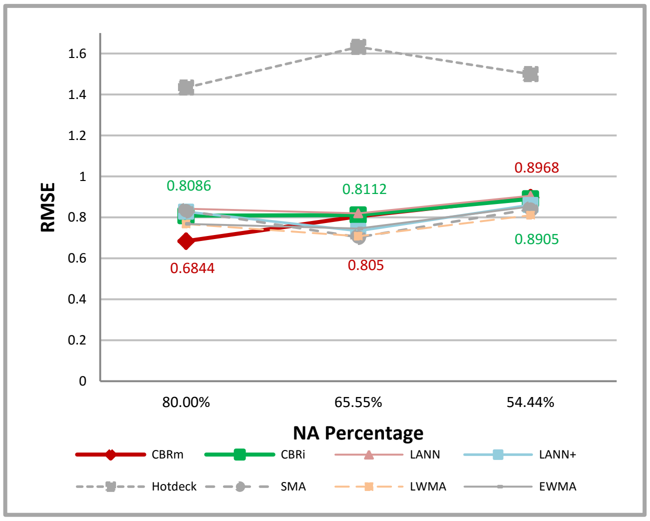


Fig. 6. Comparison with other Techniques (90 Days).

TABLE. VI. COMPARING WITH MICE AND KNN (90 DAYS)

Technique	RMSE (NAs 80%)	RMSE (NAs 65.55%)	RMSE (NAs 54.44%)
CBRm	0.6844	0.8050	0.8968
MICE	1.4063	1.3900	1.4714
KNN	1.0807	1.0751	1.2388

According to Table VI, the proposed CBRm outperformed the multivariate imputation algorithms KNN and MICE.

CBRi “unpublished” [4] despite the weaknesses mentioned, of the six problems proposed, as CBRm in two of them was among the best: in one of them it reached first place and in another it achieved third place.

LANN [3] for medium-gaps from 6 problems, in 2 of them he was among the best, it got second place twice. LANN+ [3] was a little better tan LANN, in 3 cases it was among the best getting the first, second and third place.

VI. CONCLUSION

In imputation processes of meteorological time series with medium-gaps (from 3 to 10 consecutive NAs), corresponding to time series of maximum temperatures, of the six proposed problems, in two of them CBRm was among the best: in one of them it reached the best performance and in another it achieved the third place.

Comparing CBRm with CBRi “unpublished” [4], of the 6 problems proposed in 3 cases CBRm outperformed CBRi “unpublished” [4] and in three other cases CBRi outperformed CBRm, so it is concluded that both techniques are good alternatives for the medium-gaps imputation process.

VII. FUTURE WORK

In this section, it is important to highlight the main limitation of CBRm, since it is based on historical data from the time series; it requires large amounts of historical data, something not always present in the known time series. A solution to this problem could be the complementation of the technique with other techniques until the case base has enough cases.

In the present study a medium-gaps imputation algorithm was proposed and analyzed, it would be interesting and important for future work to use gaps of 11 or more NA values (big-gaps). Also, it would be important to analyze the CBRm performance in other time series, for example in time series with no trend and no seasonality.

REFERENCES

[1] Rantou, “Missing Data in Time Series and Imputation Methods,” University of the Aegean, Samos, 2017.

[2] S. Moritz, A.Sardá, T. Bartz-Beielstein, M. Zaeffer, J, Stork, “Comparison of different methods for univariate time series imputation in R,” arxiv.org, 2015.

[3] A. Flores, H. Tito, C. Silva, “Local average of nearest neighbors: Univariate time series imputation,” International Journal of Advanced Computer Science and Applications, vol. 10, nº 8, 2019.

[4] A. Flores, H. Tito, C. Silva, “CBRi: A Case Based Reasoning-Inspired Approach for Univariate Time Series Imputation. Unpublished,” de IEEE Latin American Conference on Computational Intelligence, Guayaquil, Ecuador, 2019.

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

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

[7] K. Kaiser, O. Affuso, T, Beasley, D. Allison, “Getting carried away: A note showing baseline observation carried forward (BOCF) results can be calculated from published complete-cases results,” PMC US National Library of Medicine, 2012.

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

[9] T. Aljuaid, S. Sasi, “Proper imputation techniques for missing values in data sets,” de International Conference on Data Science and Engineering (ICDSE), Cochin, India, 2016.

[10] S. Moritz, “Package ImputeTS,” cran.r-project.org, 2019.

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

[12] A. Chaudhry, W. Li, A. Basri, F. Patenaude, “On improving imputation accuracy of LTE spectrum measurements data,” de Wireless Telecommunications Symposium, Phoenix, AZ, USA, 2018.

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

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

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

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

[17] R. Little, “A test of missing completely at random for multivariate data with missing values,” Journal of the American Statistical Association, vol. 83, nº 404, pp. 1198-1202, 1988.

[18] M. Khan, H. Hayat, I, Awan, “Hybrid case-base maintenance approach for modeling large scale case-based reasoning systems,” Human-centric Computing and Information Sciences, vol. 9, nº 9, 2019.

Performance Impact of Relay Selection in WiMAX IEEE 802.16j Multi-hop Relay Networks

Noman Mazhar[1]
Faculty of Computer Science and Information Technology
University of Malaya
Kuala Lumpur, Malaysia

Muhammad Zeeshan[2], Anjum Naveed[3]
School of Electrical Engineering and Computer Science
National University of Sciences and Technology (NUST)
Islamabad, Pakistan

Abstract—Worldwide Interoperability for Microwave Access network accepts the challenge of last mile wireless access for internet. IEEE 802.16 standard, commercially known as WiMAX provide wireless broadband experience to the end subscribers and challenges many wired solutions like Digital Subscriber Line (DSL) and cable internet. Wireless network has many inherent issues like coverage holes; capacity optimization and mobility are few of them. Adding relays to multi-hop WiMAX IEEE 802.16j network present an effective solution to address them to some extent but this amendment does not elaborate any algorithm regarding the relay selection and narrate no performance guarantees. In this work, we proposed linear model that fairly allocates wireless resources among subscribers in 802.16j network. A relay selection algorithm is also presented to optimally select nodes with higher signal-to-noise ratio as relay station for nodes with lower signal-to-noise ratio objectively maximize overall network capacity. This scheme further extends network coverage area and improves network availability. We also did extensive performance evaluation of the proposed linear model. Results show that optimal relays selection scheme do provide a substantial increase of up to 66% in overall network capacity in the fixed WiMAX network. This improvement is substantial at places where network condition is not optimal. Investigating the problem further leads to the conclusion that the relay selection criterion is the key to achieve maximum network capacity.

Keywords—WiMAX; multi-hop; wireless broadband; relay; SNR

I. Introduction

High speed internet access for the last mile has been a challenge over the years. Apart from the wired technology, inception of advanced coding schemes and antenna technology made wireless broadband a competitive solution. IEEE 802.16 Working Group (2004) commercially known as WiMAX initially formalized in 2001, however till 2004 standard based on [1] targeted only fixed applications and the standard was referred to these as fixed WiMAX. In December 2005 new amendment IEEE 802.16e [2] was launched which added a new dimension in WiMAX, mobility. Since then IEEE 802.16e-2005 forms the basis for the WiMAX solution for nomadic and mobile applications and is often referred to as mobile WiMAX. IEEE 802.16 [3] merges the fixed and mobile PHY and MAC capabilities of the network. Another amendment IEEE 802.16 [4] modified the physical layer and the MAC layer for inclusion of relays in the network known as IEEE 802.16j. The current version of IEEE 802.16 [5] has added further modification to support higher reliability networks.

Multi-hop wireless network started to get much focus in the telecommunication industry, due to rapid deployment and coverage enhancements. This also forces various product portfolios to upgrade to this technology in order to get major share of the industry in near future. There are many networks which provide multi-hop communication, but all of these have different MAC and PHY layer design due to constraint like power, mobility and transmission range. Augmenting to this, multi-hop wireless networks require complex routing algorithms. All this added to the limited use of the multi-hop networks. Introduction of relays in multi-hop wireless network have improved the network capacity and extend coverage area and simplify deployment scenario. According to Pabst R. et al. [6], relays provide communication to the nodes outside transmission range of the base station (BS), support in alleviating the range limitation for wireless broadband networks.

There are many practical scenarios as shown in the Fig. 1, where relays do play important role. Relays can be placed in different formations to enhance the system performance and coverage. In fixed infrastructure the relays are placed in stationary areas by the service provider to provide normal traffic by extenuating the loop holes and extending the coverage.

To provide relay based multi-hop support in WiMAX, a new amendment to the standard is made known as IEEE 802.16j and it defines two types of relays transparent and non-transparent relays. Transparent relays are used for improvement in capacity of overall network while the use of non-transparent relays extends the coverage area. The major difference between the two relay modes is the way framing information is transmitted. In transparent mode frame header information is not transmitted; while in non-transparent mode the frame contains the header information. The frame header encapsulates critical scheduling information which the nodes used to determine when it can transmit or receive. In relay modes there are two types of scheduling modes: centralized and distributed. Base Station (BS) control scheduling for all the nodes in centralized mode while in distributed mode relay station (RS) can take some scheduling decision for the nodes attach to it. But there are challenges faced by the system implementing 802.16j like frequency reuse, resource allocation, relay selection, performance and scheduling. IEEE

802.16j standard has enhanced BS and RS capabilities to meet multihop networks communication challenges. The standard does not discuss much about relay selection criteria. This issue has been left vendor specific. Similarly, no performance parameters are given while using relays in the WiMAX network.

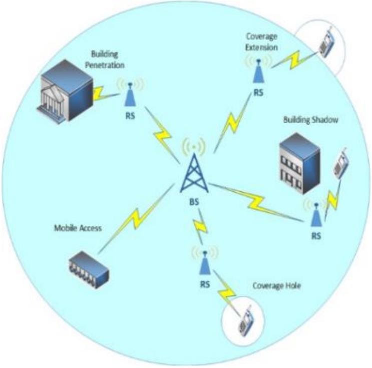


Fig. 1. Relay Application.

In this study we focus on the relay selection such that throughput of overall network is optimized in IEEE 802.16j network. In this context we propose linear model that fairly allocates the resources among subscribers in the range of BS. In fixed and nomadic infrastructure relays with centralized scheduling is the simplest and cost-effective solution to the multi-hop relay network. However distributed scheduling mode is more affective for mobile applications in order to handle coverage extension and other mobility issues. Therefore, the study focuses on centralized scheduling in fixed WiMAX network. We develop linear model using linear programming technique. We implemented the optimization model using algebraic modeling problem language (AMPL) and use simplex method for testing its results. We also develop a simulation program in c programing language to test the system model against IEEE 802.16-2004 based network without relays. Results show considerable throughput enhancement as compare to IEEE 802.16-2004. Further in this research we consider two main aspects: (i) relay selection scheme in IEEE 802.16j network, (ii) implications of relays on the overall network performance. Followings are the contributions of this work:

- We first propose a linear model to evaluate the performance of WiMAX network.
- We also develop relays selection algorithm for WiMAX multi-hop network and perform optimization of proposed linear model.
- We then perform comprehensive performance analysis of capacity gains after inclusion of relays in the WiMAX multi-hop network.
- Results show that proposed relay-based model out performs traditional WiMAX network in terms of capacity utilization.

The rest of the paper is organized as follows. In Section 2, related work has been elaborated. Section 3 presents the system model and problem formulation. Section 4 discusses the relay selection algorithm. The results and analysis of proposed algorithm are detailed in Section 5. Section 6 concludes this work.

II. Related Work

In recent years, some research tried to predict the performance of IEEE 802.16j network under different realizations and addressed issues like performance of network with inclusion of relays. Other issues include handling the loop holes and actual coverage extension.

In recent work on the performance evaluation of the WiMAX network, there are different attributes of wireless network that are exploited to enhance the network throughput. Like the work of Bonald T. et al. [10] is to determine the maximum throughput gain achievable under max-min fairness in which approximately equal performance is delivered to all subscribers. Like the Genc V. et al. [16] present an analytical modal that tries to enhance end-to-end throughput under max-min fairness constraint. Further by Genc V. et al. [17] extends the same model by incorporating the variable number of relays and transmits power; results show that about 55% to 125% of the throughput gain can be achieved subject to spatial reuse. But the MAXMIN algorithm has an issue that it may increase the throughput of some subscribers but starved the rest of the nodes. Another way is to compare the SNR for each path as done by Shrestha D. M. et al. [11] purposes more advance technique which use backward compatible signaling mechanism and introduce a centralize path selection algorithm based on ELT (estimated link throughput). ELT is based on available bandwidth and data rate for maximum throughput where data rate depends on SNR. ELT calculation for each path and signaling make base station processing more complex which compromises the overall performance of the network.

Similar work is purposed by Ann S. et al. [14] tries to find the route that reduces the latency and maximizes the network throughput. It is a centralized algorithm that makes decision based on SNR, available bandwidth, hop count but this scheme does uneven resource allocation and starves many subscribers. WiMAX provides two modes centralized and distributed scheduling. Some research uses centralized mode to achieve the performance as in Li D. et al. [18] focus on the maximum throughput in two hop fixed WiMAX network by using centralized scheduling scheme. Relay selection algorithm in such proposal is based on quality-of-service QoS, parameters including SNR and latency. But it did not warrant throughput maximization that can be achieved using this algorithm. One of the techniques is to measure the resource cost for the path, in order to select the most optimum path for the channel like the work in [20] proposes a path selection model based on radio resource cost (RRC) and the minimum the cost the best is the path. The result shows the throughput obtain using RRC scheme is much better as compare to the rest of scenarios. An interesting study use the adaptive technique according to the channel conditions as in Chang J. Y. et al. in [22] develop a deployment algorithm that work based on traffic and uniform clustering. This study takes both RS and BS placement to get the best throughput and coverage. The algorithm provides the vendors an adaptive deployment of BS and RS, further considering the environment.

Liu I. H. et al. [23] performs performance analysis and purpose scheduling algorithm that suggests that if the boundary for the zones are kept adaptive then the overall performance of the network can be enhanced notably. In support of this another study by Rajya Lakshmi et al. [28] improves the performance of MMR WiMAX networks and maintains the QoS flow requirements by using adaptive Zone size based on channel quality of each node. But if the number of subscribers increases with more variable channel conditions the purposed method will be expensive to use and may affect the overall performance. Even the power mode can be used for the performance improvement in the overall network as shown by the research by Paul A. et al. [24]; exploit sleep mode parameters of IEEE 802.16j to control or minimize the energy needs of the mobile node. They propose a scheme called energy saving centric uplink scheduling ESCS. This scheme does bandwidth allocation and sleep cycle decision algorithm. The results show that proposed ESCS provide more sleep time to the nodes, hence making them more energy efficient. Interference is the major constraint for the wide spread shared resources in WiMAX network among the devices and become challenge for the researchers. Therefore, resource allocation and sharing become an issue. In Mahb. et al develop a greedy centralized algorithm handle this issue. Another concept introduce relays in the network comes into the lime light. Initially relays were used to extend the network coverage. But IEEE 802.16j made relays more useful in the context of improving the performance and QoS of the network.

The main question is do the relays in network can improve the overall network performance or otherwise? For this research get focus on this aspect like study by Deb S. et al. [13] demonstrated the enhancement in throughput and range extension using relays at MAC layer. The results show an improvement in the median throughput of about 25%. But their analysis addresses only downlink scheduling however uplink results can be obtained similarly. Also, the relay selection for the respective subscriber remains unanswered in this work. A similar study by Genc V. et al. [15] shows that introduction of relays only improves throughput of approximately half of the coverage area of base station (BS).

Next question is, will the relay placement and selection play any role in the overall network performance. First, if we consider the relay placement lot of work shows the importance of relay placement in the context of the network overall performance like work by Chang C. Y. et al. [26] purposes a relay placement mechanism (RPM) that enhances the overall network capacity. Simulation results shows improvement in throughput and delays while maintaining a satisfactory level QoS. Also, performance evaluation in MMR networks is carried out by Ge Y. et al. [19] and they show that optimal relay selection can enhance the end-to-end throughput up to approximately 49%. Further Chang C. Y. et al. in [21] propose a relay deployment algorithm with an objective to minimize the hardware cost. Result shows improvement in the time slots allocation when relays are optimally placed but still compromise regarding hardware cost. In Chang J. Y. et al. in [27] proposed relay placement in order to improve network performance by minimizing the budget using rotational clustering algorithm. The results of the purpose scheme are compared to the RSPS and RPCC schemes. The average throughput and coverage ratio show profound improvement.

Work done by Arthi M. et al. [25] mainly focuses on the IEEE 802.16j network RS placement. The author emphasis the need of proper placement of RS in the network because improper placement may introduce multitude of issues like SNR in case of densely use of RS nodes, transmission delay in case of long spread RS nodes, Coverage holes, path selection in case of more than one option available for any node finally the link overloading such as many nodes requesting communication from one RS node. The paper selects the candidate positions for the RS using throughput-oriented method. Then develop an optimized model for the deployment of the RS in the network in order to minimize the overall budget. The model simulation result shows quite improvement in the capacity of the network. Especially when the BS an RS distance increases till a point where one can get the maximum throughput. These points become the deployment place for the RS. Relay selection is also important as the relay placement.

Relay selection plays very important role in enhancing the overall network performance as shown by the study Sreng V. et al. [7] that propose relay selection strategy based on physical distance and path-loss and conclude with the results that path-loss selection scheme consistently shows superior performance in comparison to those based on distance only. However, the study only discusses the coverage extension; impact on network throughput is not considered. Researchers devise different parameters for the selection of relays, some purpose methods such as selection based on distance from BS, simplest ways in this issue. Some studies find out that SNR (signal-to-noise ratio) should be considered as a main factor for the relay selection, therefore Hu H. et al. [8] concludes that station with maximal SNR in candidate relay stations should be chosen as relay. Some selection is based on the power criteria aspect considered by the researchers for the relay selection, the algorithm proposed by Hui T. et al. [9] choose the minimal total transmission power, set P1 and P2 for the power of the relay link and access link respectively, this find the complete path and relay will be selected automatically. Some study selects the different modes of relays for the performance improvement of the network. Such as Zhu V. et al. [12] did comprehensive performance evaluations of relays in WiMAX networks and finds out that the non-transparent relay station in IEEE 802.16j network do perform better with distributed scheduling as compare to the transparent relays. But non-transparent relays are expensive and complex, and therefore most research studies are done for transparent relays which are more economical and easier to deploy. Another relay selection criterion is the mode of working principle of the relays like amplify-forward (AF) and decode and forward (DF) relays as a Study by Swain Manoj et al. [29] propose relay selection scheme based on both AF and DF types of relays. The results show that the harmonic mean is better than min max scheme as far as SNR requirements for selection of relays is concern but for bandwidth performance min max supersedes harmonic mean. However, the results are simulated keeping fixed number of subscribers. In [30], we tried to maximize network capacity by optimizing relay selection.

In this work we consider the issues of performance and relay selection in 802.16j network, we propose model that makes relay selection with an objective to fairly maximize overall network capacity. Further we develop relay selection algorithm to ensure fair data rate for all the subscribers with in the range of BS. We explain the detail of our proposed system architecture and model in the next section.

III. IEEE 802.16J EVALUATION MODEL

In this section, we propose an optimization model for performance evaluation of IEEE 802.16j. Relays can be used to enhance network capacity of IEEE 802.16j network by increasing data rate for the subscriber and mitigating the coverage holes in the network. Simple scenario is shown in Fig. 2; it is assumed that the relay RS2 send data to SS2 and SS4 at much improved data rates due to better modulation and coding scheme causes increase in the network capacity. But here couple of questions arises; firstly, how much capacity increases in the network by using relays. Secondly what will be the relay node selection criteria? Since the IEEE802.16j left these issues vendor specific, therefore our study tries to answer these open questions. Next, we build a scenario to explain our model and demonstrate the selection criteria.

We consider a scenario in which the relay nodes are being selected based on modulation and coding scheme and we consider a system having tree topology as shown in Fig. 3. Since transparent relays assume that all the nodes can decode the framing information, therefore we assume that all the nodes are within the transmission range of base station. We have one base station and four subscribers. SS4 can communicate with the BS using four different routes; AB, CD, E and FG.

If we assume that the SS4 require fixed amount of data, for the requested demand the BS will allocate 30 slots via AB, 10 slots via CD, 20 slots via E and 15 slots via FG. The base station does this allocation based on signal to noise ratio between the base station and the nodes. In order to optimize the throughput of the network, our proposed model selects the least demanding route to fulfill the data rate requirement of SS4, which is in this case is CD with 10 slots. Rest of the routes is not considered. Since the allocation is done only for one frame at a time therefore the all the demand of the subscriber may not be fulfilled only in one frame however the allocation per frame will optimize the overall network throughput.

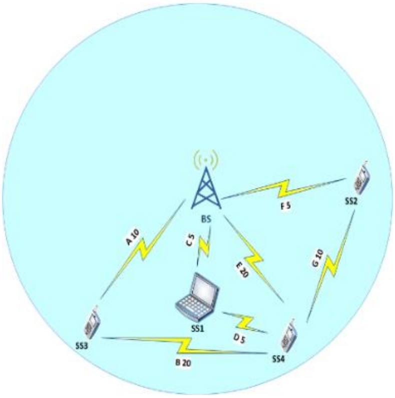


Fig. 2. Relay Network.

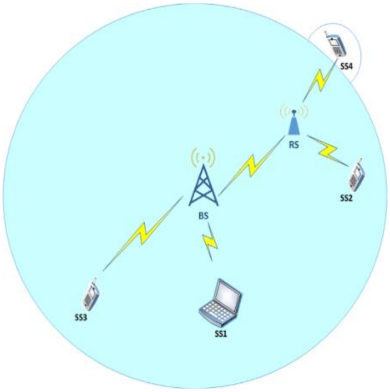


Fig. 3. Relay Scenario.

A. Parameters and Notations

We define some parameters as shown in Table I, such as S is the set of all subscribers within the coverage range of the base station; L is set of all the links between the nodes and with the base station in the system and q is total number of slots in the downlink frame. In WiMAX TDD frame comprises of slots; which is a unit of resource occupies space both in time and frequency. D_s is demand of each subscriber s, M_{ij} is data rate per slot against a specific modulation and coding scheme for the subscriber between node i and j. J is assumed to be base station node and finally K_s is a set of all the nodes except the node s itself.

B. Linear Programming Model

We employ linear programming approach therefore first we define our decision variable T_s which shows the number of slots allocated to each subscriber over a link. Since centralized scheduling and resource allocation is used there for BS has this responsibility. Where q is the number of slots allocated to the subscriber on the link e_{jks}. Here j is the base station, k belongs to the set of relays and s is among the set of subscribers. Therefore, the link e can be between base station and the subscriber or between the base station and the relay node.

TABLE. I. NOTATIONS

Parameter	Description
S	Set of all subscribers
L	Set of all links in the system
q	Total number of slots in the DL frame
d_s	Demand of subscriber in set S
m_{ij}	Data rate per slot against a specific MCS for the subscriber between node i and j
J	Base station node
k_s	Set of subscriber nodes act as relays attached to s in set S
e_{js}	Link between base station to the subscriber
e_{jk}	Link between the base station to the relay
e_{ks}	Link between the relay to the subscriber

C. Objective Function

Objective of this proposed model is to select the relay nodes such that the overall throughput of the network is maximized. Further the data rate allocated to the subscribers should be fair, so that the network resources may not consumed by few nodes in the network. We implemented the slot wise fairness as a constraint explained below. Objective function in Eq. (1), the $T_s e_{js}.m_{js}$, m_{js} is the modulation and coding scheme between base station and the subscriber, gives us the actual data rate in the form of data bits per slot for the given subscriber over the given link. This equation calculates the data rate for all the subscriber attaches directly to the base station. While $T_s e_{jk_s}.m_{jk_s}$ determines the data rate between the base station and the relay node, a relay node can be any node in the network except the requesting subscriber.

We used simplex method for the maximization function, initially the function allocate slot for the subscriber over the link between the base station and the subscriber, then further allocates the slots to the relay nodes for the given subscriber over the link between base station and the relays in the same frame. The function selects the relay nodes for the given subscribers if the number of slots allocated to the subscriber in the direct link to the base station is more than the number of slots via relay nodes. The equation $T_s e_{jk_s}.m_{jk_s}$ is multiplied by two and subtracted by $T_s e_{k_s s}.m_{k_s s}$ to show two hop communication. Since $T_s e_{k_s s}.m_{k_s s}$ shows the data rate allocated to the relays which need to be subtracted to get the actual data rate of the subscribers attach to that relay. We assume centralized scheduling; therefore, base station allocates slots both for the relay and subscriber. Proposed objective is to maximize the overall data rate of the network under given constraints. These constraints are explained one by one as below:

$$\sum_{s} T_s e_{js}.m_{js} + 2\sum_{s}\sum_{k_s} T_s e_{jk_s}.m_{jk_s} - \sum_{s}\sum_{k_s} T_s e_{k_s s}.m_{k_s s} \qquad (1)$$

$$j \in J, k_s \in K_s, s \in S$$

D. Demand Constraint

First constraint is obvious that the actual data rate allocated to each subscriber must be equal or less than its actual demand. Since the demand of the subscriber does vary therefore to keep the system safe from abnormal allocations this constraint plays a vital role. In Eq. (2), $T_s e_{js}.m_{js}$ shows data rate of the subscriber for direct link between base station and the subscriber. Similarly, $T_s e_{jk_s}.m_{jk_s}$ shows the data rate for the subscriber over the indirect link via relay. If the subscriber attaches to the base station directly then the relay link will have Ts with zero value, which further makes the data rate insignificant. Therefore, numerically we will be getting the data rate of only one link, which in our scenario is the direct one.

$$T_s e_{js}.m_{js} + \sum_{k_s} e_{jk_s}.m_{jk_s} \le d_s \qquad (2)$$

$$j \in J, k_s \in K_s, s \in S$$

E. Data Preservation Constraint

This constraint enforces the fact that amount of data received by the subscriber should be equal to the amount of data send by the base station to the relay node for the subscriber. As this is shown in Eq. (3), is the link between base station and the relay node for a subscriber. Similarly, the ekss shows the link between subscriber's relay nodes to that subscriber. In other words, it can be said that the data transmitted between the relay link and the access link for a particular subscriber should be same.

$$T_s e_{jk_s}.m_{jk_s} = T_s e_{k_s s}.m_{k_s s} \qquad (3)$$

$$j \in J, k_s \in K_s, s \in S$$

F. Resource Constraint

In the world of wireless data communication resources are the most precious and scarce thing. In proposed modal the resource we consider is the physical slots. Since q represent the total number of slots in a downlink frame. Therefore, it is checked that the total number of slots allocated to all the subscribers and relays should be equal or less than the total number of slots available i.e. q. In Eq. (4), $T_s e_{js}$ represent the number of slots allocated to the subscriber over the direct link between the base station and the subscriber. $T_s e_{jk_s}$ shows the number slots allocated to the relay nodes over the link between the base station and the relay node. $T_s e_{k_s s}$ here the Ts represents the number of slots allocated to the subscriber attach to the relay ks. Summation of all the allocated slots to all the nodes over all the links should be less or equal to q.

$$\sum_{s} T_s e_{js} + \sum_{s}\sum_{k_s} T_s e_{jk_s} + \sum_{s}\sum_{k_s} T_s e_{k_s s} \le q \qquad (4)$$

$$j \in J, k_s \in K_s, s \in S$$

G. Share Constraint

In order to make our model fair, following equations are constructed. Here in Eq. (5), u is the total number of subscribers in a network. $(q/u).m_{js}$ gives average data rate for each subscriber. $(q/u).m_{js} - d_s$ this equation checks that if the demand of subscriber is more than the average data rate available for it or vice versa. In case of more demand Eq. (5) will be executed. Eq. (6) ensures that each subscriber should get at least average data rate or more if possible. In the next section we compare our model with fixed WiMAX. The simulation results show substantial improvement in the overall network throughput. These results with given scenarios are further elaborated below:

$$(q/u).m_{js} - d_s \ge 0 \qquad (5)$$

$$j \in J, s \in S$$

$$T_s e_{js}.m_{js} + \sum_{k_s} T_s e_{jk_s}.m_{jk_s} - d_s = 0 \qquad (6)$$

$$j \in J, k_s \in K_s, s \in S$$

Where

$$(q/u).m_{js} - d_s \geq 0, \quad j \in J, s \in S$$

IV. Relay Selection Algorithm

In the proposed relay selection algorithm, all the parameters are defined in Table II. In this when our objective function got maximized the slots over all the links are assigned for a subscriber. First the slots required for the SS over direct link between BS and the SS are allocated then slots for the SS and all the other SS's other than BS are assigned. The algorithm selects link with minimum number of slots required to full fill the demand of the subscriber. This link can be direct between the BS and SS or BS to RS then to SS depend upon the number of slots required. This algorithm optimizes the slots assignment to all subscribers such that every user may get a minimum number of useful resources.

We use for loop since the function will iterate for each node in the network. The function first part that is $T_{as}.m_{js}$ assign slots to the nodes directly attached to the base station. While second part of the function $2(T_{rs}.m_{jk_s}) - T_{ks}.m_{k_s s}$ assign slots to the nodes that require relays for the communication. After that $T_{as}.m_{js} + T_{rs}.m_{jk_s} > d_s$ is a demand constraint that enforce the check that the total resource allocated to the node must not increase the demand requested by the node. Next $T_{rs}.m_{jk_s} \neq T_{ks}.m_{k_s s}$ ensure data integrity it means that the data send by the base station should be equal to the data received by the node. Following this is a resource constraint $(T_{as}.m_{js} + T_{rs}.m_{jk_s} + T_{ks}.m_{k_s s}) > q$ states that the total slots assign to all the nodes must not surpass the given maximum number of slots. Finally the share constraint insert fairness attribute in the algorithm $T_{as}.m_{js} + T_{rs}.m_{jk_s} - d_s \neq 0$ this equation forces the algorithm to assign some average resource to the node. Next section will discuss the results we obtain from this algorithm.

TABLE. II. Algorithm Parameters

Parameter	Description
J	Base Station
S	Set of all nodes
k_s	Set of all nodes except s
m_{ij}	Set of all MCS between Base station and nodes S
m_{jk_s}	Set of all MCS between Base station and nodes Ks
$m_{k_s s}$	Set of all MCS between nodes S and Ks
d_s	Demand of node s
T_{as}	number of nodes assign over the link BS and s node
T_{rs}	number of nodes assign over the link BS and Ks node
T_{ak}	number of nodes assign over the link Ks and s node

Algorithm 1. Relay Selection Algorithm

```
For Each s in S do
    Data rate for node s := max $T_{as}.m_{js} + (2(T_{rs}.m_{jk_s}) - T_{ks}.m_{k_s s})$
     if $(T_{as}.m_{js} + T_{rs}.m_{jk_s} > d_s)$ then
        return
     else
        continue
     end
     if $(T_{rs}.m_{jk_s} \neq T_{ks}.m_{k_s s})$ then
        return
     else
        continue
     end
     if $(T_{as}.m_{js} + T_{rs}.m_{jk_s} + T_{ks}.m_{k_s s}) > q$ then
        return
     else
        continue
     end
     if $(T_{as}.m_{js} + T_{rs}.m_{jk_s} - d_s \neq 0)$ then
        return
     else
        continue
     end
     if $(T_{rs} + T_{ks} < T_{as})$ then
        then k is realy node for s
     else
        No relay node for s
     end
end
```

V. Results and Discussion

For the simulation purpose we employed tree-based topology with BS at the root and centralized scheduling is assumed. Also, nodes only two hops away from the BS are taken under consideration. Since more than two hops have capacity issues depend upon subscriber density. We create different downlink scenarios to test the effectiveness of our proposed model. The scenario we assume is such that on OFDMA PHY using frequency 3.5GHz, 20MHz channel bandwidth, size is 2048 FFT, contain 1440 data subcarrier and 1/8 cyclic prefix. In addition to this other parameter like noise figure and thermal noise at transmitter are adjusted at 13db and -174dbm respectively. The frame size is assumed to be 20ms. The DL sub frame is about uses 50 of the total frame sizes.

Major parameter that we take under consideration for analysis purpose is SNR [signal to noise ratio]. SNR value determines the quality of a link ranges from 1-26, if its value is high then link is good and can transmit data at higher rate because of better modulation and coding scheme and vice versa. Our focus in this simulation is to cover all the possible scenarios that come across in networks. Table III shows the quantitative values of SNR against which the respective modulation and coding scheme is selected. In this we see that as the values of SNR increases, the better MCS are mapped against them.

TABLE. III. SIGNAL TO NOISE RATIO

Modulation	Coding Scheme	Receiver (SNR)
BPSK	1/2	3.0
QPSK	1/2	6.0
QPSK	3/4	8.5
16-QAM	1/2	11.5
16-QAM	3/4	15.0
64-QAM	2/3	19.0
64-QAM	3/4	21.0

A. Scenario 1: Random SNR for All Subscribers

First case we consider random value of SNR for all the subscribers as shown in Fig. 4. The graph shows throughput, at y-axis, achieved for different number of subscribers, shown at x-axis. In this scenario, we consider thirty subscribers against different demands. As we increase the number of users, we also vary the demand for the subscriber, but this demand remains same for all the users under consideration. The results show great increase initially as the users were less and demand was nominal too, as compare to the WiMAX network without relays. However, as the user demand increases the percentage decreases but remains up to 66%. Scenario in which we consider the demand of the entire subscriber fixed, and by only varying the SNR for different subscribers, we can see that the performance of our model represented as J is giving up to 66% increase compare to fixed WiMAX.

B. Scenario 2: 90% Subscribers with Excellent SNR

In another experiment we test a scenario in which 90% of subscribers are assumed to have SNR 15-26db, a much better channel condition. While the rest of 10% has poor channel condition or have low values of SNR. Initially we increase the number of users at the same rate as described in scenario 1. In addition to this the demand increases simultaneously. Only difference here is that the SNR is not random, but we have fixed the SNR to certain percentage of users.

Here as shown in Fig. 5, model is still able to get 8% increase of network throughput as compare to the WiMAX without relays. Since in this case if the BS coverage is good then the relays will not make much difference, however in cases like shadowing or dark spots our model still prove to be much more efficient the plain WiMAX network. The results again show that model performance is 8% more than the S.

C. Scenario 3: 10% Subscribers with Excellent SNR

Now in this experiment we swap the scenario and only 10% of subscriber have best SNR values between 15db to 26db while the rest of subscriber have SNR values between 0db to 14db. The results shown in Fig. 6 are quite convincing. This time model output shows 63% improvement in the context of throughput as compared to WiMAX without relays. Here again we have same number of subscribers which we increase gradually with different demands. Here we see the true benefit of relays in the network when the BS coverage is not good then our model can play significant role in improving the overall throughput of the network by giving alternate routes to the subscribers with low or poor SNR values.

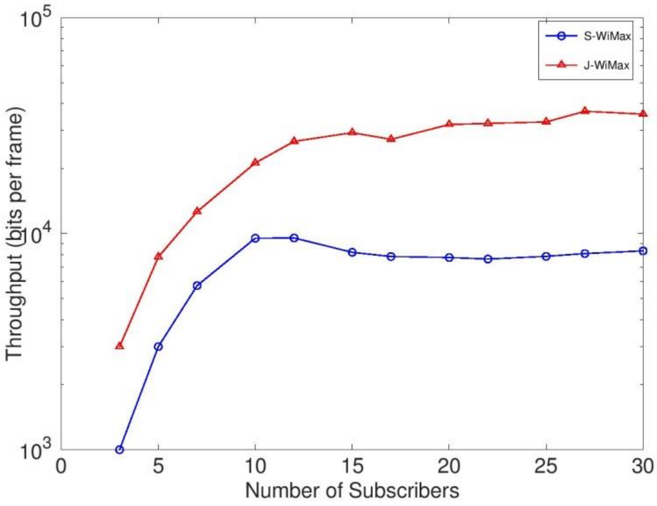


Fig. 4. Random SNR for all Subscribers.

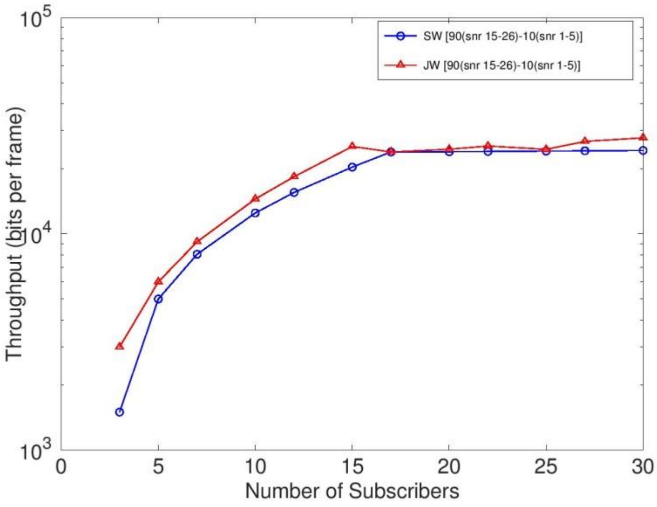


Fig. 5. 90% Subscribers with Excellent SNR.

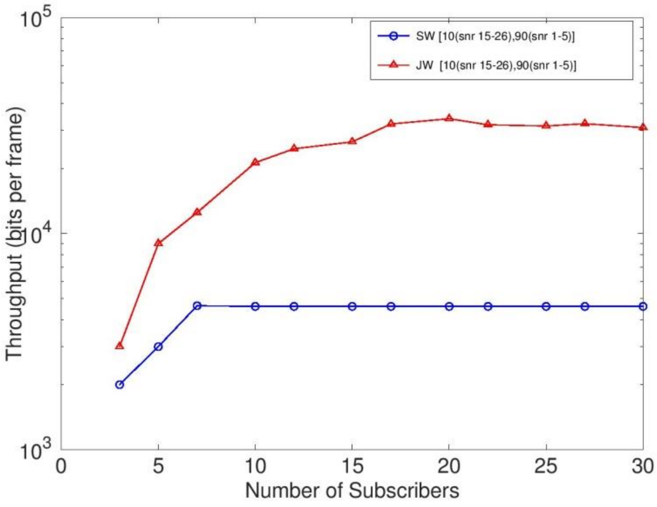


Fig. 6. 10% Subscribers with Excellent SNR.

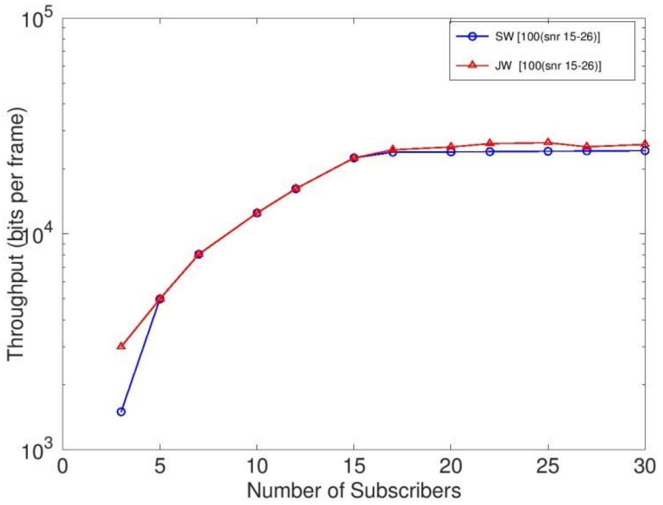


Fig. 7. 100% Subscribers with Excellent SNR.

D. Scenario 4: 100% Subscribers with Excellent SNR

Now we start to check for the extreme cases, we gave all the subscribers best SNR values 15-26db, again we increase the subscribers and demand as we did in the previous experiments. Here the results shown in Fig. 7 tell us that we can have throughput improvement by 4% using our model as compared to the WiMAX without relays. Again, we see that if we assume the network coverage to be best for all the subscribers then again the rule of relays remain quite trivial, and the results are more or less like the previous experiment as shown in Fig. 5.

E. Scenario 5: 100% Subscribers with Poor SNR

This scenario considers all subscribers with lowest SNR values 1-5. In this scenario again we increase the number of subscribers and the demand gradually, but the results shown in Fig. 8 are not much convincing. In this the results shows that if there are small number of subscribers then our model does perform much better than WiMAX without relays, however as the number of subscribers increases then there is not much difference between the WiMAX network using our model and the WiMAX network without the relays. Again, this shows that in extreme scenarios the behavior of network throughput is not much changed using relays. But such scenarios are quite difficult to get established. Even in these extreme scenarios our model still extracts some efficiency where ever and whenever is possible, therefore it is much better to use relays where ever it is feasible.

F. Scenario 6: 100% Subscribers with Good SNR

In another scenario we assume good channel conditions for all the subscribers. Now this is somewhat more realistic scenario. Here again, the number of users is the same and increased gradually while the demand got varied too. The SNR for all the subscriber is between 10-20db which automatically end up having much improved modulation and coding scheme. The results in Fig. 9 again show that the results shown by our model against the simple WiMAX without relays are 22% better. And the results got better as the number of users increases and demand of the user increases.

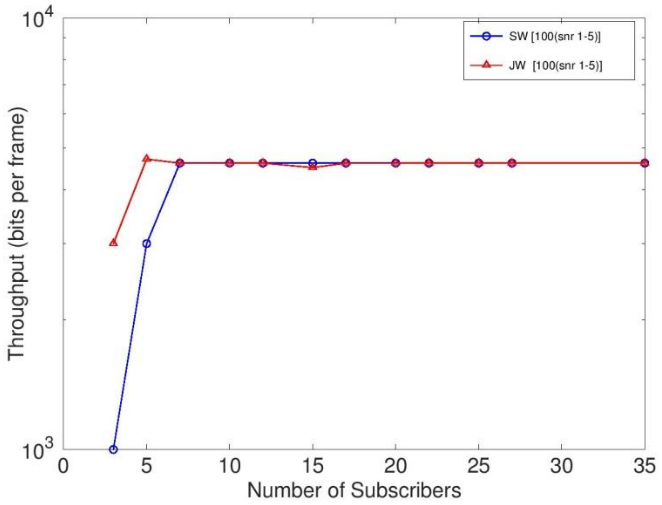


Fig. 8. 100% Subscribers with Poor SNR.

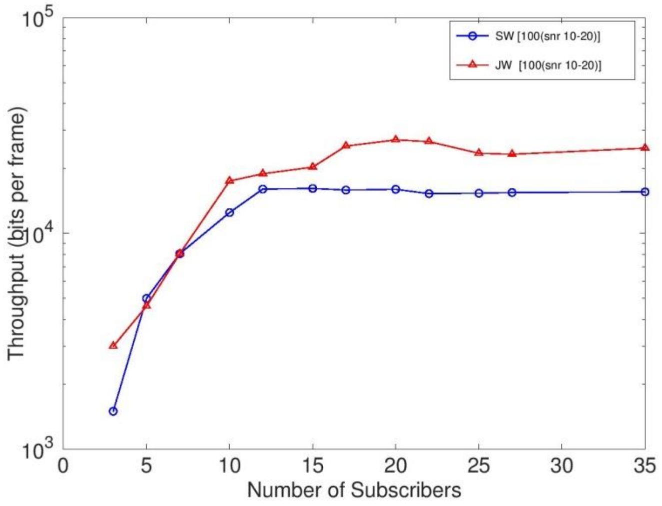


Fig. 9. 100% Subscribers with Good SNR.

VI. Conclusion

This work proposes a linear model for the IEEE 802.16j network to study the overall network capacity gains. The study helps to understand the performance effects after deploying the relays in WiMAX network. Mainly, multiple aspects of network are taken under consideration like performance changes under certain conditions and relay placements implication on overall network capacity. The study developed a linear modal for capacity evaluation and relay placement algorithm for its analysis. The results show comprehensive increase in throughput up to 66% in overall network capacity using relays in a fixed WiMAX network. In our future work, we are currently working to incorporate relay selection while satisfying quality of service (QoS). Different aspects of QoS can be explored to see how much difference it will create compared to other schemes.

References

[1] IEEE 802.16 Working Group. Ieee standard for local and metropolitan area networks-part 16: Air interface for fixed broad-band wireless access systems. IEEE Std. 802.16-2004. 2004.

[2] IEEE LAN/MAN Standards Committee. IEEE Standard for local and metropolitan area networks Part 16: Air interface for fixed and mobile broadband wireless access systems amendment 2: Physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum 1. IEEE Std 802.16 e-2005. 2006.

[3] IEEE 802.16 Working Group. Ieee standard for local and metropolitan area networks-part 16: Air interface for fixed broad-band wireless access systems. IEEE Std. 802.16-2004. 2004.

[4] IEEE 802.16 Working Group. IEEE standard for local and metropolitan area networks, part 16: Air interface for broadband wireless access systems, amendment 1: Multi-hop relay specification. IEEE Standard 802.16 j-2009. 2009.

[5] Pareit D, Lannoo B, Moerman I, Demeester P. The history of WiMAX: A complete survey of the evolution in certification and standardization for IEEE 802.16 and WiMAX. IEEE Communications Surveys & Tutorials. 2011 Oct 13;14(4):1183-211.

[6] Pabst, Ralf, Bernhard H. Walke, Daniel C. Schultz, Patrick Herhold, Halim Yanikomeroglu, Sayandev Mukherjee, Harish Viswanathan et al. "Relay-based deployment concepts for wireless and mobile broadband radio." IEEE Communications Magazine 42, no. 9 (2004): 80-89.

[7] Sreng, Van, Halim Yanikomeroglu, and David D. Falconer. "Relayer selection strategies in cellular networks with peer-to-peer relaying." In 2003 IEEE 58th Vehicular Technology Conference. VTC 2003-Fall (IEEE Cat. No. 03CH37484), vol. 3, pp. 1949-1953. IEEE, 2003.

[8] Hu, Huining, Halim Yanikomeroglu, David D. Falconer, and Shalini Periyalwar. "Range extension without capacity penalty in cellular networks with digital fixed relays." In IEEE Global Telecommunications Conference, 2004. GLOBECOM'04., vol. 5, pp. 3053-3057. IEEE, 2004.

[9] Hui, Tian, Gu Xuelin, and Zhang Ping. "The impact of relaying strategies on the performance in cellular system." In IEEE International Symposium on Communications and Information Technology, 2005. ISCIT 2005., vol. 2, pp. 1404-1407. IEEE, 2005.

[10] Bonald, Thomas, Laurent Massoulié, Alexandre Proutiere, and Jorma Virtamo. "A queueing analysis of max-min fairness, proportional fairness and balanced fairness." Queueing systems 53, no. 1-2 (2006): 65-84.

[11] Shrestha, Deepesh Man, Sung-Hee Lee, Sung-Chan Kim, and Young-Bae Ko. "New approaches for relay selection in IEEE 802.16 mobile multi-hop relay networks." In European Conference on Parallel Processing, pp. 950-959. Springer, Berlin, Heidelberg, 2007.

[12] Zhu VM, Viorel VD. Multihop relay extension for WiMAX networks—overview and benefits of IEEE 802.16 j standard. Fujitsu Sci. Tech. J. 2008 Jul;44(3):292-302.

[13] Deb, Supratim, Vivek Mhatre, and Venkatesh Ramaiyan. "WiMAX relay networks: opportunistic scheduling to exploit multiuser diversity and frequency selectivity." In Proceedings of the 14th ACM international conference on Mobile computing and networking, pp. 163-174. ACM, 2008.

[14] Ann, Sojeong, Kyung Geun Lee, and Hyung Seok Kim. "A path selection method in IEEE 802.16 j mobile multi-hop relay networks." In 2008 Second International Conference on Sensor Technologies and Applications (sensorcomm 2008), pp. 808-812. IEEE, 2008.

[15] Genc, Vasken, Seán Murphy, and John Murphy. "Performance analysis of transparent relays in 802.16 j MMR networks." In 2008 6th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks and Workshops, pp. 273-281. IEEE, 2008.

[16] Genc, Vasken, Sean Murphy, and John Murphy. "An interference-aware analytical model for performance analysis of transparent mode 802.16 j systems." In 2008 IEEE Globecom Workshops, pp. 1-6. IEEE, 2008.

[17] Genc, Vasken, Seán Murphy, and John Murphy. "Analysis of transparent mode IEEE 802.16 j system performance with varying numbers of relays and associated transmit power." In 2009 IEEE Wireless Communications and Networking Conference, pp. 1-6. IEEE, 2009.

[18] Li, Dandan, and Hao Jin. "Relay selection in two-hop IEEE 802.16 Mobile Multi-hop Relay networks." In 2009 First International Workshop on Education Technology and Computer Science, vol. 2, pp. 1007-1011. IEEE, 2009.

[19] Ge, Yu, Su Wen, and Yew-Hock Ang. "Analysis of optimal relay selection in IEEE 802.16 multihop relay networks." In 2009 IEEE Wireless Communications and Networking Conference, pp. 1-6. IEEE, 2009.

[20] Mach, Pavel, Robert Bestak, and Zdenek Becvar. "Optimization of association procedure in WiMAX networks with relay stations." Telecommunication Systems 52, no. 3, 2013.

[21] Chang, Chih-Yung, and Ming-Hsien Li. "A placement mechanism for relay stations in 802.16 j WiMAX networks." Wireless networks 20, no. 2, 2014.

[22] Chang, Jau-Yang, and Ya-Sian Lin. "A clustering deployment scheme for base stations and relay stations in multi-hop relay networks." Computers & Electrical Engineering 40, no. 2, 2014.

[23] Liu, I-Hsien, Chuan-Gang Liu, Chien-Tung Lu, Yi-Tsen Kuo, and Jung-Shian Li. "A multi-hop resource scheduling algorithm for IEEE 802.16 j relay networks." Computer Networks 67, 2014.

[24] Paul, A., Anagha, P., and Umaparvathi, M. "Energy Efficient Scheduling For Wimax Network", International Journal of Software & Hardware Research in Engineering, 3(3), 2015.

[25] Arthi, M., Jimy Jose Joy, P. Arulmozhivarman, and K. Vinoth Babu. "An efficient relay station deployment scheme based on the coverage and budget constraints in multi-hop relay networks." In 2015 International Conference on Communications and Signal Processing (ICCSP), pp. 0124-0128. IEEE, 2015.

[26] Chang, Chih-Yung, Chao-Tsun Chang, Tzu-Chia Wang, and Ming-Hsien Li. "Throughput-enhanced relay placement mechanism in WiMAX 802.16 j multihop relay networks." IEEE systems journal 9, no. 3, 2014.

[27] Chang, Jau-Yang, and Yun-Wei Chen. "A relay station deployment scheme with a rotational clustering algorithm for multi-hop relay networks." In 2016 International Conference on System Science and Engineering (ICSSE), pp. 1-4. IEEE, 2016.

[28] Lakshmi, L. Rajya. "Adaptive Zone Size Selection Method for IEEE 802.16 j Mobile Multihop Relay Networks." Wireless Personal Communications 97, no. 4, 2017.

[29] Swain, Chaudhuri Manoj Kumar, and Susmita Das. "Study and Impact of Relay Selection Schemes on Performance of an IEEE 802.16 j Mobile Multihop Relay (MMR) WiMAX Network." In Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, pp. 491-499. Springer, Singapore, 2018.

[30] N. Mazhar, “802.16j Network Performance Analysis and Relay Selection” MS. dissertation, Dept. Computing, SEECS, NUST Univ., Islamabad, Pakistan, 2011.

Evaluating Factors for Predicting the Life Dissatisfaction of South Korean Elderly using Soft Margin Support Vector Machine based on Communication Frequency, Social Network Health Behavior and Depression

Haewon Byeon[1]
Department of Speech Language Pathology
Honam University, Gwangju, Republic of Korea

Seong-Tae Kim*[2]
Department of Speech Language Pathology
Dongshin University, Naju, Republic of Korea

***Abstract*—Since health and the quality of life are caused not by a single factor but by the interaction of multiple factors, it is necessary to develop a model that can predict the quality of life using multiple risk factors rather than to identify individual risk factors. This study aimed to develop a model predicting the quality of life based on C-SVM using big data and provide baseline data for a successful old age. This study selected 2,420 elderly (1,110 men, 1,310 women) who were 65 years or older and completed the Seoul Statistics Survey. The quality of life satisfaction, a binary outcome variable (satisfied or dissatisfied), was evaluated based on a self-report questionnaire. This study performed a Gauss function among the SVM algorithms. To verify the predictive power of the developed model, this study compared the Gauss function with the linear algorithm, polynomial algorithm, and sigmoid algorithm. Additionally, C-SVM and Nu-SVM were applied to four kernel algorithm types to create eight types, and prediction accuracies of the eight SVM types were estimated and compared. Among 2,420 subjects, 483 elderly (19.9%) were not satisfied with their current lives. The final prediction accuracy of this SVM using 625 support vectors was 92.63%. The results showed that the difference between C-SVM and Nu-SVM was negligible in the models for predicting the satisfaction of life in old age while the Gaussian kernel had the highest accuracy and the sigmoid kernel had the lowest accuracy. Based on the prediction model of this study, it is required to manage local communities systematically to enhance the quality of life in old age.**

Keywords*—*C-SVM; communication frequency; life satisfaction; social network; quality of life

I. Introduction

Globally, the proportion of the elderly population is rapidly increasing due to economic growth, the advancement of medical technologies, and improved living standards. The aging speed of South Korea, in particular, is much faster than other countries in Europe, North America, Oceania, and Africa. As of 2018, the number of the elderly aged 65 or over was 7.38 million (14.3% of the total population), indicating that one out of 7 people in the population is a senior citizen [1]. If this trend persists, the proportion of the elderly will reach 20.8% in 2026 and South Koreans will enter a post-aged society.

Since nuclearized families, decreased socioeconomic capacities, and chronic degenerative diseases such as dementia have increased, the quality of life in old age is lower than that of the young and middle-aged people [2]. The elderly are particularly vulnerable to diseases. Recent studies [3, 4, 5] reported that communication issues (e.g., dementia and aphasia) and swallowing problems as well as physical problems (e.g., degenerative joint diseases) adversely affected the quality of life in old age. The physical aging and chronic diseases that people experience in old age shrink the elderly psychologically and deprive them of positive emotions [6]. Additionally, psychological aging decreases mental and neurological functions such as sensation and intelligence and it also leads to emotional changes such as anxiety and depression [6]. Furthermore, the elderly often suffer from various difficulties such as aggravated health, disabilities, higher psychological alienation and loneliness, decreased economic ability, reduced the social role, and declined informal network due to diverse factors including physical aging, chronic diseases, and loss of spouse [7].

The satisfaction of life is a subjective evaluation regarding the satisfaction and perception of one’s current life [8]. Moreover, this concept is widely used in gerontology and geriatrics studies [8]. An individual’s quality of life indicates the well-being state containing the concepts of satisfaction, happiness, and positive emotions and includes physical factors, mental factors, social factors, and personal achievements [9]. In other words, the quality of life is an index that reflects not only physical factors but also mental factors such as health, occupational factors, social interaction, happiness, and satisfaction. The satisfaction of life is an important issue in gerontology because the satisfaction of life is not obtained by the general conditions of the elderly but formed by the interaction between the physical environment and the social environment that are directly experienced by them.

*Corresponding Authors.

Previous studies [10, 11, 12] have identified a variety of factors affecting the quality of life in old age such as sociodemographic factors (e.g., gender, age, income level, and educational level), marital status, and chronic diseases. Wada et al. [13] reported that the subjective satisfaction of life correlated with physical functions and health. Recent studies showed that social networks and the frequency of contacting people around them also influenced the quality of life in old age [14, 15]. The elderly had higher satisfaction in life when they contacted family, relatives, friends, and neighbors more frequently and when they had various activities such as using senior citizen centers and community welfare centers [16].

Since health and the quality of life are caused not by a single factor but by the interaction of multiple factors, it is necessary to develop a model that can predict the quality of life using multiple risk factors rather than to identify individual risk factors [17]. Nevertheless, previous studies [18, 19] evaluating the risk factors affecting the quality of life in old age just aimed to explore individual risk factors using the generalized linear model. These studies [18, 19] mostly used logistic regression models to identify predictors. However, since regression models are used to predict the strength of the relationship between independent variables and dependent variables, this method is not appropriate to discover new predictors.

Recently, supervised learning algorithms such as the support vector machine (SVM) have been widely used in social science as a method of identifying complex factors associated with diseases and health problems [20]. Numerous studies [21, 22] have reported that the SVM has higher prediction power (accuracy) in classifying or predicting binary data than decision tree based machine learning or logistic models. Particularly, [23] showed that C-SVM, a transformation algorithm of the SVM, performs better because it makes the classification margin to classify two categories a serpentine nonlinear shape. This hyperparameter technique is drawing attention as a way to increase prediction power.

To date, it has not been tried to develop a prediction model reflecting health habits, subjective health, communication frequency, and social network, in addition to demographic factors, based on the supervised learning algorithm. This study aimed to develop a model predicting the quality of life based on C-SVM using big data and provide baseline data for a successful old age. Construction of our study is as follows. Section II explains database and analyzed variables and Section III defines C-SVM and explains the procedure of model development. Lastly, Section IV presents conclusion for future studies.

II. Methods and Materials

A. Target Subjects

The data source of this study was the 2015 Seoul Statistics Survey. The Seoul Statistics Survey was conducted with the permission of the Statistics Korea in order to measure the welfare level of Seoul's resident population, track the changes in the welfare level by sector and year, and secure basic data for establishing the welfare policies of Seoul. The purposes of the Seoul Welfare Panel Survey were to (1) measure the welfare level of the resident population of Seoul and track the changes in welfare level by year; (2) estimate the demand for welfare services by identifying the size and status of the socially vulnerable class; (3) measure the effects of the program by evaluating the accessibility and satisfaction of citizens for the existing welfare service programs; (4) present baseline data for providing comprehensive welfare program through the above survey and measurement; and (5) pave the way to revitalize social science studies based on quantitative data by providing statistical data to social welfare researchers. The Seoul Statistics Survey contains items related to the quality of life such as income, consumption, savings, debt, assets, living conditions, health, housing, elderly support, child education, disability and rehabilitation, welfare services, cultural lives, and social participation. This study used computer assisted personal interviewing for surveying that trained investigators visited the homes of subjects and conducted face-to-face interviews using laptops. This study selected 2,420 adults (1,110 men, 1,310 women) who were 65 years or older and completed the Seoul Statistics Survey as the final subjects.

B. Measurements of Variables

The definitions of the variables measured in this study are presented in Table I. The quality of life satisfaction, a binary outcome variable (satisfied or dissatisfied), was evaluated based on a question, “Are you satisfied with your life?”. Explanatory variables included gender, age (i.e., 65-74 years old and 75 years or older), the highest level of education (i.e., below elementary school, middle school, high school, and equal to or higher than collage graduation), monthly mean total household income (i.e., <2 million KRW, 2 million KRW⩽ and <4 million KRW, and 4 million KRW⩽), marital status (i.e., living with a spouse, married but not living with a spouse, and not married), current employment status (i.e., employment and unemployment), drinking frequency (i.e., less than once a week and more than once), smoking (i.e., smoking and non-smoking), walking per week (i.e., equal to or more than two days and less than one day), subjective health status (i.e., good, normal, and bad), depression (i.e., yes and no),, disease or accident experience in the past two weeks (i.e., yes and no), the frequency of meeting a neighbor (i.e., less than once a month and more than twice a month), and frequency meeting a relative (i.e., less than once a month and more than twice a month). Depression was measured using the Short Form Geriatric Depression Scale (SGDS) [24]. SGDS is easy to test the depression of the elderly, does not take much time, and is highly valid [24]. For each question, 'Yes' was given 1 point, and 'No' was given 0 points. The total depression score ranged from 0 to 15. The threshold for depression was 8, and a higher score means more severe depression. The reliability of SGDS in this study was evaluated by Cronbach's α (.881).

TABLE. I. THE DEFINITIONS OF THE VARIABLES MEASURED IN THIS STUDY

Category	Factor	Measurement
Demographic characteristics	Gender	Male, Female
	Age	65-74, 75+
	Income (Monthly mean total household income)	<2 million KRW, 2 million KRW≤ and <4 million KRW, and 4 million KRW≤
	Marital status	Not married, Married but not living with a spouse, Living with a spouse
	Education level	Below elementary school, Middle school, High school, Equal to or higher than collage graduation
	Current employment status	Employment, Unemployment
Health behavior	Current Smoking	No, Yes
	Drinking frequency	Less than once a week, More than once
	Walking per week	Less than one day, Equal to or more than two days
Social Network / Communication Frequency	Frequency of meeting a neighbor	More than twice a month, Less than once a month
	Frequency meeting a relative	More than twice a month, Less than once a month
Health status	Disease or accident experience in the past two weeks	No, Yes
Depression	Short Form Geriatric Depression Scale(SGDS)	No, Yes
Life satisfaction	Subjective Life Satisfaction	Satisfied, not satisfied

III. ANALYSIS METHODS

A. Support Vector Machine

The model for predicting the quality of life in old age was developed using the support vector machine (SVM). The SVM is a machine learning algorithm that finds the optimal decision boundary, linear separation, that divides hyperplane optimally by transforming the learning data to a higher dimension through non-linear mapping [25]. For example, A=[a,d] and B=[b,c] have non-linearly separable characteristics in two-dimension. When they are mapped to three-dimension, it has a linearly separable characteristic. Therefore, when an appropriate nonlinear mapping is used to a sufficiently large dimension, a dataset, which has two classes, can always be separated in the hyperplane. The SVM is very accurate because it can model the complex nonlinear decision-making domain, and it tends to have less over-fitting possibilities than other models, which is an advantage [25].

B. C-SVM Algorithm

The hyperplane of the SVM was used to classify linear forms. However, the use of kernel functions allows conducting nonlinear as well as the linear classification for a complex dataset (Fig. 1) [26]. Fig. 2 shows the transformed data (right) by mapping the actual data (left) to the feature space through a kernel function. The figure on the right reveals the process of deriving a linear hyperplane by conducting the SVM. C-SVM is often used as a criterion for classifying nonlinearities in a complex dataset (Fig. 3).

This study performed the radial basis function (a Gauss function) among the SVM algorithms. To verify the predictive power of the developed model, this study compared the Gauss function with the linear algorithm, polynomial algorithm, and sigmoid algorithm. Additionally, C-SVM and Nu-SVM were applied to four kernel algorithm types to create eight types, and prediction accuracies of the eight SVM types were estimated and compared. The analysis was conducted using R version 3.4.3.

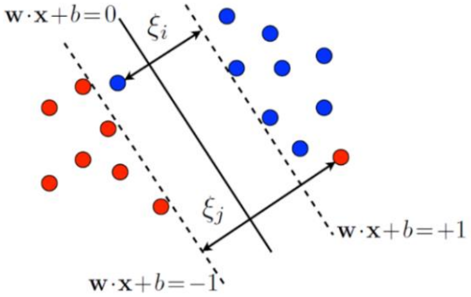

$\mathbf{w}\cdot\mathbf{x}+b=0$
ξ_i
ξ_j
$\mathbf{w}\cdot\mathbf{x}+b=+1$
$\mathbf{w}\cdot\mathbf{x}+b=-1$

Fig. 1. The Concept of Kernel Functions [26].

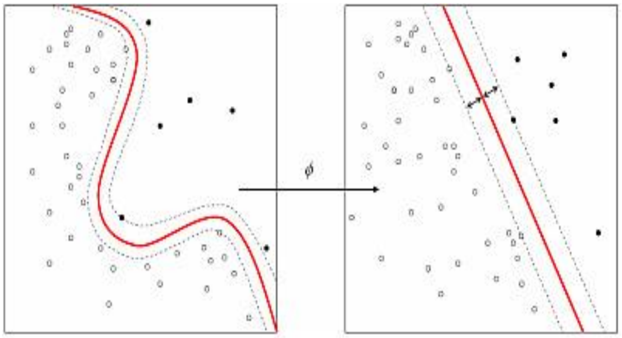


Fig. 2. The concept of soft-margin [26]

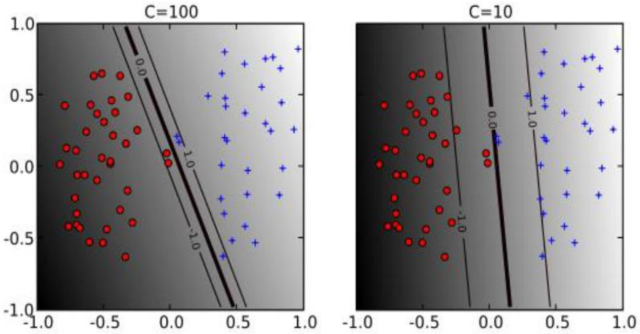


Fig. 3. Effect of Soft-margin [26]

IV. Results

A. General Characteristics of Subjects

Among 2,420 subjects, 483 elderly (19.9%) were not satisfied with their current lives (Table II). The results of chi-square test showed that age, gender, education level, economic activity, living with a spouse or not, subjective health status, walking practice per week, depression, illness and accident experience within the past two weeks, the frequency of meeting a neighbor, and the frequency of meeting a relative were significantly different between the elderly satisfied with their lives and those not satisfied with their lives (p <0.05). The ratio of life dissatisfaction were high when subjects were 75 years old or older (37.5%), males (23.1%), elementary school graduates or below (26.1%), unemployed (26.1%), separated from spouses (32.3%), poor subjective health (20.4%), senior citizens who walked one day or less per week on average (25.7%), depression (66.3%), senior citizens who experience a disease or an accident in the past two weeks (29.5%), when the frequency of meeting a neighbor is less than once a month (35.2%), when the frequency of meeting a relative is less than once a month (37.2%).

B. Predictors of Life Dissatisfaction based on C-SVM

The function weights of the SVM based on the Gaussian kernel algorithm are presented in Table III. Although it is impossible to simply compare the magnitudes (priorities) of variables using the function weights of C-SVM, it is possible to determine whether the predictor is a preventative factor (positive relationship) or a risk factor (negative relationship). This C-SVM based model for predicting the quality of life in old age derived both preventive and risk factors. Since the objective of this study was to explore the main predictors of life dissatisfaction in old age, this study only evaluated factors affecting life dissatisfaction. It was found that the predictors of life dissatisfaction in old age were 75 years old or older, currently unemployed, household income equal to or below 4 million KRW, middle school graduate or below, male, poor subjective health, depression, experienced a disease or accident in the past two weeks, walk one day or less per week, and the frequency of meeting a relative or neighbor equal to or less than once a month. The final prediction accuracy of this SVM using 625 support vectors was 92.63%.

C. Ccuracy of Predicting the Satisfaction of Life in Old Age According to the Classification Algorithm of the SVM

The accuracy of predicting the satisfaction of life according to the classification algorithm of the SVM is presented in Table IV. The model fitness of the SVM may vary depending on a kernel type. Therefore, this study compared the prediction accuracy of diverse algorithms (i.e., Gaussian, linear, polynomial, and sigmoid) to evaluate the prediction accuracies of models according to a kernel type. Moreover, this study examined two SVM types (C-SVM and Nu-SVM), so this study compared the prediction accuracies of eight different SVM types. The results showed that the difference between C-SVM and Nu-SVM was negligible in the models for predicting the satisfaction of life in old age while the Gaussian kernel had the highest accuracy and the sigmoid kernel had the lowest accuracy.

TABLE. II. Characteristics of Participants based on Life Satisfaction, n (%)

Variables	life satisfaction		p
	Satisfied (n=1,937)	Not satisfied (n=480)	
Age			<0.001
65-74	1,333 (75.3)	438 (24.7)	
75+	404 (62.5)	242 (37.5)	
Gender			<0.001
Male	854 (76.9)	256 (23.1)	
Female	1,083 (82.7)	227 (17.3)	
Education			<0.001
Elementary school graduation and below	675 (79.9)	238 (26.1)	
Middle school graduation	243 (78.1)	68 (21.9)	
High school graduation	528 (83.7)	103 (16.3)	
College graduation and above	489 (86.9)	74 (13.1)	
Economic activity			<0.001
Employed	1,040 (86.1)	168 (13.9)	
Not-employed	894 (73.9)	315 (26.1)	
Household Income			0.282
Below 2 million KRW	1,039 (79.4)	270 (20.6)	
Between 2 and 4 million KRW	549 (79.9)	138 (20.1)	
4 million KRW or above	336 (83.0)	69 (17.0)	
Spouse			<0.001
Cohabitation	1,397 (81.2)	324 (18.8)	
Separation	251 (67.7)	120 (32.3)	
Bereavement	284 (88.2)	38 (11.8)	
Drinking frequency			0.710
Once a week or less	1,547 (79.9)	389 (20.1)	
Once a week or more	388 (80.7)	93 (19.3)	
Smoking			0.179
Current smoker	347 (82.4)	74 (17.6)	
Non-smoker	1,590 (79.5)	409 (20.5)	
Subjective health status			0.003
Good	549 (82.3)	118 (17.7)	
Average	700 (83.4)	139 (16.6)	
Poor	728 (79.6)	186 (20.4)	
Walking per week			<0.001
One day or more	1,367 (82.7)	286 (17.3)	
Less than one day	569 (74.3)	197 (25.7)	
Depression			<0.001
Yes	110 (33.6)	217 (66.3)	
No	1,627 (77.8)	463 (22.2)	
Disease or accident experience in the past two weeks			0.011
Yes	304 (70.5)	127 (29.5)	
No	1,573 (79.1)	416 (20.9)	
Frequency of meeting a neighbor			0.001
Less than once a month	426 (64.8)	231 (35.2)	
Twice or more per month	1,440 (82.3)	309 (17.7)	
Frequency of meeting a relative			0.001
Less than once a month	301 (62.8)	178 (37.2)	
Twice or more per month	1,654 (85.3)	284 (14.7)	

TABLE. III. VALUES OF FUNCTION WEIGHTS

Age	
65-74	.011
75+	-.032
Gender	
Male	-.019
Female	.005
Education	
Elementary school graduation and below	-.030
Middle school graduation	-.028
High school graduation	.003
College graduation and above	.007
Economic activity	
Employed	.025
Not-employed	-.033
Household Income	
Below 2 million KRW	-.007
Between 2 and 4 million KRW	-.007
4 million KRW or above	.005
Spouse	
Cohabitation	.017
Separation	-.027
Bereavement	.019
Drinking frequency	
Once a week or less	.030
Once a week or more	.028
Smoking	
Current smoker	.007
Non-smoker	.009
Subjective health status	
Good	.007
Average	.004
Poor	-.011
Walking per week	
One day or more	.020
Less than one day	-.034
Depression	
Yes	-.037
No	.018
Disease or accident experience in the past two weeks	
Yes	-.023
No	.015
Frequency of meeting a neighbor	
Less than once a month	-.038
Twice or more per month	.015
Frequency of meeting a relative	
Less than once a month	-.031
Twice or more per month	.018
Number of support vectors: 625	

TABLE. IV. THE ACCURACY OF PREDICTING THE SATISFACTION OF LIFE ACCORDING TO THE CLASSIFICATION ALGORITHM OF THE SVM, %

Type of SVM	Type of Kernel			
	Linear	Polynomial,	Gaussian	Sigmoid
C-SVM	91.25	90.58	92.63	89.88
Nu-SVM	91.14	90.77	92.10	88.85

V. DISCUSSION

This study developed an SVM-based model for predicting the satisfaction of life in old age using the data of the Seoul Statistics Survey. It was found that 20% of the surveyed elderly were dissatisfied with their current lives. “The Satisfaction of Life of Elderly Population Groups [27]” published by Statistics Korea in 2018 showed that the satisfaction of life (⩾50 years old) of South Korean was 5.4 out of 11 points, which was 1 point lower than the average of OECD countries (6.4 points), even though it compared with cannot be directly the results of this study. It was ranked as 28thamong35OEC Dmembercountries:2pointslowerthanDenmark (7.6 points; the top score), and even lower than Japan (5.8 points) [27]. If South Korean enters a post-aged society with continuing this trend, it is more likely to decrease the quality of life in old age. Therefore, it is necessary to prepare polices at a society level and take active measures to improve the quality of life in old age.

The results of this study revealed that health status (e.g., subjective health status, depression, and the experience of a disease or accident in the past two weeks), health habits (number of walking per week), and social network/ communication frequency (the frequency of meeting a neighbor and the frequency of meeting a relative) were main factors for predicting the quality of life in old age in addition to demographic factors (e.g., age, gender, education, economic activity, and marital status). Numerous previous studies evaluating the quality of life in old age reported that health status [30] and marital status [31] were main factors affecting the quality of life and these results, in addition to sociodemographic characteristics such as economic level [28] and age [29], agree with the results of this study. Previous studies on the satisfaction of life related to the economic stability of the elderly [32, 33] showed that those with low incomes generally were recipients of national basic livelihood guarantees and they were not satisfied with their living conditions when they did not have a spouse. As of 2018, the poverty rate of the elderly (⩾65 years) in South Korea was 48.8%, which is four-folds of the average (12.1%) of OECD countries [27]. Therefore, economic support is needed to improve the quality of life in old age.

Previous studies [32, 33, 34, 35] indicated that the better health status of the elderly increased the level of their life satisfaction while functional impairment (e.g., communication problem) that adversely influenced the interaction with others negatively affected the satisfaction of life. In particular, it was found that health status had a greater effect on declining life satisfaction in female elderly than male elderly and older elderly than younger elderly [33]. Additionally, many studies showed that marital status was a major factor affecting the quality of life [34]. The elderly living with spouses had a higher level of life satisfaction than those who were unmarried or those who lose their spouses [35]. Particularly, marital status was a very important factor in determining life satisfaction for older elderly than younger elderly.

An interesting finding of this study was that social networks and communication frequency were major predictors for the quality of life in old age. Kim et al. [36] also reported that the satisfaction of life increased with more frequent contacts and higher quality contacts with family members, friends, and neighbors (so-called better social networks). It could be because the elderly would require more social support than younger people in order to relieve loneliness and physical unwellness [36]. It is particularly well known that emotional support from the family has an important effect on resolving the loneliness of the elderly [16]. It has been reported that the elderly who receive more support from the family have better emotional health, better life satisfaction, and lower loneliness [36]. Therefore, in order to improve the quality of life in old age, it is necessary to establish a system that can increase the frequency of regular contacts (communication) with the members of local community groups such as volunteers as well as relatives, friends, and neighbors.

Another major finding of this study was that the prediction accuracy of C-SVM's Gaussian kernel was higher than that of linear kernel, polynomial kernel, and sigmoid kernel algorithms. The performance of nonlinear SVM depends on the kernel functions applied to the algorithms and parameters composing them. The Gaussian kernel is an algorithm that maps the data to a characteristic space of infinite dimension. Author in [21] also proved that it is an algorithm with high prediction accuracy. Therefore, it is believed that using a Gaussian kernel based C-SVM algorithm will be more effective for subsequent studies to develop models for predicting binary variables than using a sigmoid algorithm.

The results of this study would provide an important basis that must be considered for developing health policies for successful aging. Based on the prediction model of this study, it is needed to manage local communities systematically to enhance the quality of life in old age.

REFERENCES

[1] Statistics Korea, statistics of elderly persons, Statistics Korea, Daejeon, 2018.

[2] A. Bowling, Ageing well: Quality of life in old age, McGraw-Hill Education, London, 2005.

[3] P. H. Chen, J. S. Golub, E. R. Hapner, and M. M. Johns, Prevalence of perceived dysphagia and quality-of-life impairment in a geriatric population. Dysphagia, vol. 24, no. 1, pp. 1-6, 2009.

[4] R. Manrique-Huarte, D. Calavia, A. H. Irujo, L. Girón, and M. Manrique-Rodríguez, Treatment for hearing loss among the elderly: auditory outcomes and impact on quality of life. Audiology and Neurotology, vol.21(Suppl. 1), pp. 29-35, 2016.

[5] M. Pigliautile, F. Chiesi, C. Primi, S. Inglese, D. Mari, D. Simoni, E. Mossello, and P.Mecocci,Validation study of the Italian version of Communication Activities of the Daily Living (CADL2) as an ecologic cognitive assessment measure in older subjects.NeurologicalSciences,e-pub:doi.org/10.1007/s10072-019-03937-w, 2019.

[6] P. H. Noël, J. W. Williams, J. Unützer, J. Worchel, S. Lee, J. Cornell, W. Katon, L.H.Harpole, and E.Hunkeler, Depressionand comorbid illnessin elderly primary carepatients: impacton multiple domains of health status and well-being. The Annals of Family Medicine, vol.2, no.6, pp.555-562, 2004.

[7] J. B. Unger, G. McAvay, M. L. Bruce, L. Berkman, and T. Seeman, Variation in the impact of social network characteristics on physical functioning in elderly persons: MacArthur Studies of Successful Aging. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, vol. 54, no. 5, pp. S245-S251, 1999.

[8] E. Diener, R. E. Lucas, and S. Oishi, Subjective well-being: The science of happiness and life satisfaction. Handbook of positive psychology, vol. 2, pp. 63-73, 2002.

[9] U. Schimmack, P. Radhakrishnan, S. Oishi, V. Dzokoto, and S. Ahadi, Culture, personality, and subjective well-being: Integrating process models of life satisfaction. Journal of personality and social psychology, vol. 82, no. 4, 582-593, 2002.

[10] J. Butler, and J. Ciarrochi, Psychological acceptance and quality of life in the elderly. Quality of life Research, vol. 16, no.4, pp. 607-615, 2007.

[11] S. Y. Park, A study on depression, ADL, IADL, and QOL among community-dwelling, low income elderly. Journal of Korean Public Health Nursing, vol. 23, no. 1, pp. 78-90, 2009.

[12] A. Öztürk, T. T. Şimşek, E. T. Yümin, M. Sertel, and M. Yümin, The relationship between physical, functional capacity and quality of life (QoL) among elderly people with a chronic disease. Archives of Gerontology and Geriatrics, vol. 53, no. 3, pp. 278-283, 2011.

[13] T. Wada, M. Ishine, T. Sakagami, K. Okumiya, M. Fujisawa, S. Murakami, K. Otsuka, S. Yano, T. Kita, and K. Matsubayashi, Depression in Japanese community-dwelling elderly—prevalence and association with ADL and QOL. Archives of Gerontology and Geriatrics, vol. 39, no.1, pp, 15-23, 2004.

[14] H. Byeon, Developing a model to predict the social activity participation of the senior citizens living in South Korea by combining artificial neural network and quest algorithm. International Journal of Engineering & Technology, vol. 8, no. 1.4, pp. 214-221, 2019.

[15] A. P. Lane, C. H. Wong, Š. Močnik, S. Song, and B. Yuen, Association of Neighborhood Social Capital With Quality of Life Among Older People in Singapore. Journal of Aging and Health, vol. e-pub: doi.org/10.1177/0898264319857990, 2019.

[16] O. M. R. Gouveia, A. D. Matos, and M. J. Schouten, Social networks and quality of life of elderly persons: a review and critical analysis of literature. Revista Brasileira de Geriatria e Gerontologia, vol. 19, no. 6, pp. 1030-1040, 2016.

[17] H. Byeon, and S. Kim, Development of risk prediction model for stroke among Korean older adults using quest algorithm: a community-based cross-sectional study. International Journal of Applied Engineering Research, vol. 10, no. 79, pp. 93-96, 2015.

[18] S. G. Kumar, A. Majumdar, and G. P, Quality of life (QOL) and its associated factors using WHOQOL-BREF among elderly in urban Puducherry, India. Journal of clinical and diagnostic research: JCDR, vol. 8, no. 1, pp. 54-57, 2014.

[19] J. Butler, and J. Ciarrochi, Psychological acceptance and quality of life in the elderly. Quality of life Research, vol. 16, no. 4, pp. 607-615, 2007.

[20] S. Khan, and T. Yairi, A review on the application of deep learning in system health management. Mechanical Systems and Signal Processing, vol. 107, pp. 241-265, 2018.

[21] H. Byeon, Model development for predicting the occurrence of benign laryngeal lesions using support vector machine: focusing on South Korean adults living in local communities. International Journal of Advanced Computer Science and Applications, vol. 9, no.10, pp. 222-227, 2018.

[22] J. Chorowski, J. Wang, and J. M. Zurada, Review and performance comparison of SVM-and ELM-based classifiers. Neurocomputing, vol. 128, pp. 507-516, 2014.

[23] Y. C. Wu, Y. S. Lee, and J. C. Yang, Robust and efficient multiclass SVM models for phrase pattern recognition. Pattern recognition, vol. 41, no. 9, pp. 2874-2889, 2008.

[24] N. Herrmann, N. Mittmann, I. L. Silver, K. I. Shulman, U. A. Busto, N. H. Shear, and C. A. Naranjo, A validation study of the Geriatric Depression Scale short form. International Journal of Geriatric Psychiatry, vol. 11, no. 5, pp. 457-460, 1996.

[25] B. Scholkopf, and A. J. Smola, Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, Cambridge, 2001.

[26] Available:https://ratsgo.github.io/machine%20learning/2017/05/29/SVM2/

[27] Statistics Korea, Basic Research on Life Satisfaction of the Elderly, Statistics Korea, Daejeon, 2018.

[28] M. H. Alshamali, M. M. Makhlouf, M. Rady, A. Selim, N. Abdel, S. Ismail, and M. Fawaz, Quality of life and its predictors among Qatari elderly attending primary health care centers in Qatar. Middle East Journal of Family Medicine, vol. 7, no. 10, pp. 9-19, 2019.

[29] K. H. Jo, and H. J. Lee, Factors related to life satisfaction in young-old, old, and oldest-old women. Journal of Korean Academy of Nursing, vol. 39, no. 1, pp. 21-32, 2009.

[30] C. W. Han, E. J. Lee, T. Iwaya, H. Kataoka, and M. Kohzuki, Development of the Korean version of Short-Form 36-Item Health Survey: health related QOL of healthy elderly people and elderly patients in Korea. The Tohoku journal of experimental medicine, vol. 203, no. 3, pp. 189-194, 2004.

[31] T. Rakhshani, D. Shojaiezadeh, K. B. Lankarani, F. Rakhshani, M. H. Kaveh, and N. Zare, The association of health-promoting lifestyle with quality of life among the Iranian elderly. Iranian Red Crescent Medical Journal, vol. 16, no. 9, e-pub: doi.10.5812/ircmj.18404 2014.

[32] Y. S. Kim, and K. H. Suh, Living arrangement, satisfaction with living, end depression among the Korean elderly. Korean Journal of Gerontological Social Welfare, vol. 18, no. 1, pp. 273-290, 2002.

[33] J. D. Kwon, and J. Y. Cho, A study of factors influencing the life satisfaction of the aged. Journal of the Korea Gerontological Society, vol. 20, no. 3, pp. 61-76. 2000.

[34] H. R. Hedayati, N. Hadi, L. Mostafavi, A. Akbarzadeh, and A. Montazeri, Quality of life among nursing home residents compared with the elderly at home. Shiraz E-Medical Journal, vol. 15, no. 4, e-pub: doi.10.17795/semj22718, 2014.

[35] M. Tajvar, M. Arab, and A. Montazeri, Determinants of health-related quality of life in elderly in Tehran, Iran. BMC public health, vol. 8, no. 1, e-pub: doi.org/10.1186/1471-2458-8-323, 2008.

[36] H. K. Kim, M. Hisata, I. Kai, and S. K. Lee, Social support exchange and quality of life among the Korean elderly. Journal of Cross-Cultural Gerontology, vol. 15, no. 4, pp. 331-347, 2000.

How to Improve the IoT Security Implementing IDS/IPS Tool using Raspberry Pi 3B+

Ruíz-Lagunas Juan Jesús[1]
Departamento de Sistemas y Computación, TecNM/I.T. Morelia and Universidad Vasco de Quiroga
Morelia, México

Antolino-Hernández Anastacio[2], Torres-Millarez Cristhian[5], Paniagua-Villagómez Omar[6]
Departamento de Sistemas y Computación
TecNM/I.T. Morelia, Morelia, México

Reyes-Gutiérrez Mauricio René[3]
Facultad de Ingeniería Eléctrica-UMSNH/Departamento de Sistemas y Computación TecNM/I.T. Morelia
Morelia, México

Ferreira-Medina Heberto[4]
Unidad de TICs, IIES-UNAM and Departamento de Sistemas y Computación, TecNM/I.T. Morelia
Morelia, México

***Abstract*—This work shows a methodology of implementation and testing of the system is proposed and tested with a prototype; it is constructed with sensors and actuators that allow monitoring the behavior of the system in an environment under threats. We used an IDS / IPS as a protection tool for IoT systems, based on Raspberry Pi and Raspbian operating system. It is described in a block diagram the testing method used. We implemented the IDS/IPS Snort tool in an embedded platform Raspberry. It presents also the state of the art of cloud frameworks that have the same objective of protecting. The main contribution is the implemented testing method for Snort that can be used with security rules in other applications of embedded IoT devices.**

***Keywords*—Security IoT; IDS/IPS software; Pentesting tools; smart cities; prototype Raspberry**

I. Introduction

Today, Information Technology (IT) is immersed in the use and exploitation of telecommunications networks, to which more devices are connecting every day to communicate with each other (Peer to Peer) and to a central device (client-server). Internet of Things (IoT) offers wide range state of the art solutions, using sensors and/or devices, which monitor to control certain events, giving rise to new challenges to IT security, since information gathered can be compromised by this variety of components. Internet of Things (IoT) is a concept defined by Kevin Ashton in 1999, which describes a network that connects people and objects [1]. These objects, right now, allow computers to have "sensors", which facilitate them, not only to process information but gather more information through these devices, allowing applications to be even more "intelligent", since it is possible to make decisions in real-time, based on a major quantity of information. It is possible to assure that since the first implementation of IoT up until date there are millions of sensors interconnected.

The concept of "smart" devices is inherent in connectivity to obtain benefits from the information [2]. Industry automation, and monitoring systems are the main reasons for this technology's success, so data networks are being unified with production networks to achieve these benefits. It has been estimated that in 2018 there were more than 7 billion IoT devices [3]. It is estimated that by 2020 there will be more than 10 billion and more than 22 billion by 2025. This worldwide trend is due to the growing demand to connect devices to the networks.

With this trend, has been observed clearly that the way of the information interchange with technology will change, but at what cost? Due to the demand for interconnection, many IoT developers do not consider the security in communication for many reasons: Amongst these, we have processing costs, training and algorithm implementation.

IoT devices are considered to have many weaknesses in information security since their development. The following are examples described below:

- Passwords stored in plain text.
- Outdated firmware and not encryption.
- Video streaming without encryption.
- Communication between devices and servers in plain text.
- Over-shared data (influence of cloud utilization).
- Development bugs in the firmware.
- Use of default passwords.
- Devices have a direct interface to the internal network, but they can be connected to the Internet, making increase an attack risk exponentially.

Sponsors TecNM/I.T. Morelia and IIES-UNAM, México

In addition, it has known that hackers are exploiting these vulnerabilities with current tools and techniques of their own to achieve that goal. One of the most recent tools to detect vulnerabilities in IoT devices is *Autosploit* [4], since it uses artificial intelligence in its algorithms [5].

The main contribution of this paper is to present a testing method for IDS/IPS and the comparison of its response implemented on the Raspberry platform to Nmap and Metasploit of network attacks. The paper is organized as follows: Section II deals with actual security in IoT, in Section III describes IDS/IPS tools, in Section IV describes the methodology, Section V shows the results about this work, and finally, in Section VI shows conclusions.

II. Security in IoT

As a response to exposed security problems, the main software developers propose various strategies to guarantee security in interconnection components. According to [6] the expansion of IoT, it has been developing in the following areas, see Table I.

TABLE. I. IoT Expansion Areas based on [6]

IoT development areas	**Description**		
	Elements	***Opportunities***	***Challenges***
Smart life	- Health care. - Consumer and Retail businesses. - Bank Convergence. - Security. - Public services	Technologies that promote simplifications in the lives of the users.	Ensure secure information and data exchange.
Smart mobility	- Intelligent vehicles - Urban mobility. - Intercity mobility. - Rate management and payment solutions. - Distribution and logistics. -Fleet management.	Real-time solutions that make mobility simpler and transport reliable.	Secure interconnection and secure real-time monitoring and activation.
Smart Cities	- Intelligent infrastructure management. - Cross-agency collaboration using the cloud. - Data collection in real time and quickly. - Better planning of cities. -Network utilities. -Construction Development.	The innovations will aim to improve the quality of life in the city. Using sensors and systems that help in decision-making.	Ensure that users exchange information in real time and keep their data protected against hackers.
Smart manufacturing	- Machine learning. -Communication between machines. - Network interconnection. -Optimization of processes. -Proactive asset management. - Improve infrastructure integration.	Smart solutions to optimize production processes, controls and quality.	Keep process information safe, interconnections between machines using secure protocols.

III. Frameworks for Security in IoT

A. Frameworks in the Cloud

According to [7] the challenge in IoT, security will be to build services that can be integrated into different software solutions. In platforms described in [8], [9], [10], [11], [12] and [13] a set of services oriented to software and hardware solutions to offer layers of security at different levels, working from the platform (**PaaS**, Platform as a Service), and services (**SaaS**, Software as a Service), which are offered with auto-service, cost and on-demand schemes. Table II shows the main features of these proposals that are a reference as a framework and cloud computing.

The list of platforms shown in Table III is compared by the type of framework and the security mechanisms they offer with encryption method, in a mobile application (App) or any API.

Despite the implementation of different security schemes both in the device, in the interconnection and in the cloud, these do not guarantee that the IoT components are free from attacks using the different layers of the communication protocols [20].

TABLE. II. MAIN FRAMEWORKS IN THE CLOUD THAT OFFER SECURITY FOR IOT

Framework	Description	Elements	Security
Amazon Web Services IoT [8]	It offers a cloud platform, as well as IoT hardware, operating system, software and cloud connectivity services, security layers, monitoring, and administration software for PaaS services.	- Software for devices. - Service control. - Data services. - Secure interconnection services.	Oriented to platforms with clients and servers using secure interconnection mechanisms, it offers cryptography and monitoring services.
Microsoft Azure IoT Hub [9]	They offer a cloud platform with open and flexible services to connect securely, monitor and manage IoT devices and develop applications using open source SDK (Development Kit) and multiple protocols. Working under a SaaS scheme.	- Device layer. -Interconnection layer. - Cloud access layer. - Hub layer. - Back-End for Apps.	It offers an exchange of information with the devices, using languages such as NodeJS, .Net, Java, Python, Android, IOS and C. It establishes layers of security for the connection.
Oracle Internet of Things Cloud Services [10]	It offers a PaaS scheme, which allows you to connect IoT devices to the cloud, analyze data in real time to integrate it into business applications, allows you to establish web services and any other Oracle proprietary service.	- Software for devices. - Hub for access to cloud services. - Offers a wide range of SaaS.	It offers connectivity with IOS, Android and any device that uses Java, Posix C, and the RESTful protocol offers a cryptography scheme.
Watson IoT Platform [11]	It offers a PaaS-based connectivity scheme, offers a firmware to be installed on different platforms and achieve connectivity and the use of services.	- Use of IBM cloud. - Device management. -Platform services. - Administration services. -Blockchain services.	Uses protocols for secure interconnection using a gateway, uses node.js, java, and JS languages. Offers blockchain and crypto services.
Xively, Google [12]	Google Cloud IoT offers a set of tools to, connect, process, store and analyze data both in the perimeter and in the cloud, they are PaaS services.	- Software for sensors - Cloud Connection (Edge). - Android, CPU, GPU and TPU support. - Offers real-time analysis service. -Data usage services.	It offers scalable and managed services, integrates Artificial Intelligence functions, and uses a communication protocol with security schemes (MQTT, Machine-to-Machine protocol).
Samsung Artik [13]	It offers connectivity services based on PaaS, to connected smart devices, as well as connected homes and smart cities, allows connectivity under the concept known as D3 (Data Driven Development).	- Software for sensors and devices. - Storage services and location of services (brokerage). - Support for third-party users and Apps.	It offers Big Data usage scheme, it uses REST, Web Sockets, MQTT and CoAP protocols (Protocol for restricted devices) to exchange information.

TABLE. III. PLATFORM COMPARISON FOR IOT

No.	Framework	Firmware owner	Security mechanism
1	Amazon Web Services IoT	Yes	Encryption
2	Azure IoT Hub	No	In app
3	Oracle IoT Cloud Services	No	In app
4	Watson IoT Platform	Yes	Blockchain
5	Xively	No	In app
6	Samsung Artik	Si	Encryption
7	Carriots [14]	Si	API keys
8	Adafruit.io [15]	No	In app
9	Ubidots [16]	No	In app
10	MyDevices Cayenne [17]	No	In app
11	Macchina IO [18]	No	In app
12	ThingSpeak [19]	No	In app
13	Arduino IoT Cloud [20]	Si	Encryption

IV. Intrusion Prevention and Detection Systems (IDS/IPS)

The IDS and IPS are complements to improve security on host systems (HIDS/IPS), mainly for embedded IoT devices, to establish which of the current tools is most suitable for IoT, a comparison was made between the main open source IDS/IPS offered on the Internet. It is based on the sum of the indicators achieved (functionality, usability, reliability, performance, supportability), each indicator contributes 5 points; 0 doesn't accomplish, 1 is mentioned, 2 slightly accomplished, 3 accomplished, 4 very accomplished, 5 Extremely accomplished. Fig. 1 shows this comparison adding the indicators achieved for the following tools IDS and/or IPS: Snort [21], Suricata [22], Broids [23], OSSec [24], OS Tripwire [25], Aide [26], Samhain [27], Fail2ban [28], Sagan [29].

As we can observe Snort and Sagan, are the best tools evaluated.

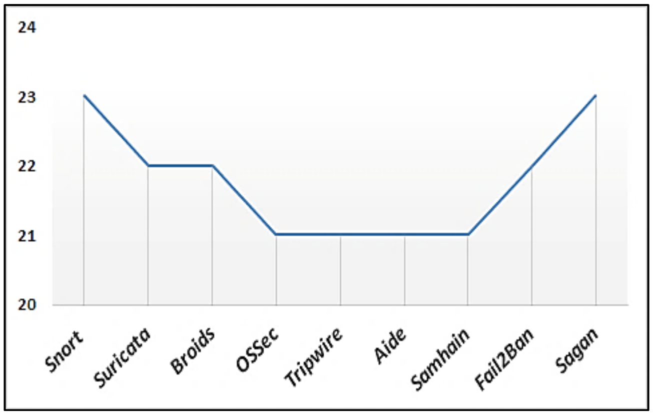


Fig. 1. Comparison between the main IDS/IPS using FURPS.

V. Methodology

The research methodology used is based on an experimental and applied method, therefore the process compromises several steps, which they are described in next lines. The implementation of an IDS/IPS as a security scheme on a Raspberry Pi3B+ card, is a relatively simple process, however, it is necessary to evaluate the operation of the system, to develop adequate detection rules using Snort and Sagan, to improve the embedded system in the management and monitoring the network traffic and the internal state of the device.

Fig. 2 shows the methodology in block diagram used to design the prototype and pentesting probes. The methodology proposes a reviewing of the state of the art in IoT security context, and related or similar projects, then a prototype is implemented with the IDS/IPS tool installed for pentesting and monitoring threat behavior in these devices. At the end, feedback is proposed to improve the prototype's components and software tools for security.

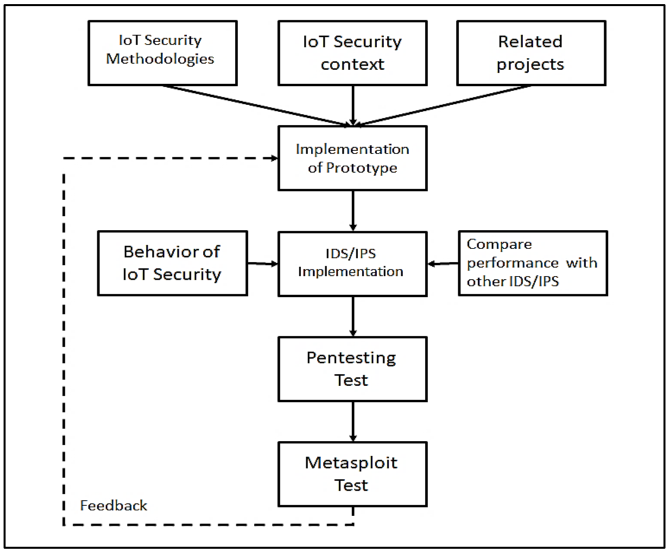


Fig. 2. Methodology in Block Diagram Implemented.

A. IoT Prototype Implementation

A system based on Raspberry Pi3 + was built, with the Raspbian Operating System, using Python language and compatible components for Raspberry card [30]. They were assembled to verify system performance against attacks. The components used are described in Table IV.

The IDS / IPS system was installed and configured in the prototype and the functional tests of each of the sensors and actuators were performed, as shown in the diagram in Fig. 3.

TABLE. IV. IoT Components Assembled in the Prototype

No.	Components	Type	Function
1	Raspberry Pi3+	Card	Host IoT system with OS Raspbian 4.19
2	Adc1	Sensor	I^{2}C, a signal for detecting AC *I*
3	Adc2	Sensor	I^{2}C, for detecting DC *I*
4	Adc3	Sensor	I^{2}C, for detecting AC Voltage
5	Adc4	Sensor	I^{2}C, , for detecting DC Voltage
6	BH1750FVI	Sensor	I^{2}C, for detecting light
7	PIR	Sensor	Detect presence /absence, ON/OFF
8	FZ0430	Sensor	Detect DC voltage
9	MCP3424	Card	Analog-Digital Converter with I^{2}C
10	Relay 2	Card	Relay 2 canals 5v
11	LED	Bulb	127v bulb
12	Electromagnet	Actuator	Opening device
13	Motor	Actuator	12v motor
14	Ov5647	Actuator	Infrared night vision camera with IR sensor, 5MP

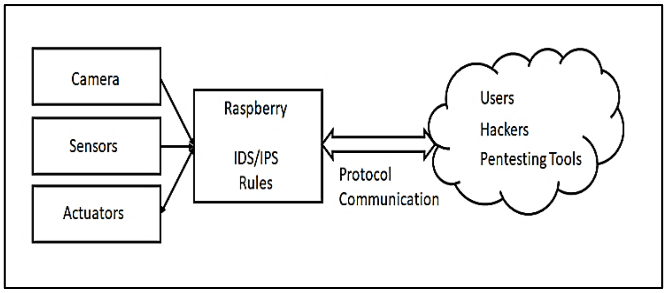


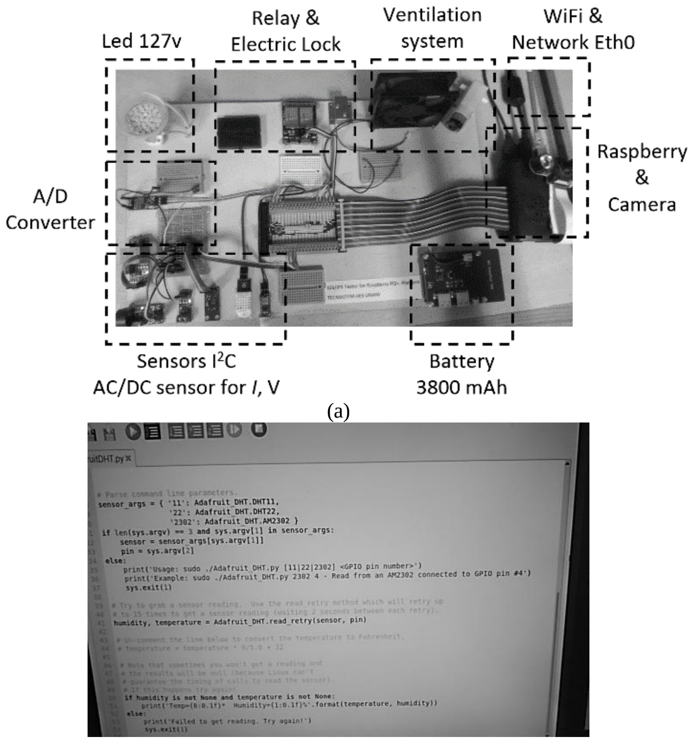
Fig. 3. System Design View and Attack Tests.

The designed prototype was integrated with the sensors and actuators to form an embedded and functional system, as shown in Fig. 4(a)). Tests were carried out with scripts in Python v3 language as shown in Fig. 4(b)).

B. IDS/IPS Implementation

Snort automates and simplifies intruder detection, using rules that describe the behavior of different attacks. The installation procedure in Raspbian is described below (it is important to have the equipment connected to the network):

```
#sudo apt-get update
#sudo apt-get install snort snort-common snort-common-libraries snort-rules-default libpcap-dev
#sudo dpkg-reconfigure snort
// sudo command vi /etc/snort/snort.debian.conf parameters
#sudo vi /etc/snort/snort.conf
// Review the configuration of the nine sections to adapt the operation of the IDS
#sudo rc.d stop snort
#sudo rc.d start snort
```



(a)

```
# Parse command line parameters.
sensor_args = { '11': Adafruit_DHT.DHT11,
                '22': Adafruit_DHT.DHT22,
                '2302': Adafruit_DHT.AM2302 }
if len(sys.argv) == 3 and sys.argv[1] in sensor_args:
    sensor = sensor_args[sys.argv[1]]
    pin = sys.argv[2]
else:
    print('Usage: sudo ./Adafruit_DHT.py [11|22|2302] <GPIO pin number>')
    print('Example: sudo ./Adafruit_DHT.py 2302 4 - Read from an AM2302 connected to GPIO pin #4')
    sys.exit(1)

# Try to grab a sensor reading.  Use the read_retry method which will retry up
# to 15 times to get a sensor reading (waiting 2 seconds between each retry).
humidity, temperature = Adafruit_DHT.read_retry(sensor, pin)

# Un-comment the line below to convert the temperature to Fahrenheit.
# temperature = temperature * 9/5.0 + 32

# Note that sometimes you won't get a reading and
# the results will be null (because Linux can't
# guarantee the timing of calls to read the sensor).
# If this happens try again!
if humidity is not None and temperature is not None:
    print('Temp={0:0.1f}*  Humidity={1:0.1f}%'.format(temperature, humidity))
else:
    print('Failed to get reading. Try again!')
    sys.exit(1)
```

(b)

Fig. 4. Prototype Design and Testing, a) Component Integration and b) Sensors Testing in Raspbian with Python Language.

C. Pentesting

IDS/IPS operation tests were performed using Nmap and Metasploit. With Nmap, the following instructions were applied:

```
#nmap –f –sS –sV –script auth 192.168.0.9
```

Fig. 5 shows the vulnerabilities detected by Nmap, in the active services that use authentication in the prototype.

Fig. 6 shows a list of all the vulnerabilities detected by Nmap in the prototype's active services.

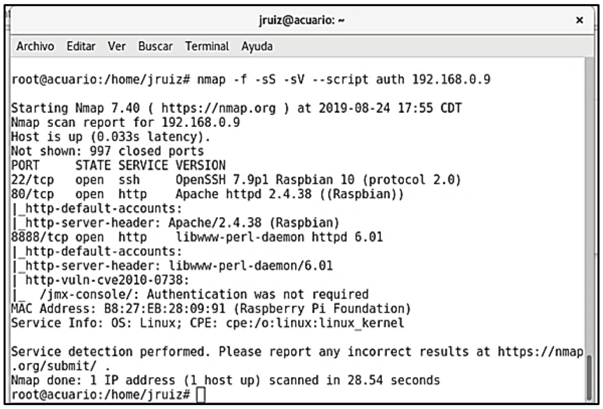
```
#nmap –f –script vuln 192.168.0.9
```

Fig. 7 shows the traffic detected and blocked by Snort against attacks, in the scanning of vulnerabilities with Nmap. The metric used is the type of traffic per app.

D. Vulnerability Test

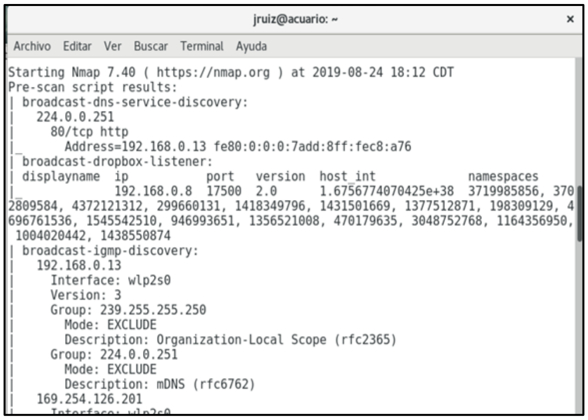
Using Metasploit tool, the following exploits were applied: DDoS attack on port 80, which consists of saturating packets to that service, with the goal of denying the service to users.

```
#msfconsole
#use auxiliary/dos/tcp/synflood
#set RPORT 80
#set RHOST 192.168.0.9
#Run
```



```
root@acuario:/home/jruiz# nmap -f -sS -sV --script auth 192.168.0.9

Starting Nmap 7.40 ( https://nmap.org ) at 2019-08-24 17:55 CDT
Nmap scan report for 192.168.0.9
Host is up (0.033s latency).
Not shown: 997 closed ports
PORT     STATE SERVICE VERSION
22/tcp   open  ssh     OpenSSH 7.9p1 Raspbian 10 (protocol 2.0)
80/tcp   open  http    Apache httpd 2.4.38 ((Raspbian))
|_http-default-accounts:
|_http-server-header: Apache/2.4.38 (Raspbian)
8888/tcp open  http    libwww-perl-daemon httpd 6.01
|_http-default-accounts:
|_http-server-header: libwww-perl-daemon/6.01
| http-vuln-cve2010-0738:
|_  /jmx-console/: Authentication was not required
MAC Address: B8:27:EB:28:09:91 (Raspberry Pi Foundation)
Service Info: OS: Linux; CPE: cpe:/o:linux:linux_kernel

Service detection performed. Please report any incorrect results at https://nmap
.org/submit/ .
Nmap done: 1 IP address (1 host up) scanned in 28.54 seconds
root@acuario:/home/jruiz# 
```

Fig. 5. Scanning Test with Nmap.



```
Starting Nmap 7.40 ( https://nmap.org ) at 2019-08-24 18:12 CDT
Pre-scan script results:
| broadcast-dns-service-discovery:
|   224.0.0.251
|     80/tcp http
|_      Address=192.168.0.13 fe80:0:0:0:7add:8ff:fec8:a76
| broadcast-dropbox-listener:
| displayname  ip           port   version  host_int              namespaces
|_             192.168.0.8  17500  2.0      1.6756774070425e+38  3719985856, 370
2809584, 4372121312, 299660131, 1418349796, 1431501669, 1377512871, 198309129, 4
696761536, 1545542510, 946993651, 1356521008, 470179635, 3048752768, 1164356950,
 1004020442, 1438550874
| broadcast-igmp-discovery:
|   192.168.0.13
|     Interface: wlp2s0
|     Version: 3
|     Group: 239.255.255.250
|       Mode: EXCLUDE
|       Description: Organization-Local Scope (rfc2365)
|     Group: 224.0.0.251
|       Mode: EXCLUDE
|       Description: mDNS (rfc6762)
|   169.254.126.201
|     Interface: wlp2s0
```

Fig. 6. Nmap Vulnerability Results.

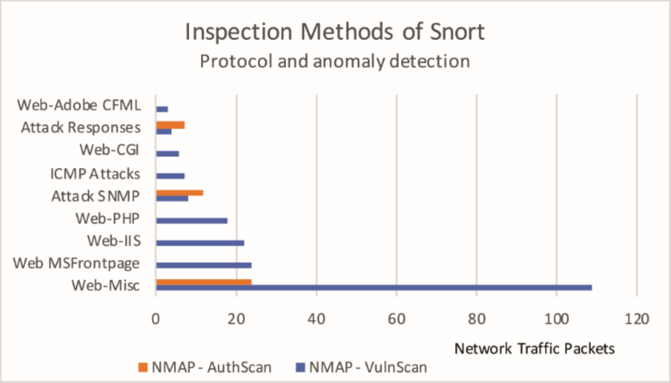


Fig. 7. Traffic Generated and Vulnerabilities Detected by Scanning.

The next step was developing the brute force attack, which consists of using a dictionary attack for breaking the password of a user account in the attacked system.

```
#Msfconsole
#Search ssh
#Use auxiliary/scanner/ssh/ssh_login
#Show options
#Set BLANK_PASSWORDS true
#Set PASS_FILE /root/Escritorio/pass.txt
#Set USER_FILE /root/Escritorio/users.txt
#Set RHOSTS 192.168.0.9
#Run
```

Intrusion attempts, generated by Metasploit, detected and blocked by the Snort tool.

Fig. 8 shows the amount of traffic generated by the attack trying to hack and block the web service this traffic was blocked by the IDS/IPS.

Fig. 9 shows the amount of traffic generated by trying to hack and compromise the ssh service, using a keys dictionary. The metric was the traffic generated by the Metasploit tool.

To configure the IPS is necessary to activate two basic elements, the whitelist (allowed hosts) and the blacklist (banned attacker hosts). Finally, add the preprocessing directives so that the IDS automatically applies the rules:

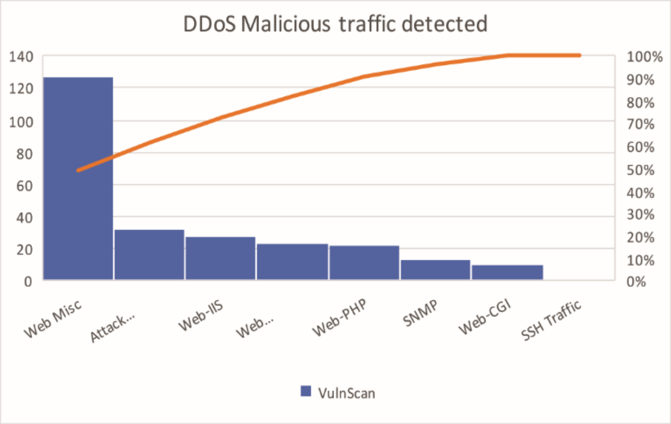


Fig. 8. Result of the DDoS Attack for Web Service Detected with Snort.

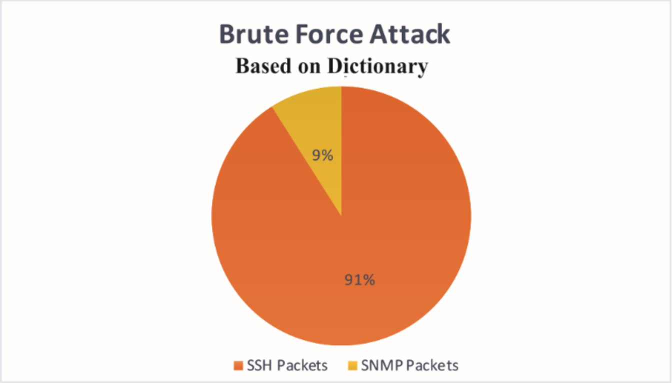


Fig. 9. Attack on SSH Service Detected with Snort Tool.

```
//Configure Snort IPS (edit snort.conf)
#sudo vi /usr/local/etc/snort/snort.conf
Add -ipvar HOME_NET 192.168.0.0/24 –make this match your
internal network;
Add -ipvar EXTERNAL_NET !$HOME_NET //IPs of network
home
Add -var RULE_PATH rules
Add -var WHITE_LIST_PATH rules //IPs from host allowed
Add -var BLACK_LIST_PATH rules
Add this to the end after “decompress_depth 65535”
max_gzip_mem 104857600
-Add this line -output unified2: filename snort.log, limit 128
-delete or comment out all of the “include $RULE_PATH” lines
except:
#include $RULE_PATH/local.rules
#include $RULE_PATH/snort.rules–add after local.rules7.
//Now the following rules are uncommented, for:
preprocessor normalize_ip4
preprocessor normalize_tcp: ips ecn stream
preprocessor normalize_icmp4
preprocessor normalize_ip6
preprocessor normalize_icmp6.
```

VI. Results

As described, the trend of using IoT components in industry 4.0 is perhaps the most complex challenge for security, it is changing the way that information is generated and exchanged. The problem observed is that due to the rapidity with which IoT devices are produced and used, due to demand, the communication security between the components is not properly established. The origin of the threats in the IoT derives from lack of training, investment, staff capacity, and security schemes.

It has demonstrated that the components, which were integrated into an IoT system, in their wild they were weak about security characteristics. They were not built considering security parameters. So we made some penetration testing to demonstrate their behavior under some kind of attacks. We used at the same time an IDS/IPS tools in these tests, to demonstrate that is necessary to support and help an IoT system. The attacks went into the system in a direct way, only the IDS/IPS system helped to detect them. With Nmap, we obtained a list of vulnerabilities in the prototype then we did a DDoS attack on port 80, we also used brute force attack trying to guess a user’s password over SSH service. In these attacks that were tested, Snort detected the unusual traffic and behavior and it sent messages and warnings.

It is worth mentioning that initially in the development of the project, the manufacturing state of the prototype components was analyzed and the sensors and actuators were implemented. With this, it was determined that despite these components, it was advisable to improve its security with specialized tools; this phase of work was done with the Snort software. The Sagan tool proved to be more demanding in the use of memory and processing, recommending its implementation in multi-threaded architectures that can support the demand.

VII. Conclusion and future work

In this work, it has demostrated how strong or weak are the IoT components against some common attacks which can be found on the Internet. We found that the components of an IoT system were built with no considerations on security schemas and response against common attacks. So, is recommend installing an IDS/IPS to secure an IoT system, to prevent and warning against some cyber attacks.

The implementation of Snort as an intruder detection system allowed real-time detection of port scanning and attempts to breach the system from other hosts; providing an opportunity to measure how systems are compromised. This provides a new opportunity for an investigation to model this behavior. Finally, it is important considering the stabilization of rules for IDS/IPS so that the system permits secure communication without repudiation.

Our future works can treat over new penetration tests, set up new rules in the IDS/IPS, encrypted messages among wireless components and integrate all components and tools to implement a platform of IoT secure scheme.

Acknowledgments

We thank Tecnológico Nacional de México/Instituto Tecnológico de Morelia, as well as Universidad Nacional Autónoma de México, for the support granted for this research, together with project 5774.19-P "Development and implementation of a secure IoT architecture, based on penetration tests, using fuzz models and detection systems with an IDS and IPS" Spanish translate "Desarrollo e implementación de una arquitectura IoT segura, basado en pruebas de penetración, utilizando modelos fuzz y sistemas de detección con un IDS e IPS". We also thank professors Abel A. Pintor from ITMorelia and from IIES-UNAM who helped the development of the prototype, Diego Cabrer, Atzimba López, Alberto Valencia, and Yumi Tzib.

References

[1] Haroon A., Naeem W., Shah M. A., Kamram M., Asim Y. & Javaid Q. "Constraints in the IoT: The World in 2020 and Beyond". International Journal of Advanced Computer Science and Applications,Vol. 7, No. 11, 2016.

[2] Santoso F. K. & Vun N. C. H. "Securing IoT for smart home system". International Symposium on Consumer Electronics (ISCE), 2015. IEEE. ISBN: 978-1-4673-7365-4. DOI: 10.1109/ISCE.2015.7177843.

[3] Qinghe D., Houbing S. & Xuejie Z. "Social-Feature Enabled Communications Among Devices Toward the Smart IoT Community". IEEE Communications Magazine. Volume 57 Issue 1. January-2019. DOI: 10.1109/MCOM.2018.1700563.

[4] Rouhiainen Tuukka. "Scanning the Internet to find security loopholes". Proceedings of the Seminar in Computer Science: Internet, Data and Things (CS-E4000). Computer Science at Aalto University. 2018.

[5] Mosca, D. "Hacking the internet of things just got easier – it's time to look at your security". [Online]. Available: https://www.computerweekly.com/opinion/Hacking-the-Internet-of-Things-just-got-easier-its-time-to-look-at-your-security. [Accessed April, 2019].

[6] Rishi Rahul & Saluja Rajeev. "Future IoT". Ernst & Young Associates LLP, Published in India. 2019

[7] Lovejoy C., Watson R. & Pizzala J.. "Internet of Things and Operating Technology Security". [Online]. Available: https://www.ey.com/en_gl/advisory/iot-operating-technology-security. [Accessed August 2019].

[8] Amazon Web Services, "Internet de las cosas, Plataforma como servicio AWS IoT," 2019. [Online]. Available: https://aws.amazon.com/es/iot/. [Accessed: 23-Apr-2019].

[9] Microsoft. (2019). IoT Hub | Microsoft Azure. Retrieved April 23, 2019, from https://azure.microsoft.com/es-mx/services/iot-hub/

[10] Oracle, "Internet of Things | Oracle Cloud," 2019. [Online]. Available: https://cloud.oracle.com/iot. [Accessed: 23-Apr-2019].

[11] IBM, "IBM Watson Internet of Things (IoT)", 2019. [Online]. Available: https://www.ibm.com/mx-es/internet-of-things. [Accessed: 23-Apr-2019]

[12] G. I. Xively, "IoT Platform for Connected Devices", 2019. [Online]. Available: https://xively.com/. [Accessed: 23-Apr-2019].

[13] Samsung Co., "IoT Cloud Platform, Samsung ARTIK cloud services", 2019. [Online]. Available: https://artik.cloud/. [Accessed: 02-May-2019].

[14] Altair Engineering Inc., "Altair SmartWorks" 2019. [Online]. Available: https://www.altairsmartworks.com/index.php/. [Accessed: 02-May-2019].

[15] Adafruit, "Welcome to Adafruit IO" 2019. [Online]. Available: https://io.adafruit.com/. [Accessed: 02-May-2019].

[16] Ubidots, "IoT platform Ubidots" 2019. [Online]. Available: https://ubidots.com/. [Accessed: 02-May-2019].

[17] MyDevices, "The IoT Solutions Company" 2019. [Online]. Available: https://mydevices.com/. [Accessed: 02-May-2019].

[18] Macchina.io, "IoT Edge Device Software Development and Secure Remote Access Solutions", 2019. [Online]. Available: https://macchina.io/. [Accessed: 02-May-2019].

[19] ThingSpeak, "IoT Analytics, ThingSpeak Internet of Things", 2019. [Online]. Available: https://thingspeak.com/. [Accessed: 02-May-2019].

[20] Arduino, "Arduino" 2019. [Online]. Available: https://www.arduino.cc/en/IoT/HomePage. [Accessed: 02-May-2019].

[21] Tomas Zitta, "Penetration Testing of Intrusion Detection and Prevention System in Low-Performance Embedded IoT Device" (2018). IEEE-Xplore. Retrieve: https://ieeexplore.ieee.org/document/8624734. [Accessed: July-2019].

[22] Cisco Systems, "Snort - Network Intrusion Detection & Prevention System." [Online]. Available: https://snort.org/. [Accessed: 13-Jul-2019].

[23] Project Suricata, "Suricata, Open Source IDS/IPS/NSM engine." [Online]. Available: https://suricata-ids.org/. [Accessed: 13-Jul-2019].

[24] The Zeek Network Security Monitor, "The Zeek Network Security Monitor." [Online]. Available: https://www.zeek.org/index.html. [Accessed: 13-Jul-2019].

[25] OSSEC Project Team, "OSSEC -World's Most Widely Used Host Intrusion Detection System-" [Online]. Available: https://www.ossec.net/. [Accessed: 13-Jul-2019].

[26] Tripwire, "Cybersecurity and Compliance Solutions". [Online]. Available: https://www.tripwire.com/. [Accessed: 14-Jul-2019].

[27] Linux, "Intrusion detection with AIDE". [Online]. Available: https://www.linux.com/news/intrusion-detection-aide. [Accessed: 14-Jul-2019].

[28] Samhain design labs, "Samhain Labs" [Online]. Available: https://www.la-samhna.de/samhain/index.html. [Accessed: 18-Jul-2019].

[29] Fail2ban Project, "Fail2ban." [Online]. Available: https://www.fail2ban.org/wiki/index.php/Main_Page. [Accessed: 19-Jul-2019].

[30] Q. I. S. Sagan Project, "The Sagan Log Analysis Engine | Quadrant Information Security." [Online]. Available: https://quadrantsec.com/sagan_log_analysis_engine/. [Accessed: 19-Jul-2019].

Intrusion Detection System based on the SDN Network, Bloom Filter and Machine Learning

Traore Issa[1]
Institute of Mathematics research (IMAR)
Computer Science Laboratory Telecom Networks
Felix Houphouet-Boigny University
08 BP 2035 Abidjan 08, Cote d'Ivoire

Kone Tiemoman[2]
Virtual University of Cote d'Ivoire
Computer Science Laboratory Telecom Networks
28 BP 536 28 Abidjan
Cote d'Ivoire

***Abstract*—The scale and frequency of sophisticated attacks through denial of distributed service (DDoS) are still growing. The urgency is required because with the new emerging paradigms of the Internet of Things (IoT) and Cloud Computing, billions of unsecured connected objects will be available. This document deals with the detection, and correction of DDoS attacks based on real-time behavioral analysis of traffic. This method is based on Software Defined Network (SDN) technologies, Bloom filter and automatic behaviour learning. Indeed, distributed denial of service attacks (DDoS) are difficult to detect in real time. In particular, it concerns the distinction between legitimate and illegitimate packages. Our approach outlines a supervised classification method based on Machine Learning that identifies malicious and normal packets. Thus, we design and implement Defined (IDS) with a great precision. The results of the evaluation suggest that our proposal is timely and detects several abnormal DDoS-based cyber-attack behaviours.**

Keywords—Distributed denial of service; intrusion detection software; software defined network; machine learning; synchronize; acknowledgment; bloom filter

I. Introduction

Over the past decade, DDoS attacks have been a powerful threat to the security of many Internet service providers, and have resulted in economic losses for them. DoS attacks cause a denial of service to legitimate requests by depleting network resources and services. To maximize impact, the attack will be launched from distributed sources, called attacks through denial of distributed service. In most cases, these attacks are launched by botnets. The largest DDoS attack on the latest records occurred in February 2018 as revealed by the Git Hub. The attack came from more than thousand different European Union countries out of tens of thousands of single endpoints. This was the one amplification attack using Memcached technology that peaked at 1.35Tbps. Another major DDoS attack is the Mirai [1] botnet attack that was used in a high volumetric DDoS of about 1.1 Tbps that destroyed a large part of Dyn's database in October 2016. Mirai has successfully ordered nearly 100,000 robots by exploiting the low security of cameras, home routers, digital recorders and printers with default credentials used for their telnet ports.

Many methods are used to block DDoS attacks, including some:

- The signature-based approach: it requires an a priori knowledge of the elements related to the signature of attacks, see SNORT [2]. Signatures are manually built by security experts. The authors of [3] analyze previous attacks to look for a match with incoming traffic to detect intrusions. Signature-based techniques are only effective in detecting the traffic of known DDoS attacks; while new attacks or even slight variations of old attack go unnoticed.
- Anomaly-based detection: the anomaly-based system uses a different method. It treats any network connection that violates the normal profile as an anomaly. The anomaly is revealed if incoming traffic deviates significantly from normal profiles, see [4] and [5]. To detect DDoS attacks, it is first necessary to know the overall normal behaviour of the system traffic and then to find deviations from this behaviour. The anomaly-based technique can detect new attacks. However, it can initiate many false alarms.
- Packet filtering: packets entering and leaving the network protect the network against attacks from any source. This technique uses server firewalls, router based packet filtering [6]. This requires the installation of filter input and output packets on all routers. It is used to filter the spoofed IP address, but approaches to prevent it need a global implementation that is not practical [7].

In this article, we set up IDS capable of detecting anomalies based on Machine Learning techniques. The volume of data to be studied is enormous, so we use SDN technology for efficient data processing. We also used the Bloom filter, which is a probabilistic structure for storing and accessing data efficiently. This document is structured as follows: Section 2 describes some approaches used to solve DDoS attack problems. Section 3 outlines our method of resolution and then Section 4 illustrates the results and discussion. Finally, the conclusion is presented in Section 5.

II. Related Work

An attack through denial of Distributed Service (DDoS) is a flood attack using several controlled sources, called Botnets or Zombies, to disable a service and prevent legitimate users from using it.

A. *How a SYN Flood Attack Works*

SYN flood attacks work by exploiting the process of establishing a TCP connection. Under normal conditions, the TCP connection has three distinct processes for establishing a connection.

- First, the client sends a SYN packet to the server to establish the connection.
- The server then responds to this initial packet with a SYN / ACK packet, in order to acknowledge receipt of the communication.
- Finally, the client sends back an ACK packet to acknowledge receipt of the packet from the server. After completing this sequence of sending and receiving packets, the TCP connection is open and capable of sending and receiving data.

To create a denial of service, an attacker exploits the fact that after receiving an initial SYN packet, the server responds with one or more SYN / ACK packets and waits for the last step of making contact, see Fig. 1.

- The attacker sends a high volume of SYN packets to the target server, often with spoofed IP addresses.
- The server then responds to each connection request and leaves an open port ready to receive the answer.
- The server waits for the last ACK packet, which never arrives, the attacker continues to send more SYN packets. The arrival of each new SYN packet forces the server to temporarily maintain a new port connection open for a period of time. Once all available ports have been used, the server can no longer operate normally.

By repeatedly sending SYN initial connection request packets, the attacker is able to overwhelm all available ports on a target server computer, resulting in the target device responding to legitimate traffic slowly or not at all.

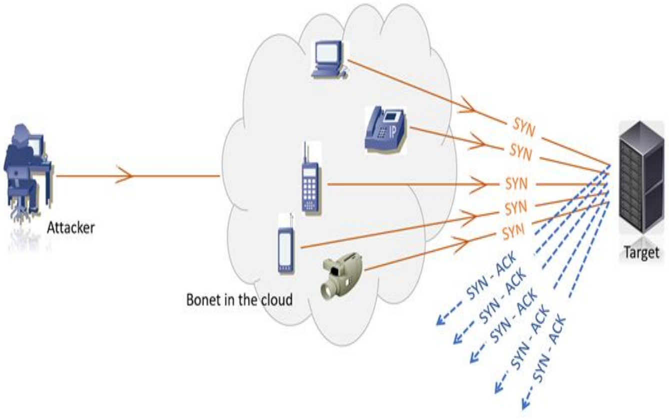


Fig. 1. DDOS Attacks Architecture.

B. *Defensive Mechanisms*

The DDoS defence mechanisms [7] and [8] are classified into two of the four main categories: source identification, attack detection, reactivity and attack prevention.

- The identification of the attack source uses mechanisms to find the source IP address to block them. The trace back investigation [9] is the most popular mechanism to identify the attacker's source IP address.
- Attack detection detects the DDoS attack when it occurs. Some defence mechanisms are MULTOPS [10] and anomaly-based detection [11].
- Reactivity to an attack aims to reduce or eliminate the effect of the attack [12]. Two main approaches [13] are taken to respond to DDoS attacks and network resource management.
- Attack prevention tries to stop the attack before it occurs. The attack does not reach the target host. Some examples are Ingress/Egress 6 filters on routers, packet filtering on routers [14], and automatic learning to detect anomalies.

The first three methods have proven their effectiveness, but they are reactive, the damage has already been done. The latter approach is proactive. It has proven its usefulness. However, it requires a lot of calculation, large data to store and managed. Our approach is proactive, because the objective is to ensure quality of service without it being blocked by a DDoS attack. Thus, this paper proposes a new IDS network paradigm based on Machine Learning to solve the network control problem. The main contribution of this article can be summarized as follows:

- Use SDN [15], automatic learning and Bloom filter [16] to set up a high-performance network and effective real-time security,
- Provide a DDoS attack detection architecture by leveraging incoming flow monitoring capability to filter traffic and establish legitimate TCP connections.

Implement a proactive IDS capable of automatically making decisions related to several behavioural parameters. These decisions are based on the set of rules predefined by the administrator.

To do this, we ensure on the one hand, the storage of IP packet information in a compact way in the address intended for this purpose, and on the other hand, the calculation of automatic learning on dedicated servers in an architecture combining SDN and Machine Learning.

III. Methodology

A DDoS attack caused by botnets generates a lot of resources; a traditional router can hardly predict the attack. The router performs calculations to route packets, assigns priorities, makes routing decisions, and enforces rules specified by the administrator. Thus, it can only be changed manually by the administrator, which obviously takes time and does not lend itself to rapid context changes. With the SDN, these changes are automated and even programmable.

A. SDN Architecture

The data plan and the control plan are increased tenfold. Thus, the administrator defines the rules in the controller, and they are instantly transmitted in the network equipment.

Fig. 2 illustrates the SDN architecture, which consists of three layers. The lowest layer is the infrastructure layer, also called the data plan. It includes the elements of the transfer network. The responsibilities of the routing plan are mainly data transfer, as well as monitoring of local data transmission, information and statistical collection.

The layer above is the control layer, also called the control plan. He is responsible for the programming and management of the routing plan. To this end, it uses the information provided by the transmission plan and defines the operation and routing of the network. It includes one or more controllers that communicate with the elements of the transmission network through standardized interfaces, known as southbound interfaces.

The application layer contains network applications that can introduce new network functionality through APIs, such as security and management, transfer schemes or control layer support in network configuration. It has an abstract and global view of the network from the controllers and uses this information to provide appropriate advice to the control layer. The interface between the application layer and the control is called the northward interface.

Northbound APIs can be used to facilitate innovation and enable efficient network orchestration and automation to align with the needs of different applications through SDN network programmability. We will use this property of the application layer to implement a TCP flooding attack detection module.

B. Bloom Filter: Data Storage

A Bloom filter is a probabilistic structure that allows the efficient storage of a set of elements [16]. It consists of a vector of m bits and a set of k hash functions. Initially all bits are at 0. To insert an element into a filter, the k hash functions are calculated on it and their results determine the positions of the bits set to 1. To test if an element belongs to a filter, simply calculate its k hash functions on the element and check if all the bits at the corresponding positions are at 1. If not, it is certain that the element is not in the filter.

However, there is still a probability of false positives: it is possible that all the corresponding bits have been set to 1 by other stored elements, and therefore to detect a tested element when it is not in the filter. The probability of false positives as a function of the number of elements stored n and the size of the filter is given by the formula:

For any pair of integers (m, k) :

$$P \gg (1- e- \frac{km}{n})^{k} \tag{1}$$

To maintain the same false positive rate with an increasing number of elements, it is necessary to increase the number of bits and hash functions, which results in higher memory consumption and increased computing costs.

The Bloom filter stores in the form of a table of bits that represent the IP addresses considered malicious, see Fig. 3.

Consider $\mathcal{F}$ = {ip_1, ip_2, ..., ip_n}, the n IP addresses that describe the array of n bits. Initially all bits are at 0.

Let $\mathcal{H}$ = {h_1, h_2, ..., h_k}, all the independent hash functions stored in p.

For each ip_x on $\mathcal{F}$:

$$h_j(ip_x) = 1 \text{ for } 1 \leqslant j \leqslant k. \tag{2}$$

To check if an attack suspect ip_x address is in $\mathcal{F}$. We check that all $h_j(ip_x) = 1$, otherwise $h(ip_x) = 0$ is not malicious. This process can generate false positives. In other words, it can happen that for an ip_x address we have $h_j(ip_x) =1$ while it is not malicious.

In our approach, false positives are negligible because the probability of their existence is low. Indeed, let us consider m, the size of the Bloom filter, n the number of hash functions. Let X be a random variable representing all the bits. Thus, the false positive rate can be evaluated by:

$$P(X = 0) = (1 - \frac{1}{m})^{nk} \tag{3}$$

In [17] have shown that this rate is very low because:

$$k = \ln(2)\frac{p}{n} \tag{4}$$

When k=10 and p=20n, the probability of a false positive is 0.0000889. This result justifies the use of the Bloom filter in the detection of DDoS attacks based IDS architecture.

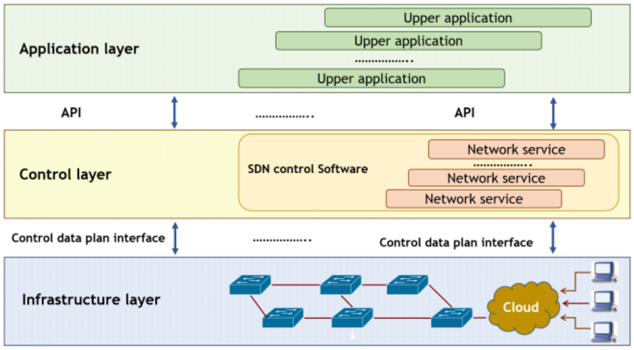


Fig. 2. Architecture SDN.

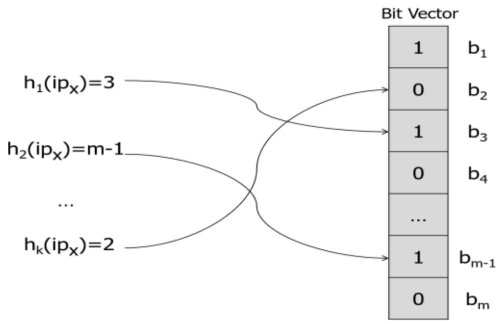


Fig. 3. IP Address Hash Functions.

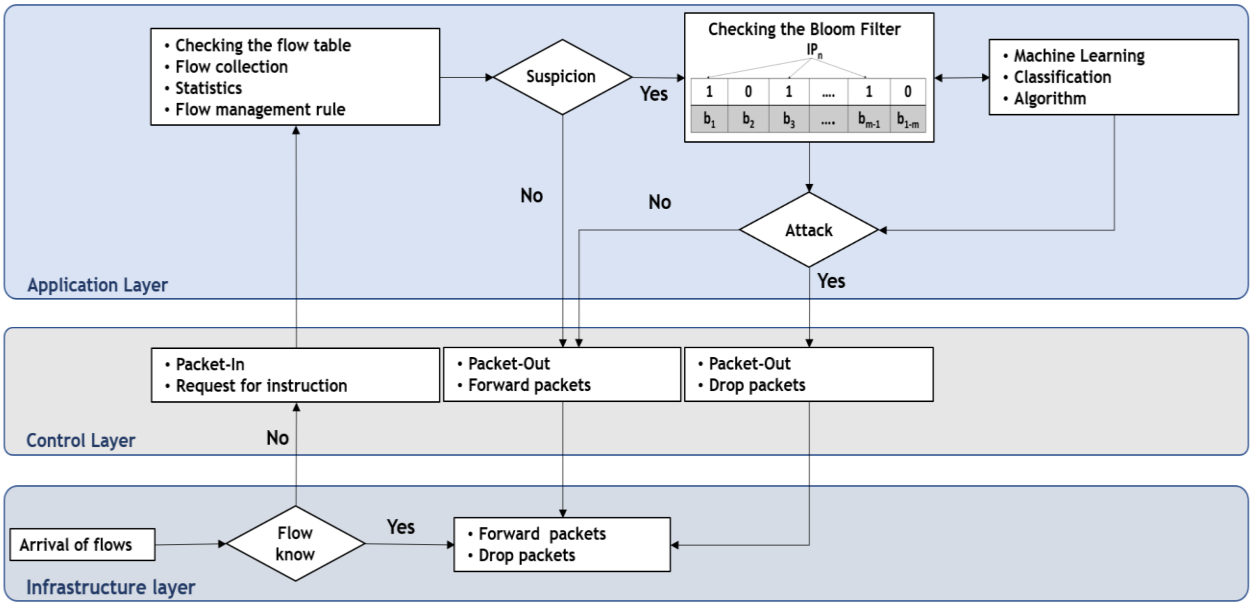


Fig. 4. IDS Architecture for DDOS Attack Detection.

In our approach, we have combined the advantages of the behavioural filter and Machine Learning as shown in Fig. 4.

In our method, network traffic must be collected from switches and then used to build the drive and classification set. The management of packets entering the network is presented in Fig. 4. When a new packet arrives at the switch, if it belongs to an existing flow in the flow table, it updates the flow statistics otherwise, a "Packet-In" message is sent to the Openflow controller. The controller responds with a "Packet-out" message on the attitude to be followed according to the pre-established rules.

Switches in the data plan uses tables to route packets. This is possible by using entries in flow tables and a packet processing process. According to [17] an entry in the flow table consists of seven fields: Match Fields, Priority, Counters, Instructions, Timeouts, Cookie, Flags.

The Counters field allows you to know the total number of packets processed for an entry. Counters can be maintained for each flow table, number of packets or bytes, flow entry, port, queue, duration during which the entry was activated.

C. ADIS: DDoS Attack Detection Module

In this section, we describe the mechanism for detecting and preventing attacks.

1) DDoS defense architecture

a) The data plan: When a new packet arrives at the switch, it checks whether the packet header matches an entry in its flow table.

If it finds an entry, it processes the packet as defined in the corresponding entry. Otherwise, he forwards the packet to the controller in order to receive instructions after a thorough investigation.

b) The control plan: After receiving a new packet, the controller processes, calculates and creates a new flow entry, which it then sends to the switch. The switch receives the message from the controller, adds the new entry to its flow table, and manages the packet as defined in the entry [18]. When the packet is unknown to the controller, the Openflow protocol sends the packet header to the Identity Attack and Storage Detection (ADIS) module in the application plan.

c) The application plan: The ADIS module of the application layer is designed to analyze the SDN network flow tables and collect traffic flows by inspecting the IP header {src_ip, src_port, dst_ip, dst_port, protocol}. Each flow can be represented by a set of statistical characteristics, such as DurationSeconds, packetCount and byteCount, etc. The ADIS module checks if the IP address of the packet is stored in the Bloom filter (attacker database). Failing this, a deep analysis based on the number of packets sent per second by the source to the Openflow switch classifies the category of the source IP address. In the following section, we propose a classification algorithm.

2) Data classification algorithm: In order to detect the DDoS attack, the IDS must be supplied with traffic information related to the following parameters: src_ip, src_port, dst_ip, dst_ip, dst_port, protocol, DurationSeconds, packetCount and byteCount, etc. It uses a Machine Learning (ML) classification model to detect attack activity. In our example, we will use the following models : linear discriminant analysis (LDA), k-nearest neighbors (KNN) and Support vector machine (SVM).

These models can learn the pattern with few training samples and produce an accurate classification by reducing false positives.

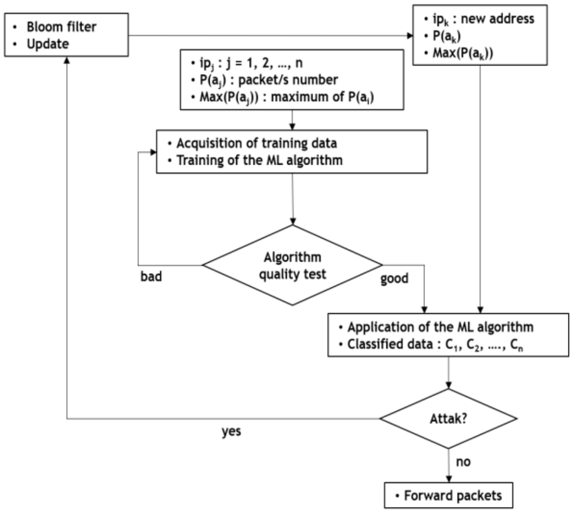

• Bloom filter
• Update
• ip_k : new address
• P(a_k)
• Max(P(a_k))
• ip_j : j = 1, 2, ..., n
• P(a_j) : packet/s number
• Max(P(a_j)) : maximum of P(a_i)
• Acquisition of training data
• Training of the ML algorithm
Algorithm quality test
bad
good
• Application of the ML algorithm
• Classified data : C_1, C_2, ..., C_n
Attak?
yes
no
• Forward packets

Fig. 5. Classification Algorithm.

Using the recognized model in Fig. 5, the IDS perform category prediction for new unknown traffic samples. The classification results for the test data points would be either normal or attack.

This security approach integrates a cognitive layer above the control layer and thus allows artificial/automated intelligence to be naturally introduced into network management. In this context, IDN (Intelligence Defined Networks) approaches are presented as an evolution of the SDN.

D. *Implémentation*

In this section, we present our experimental design and describe the experimental results of our methods. Then we compare the models against the error rates as defined above.

1) Simulation environment: To evaluate our approach, we chose the Kali-Linux simulator. The official version of Kali-Linux has several modules to simulate hacking and computer attacks. In our experience, we install hping3 which is a package available by default on Kali-Linux. The packet flow is captured by Wireshark and Tshark for analysis on the I/O graph. This data is collected at the SDN Openflow switches for analysis.

2) Data collection and analysis: Incoming packets are captured by Wireshare, two processes are performed. The first consists in performing an analysis of SYN, SYN-ACK and ACK exchanges. The second based on automatic learning allows a classification of the captured data into clusters classified according to the analysis.

In Section IV, we use R Studio software to classify the data. We use information on initiated connections: source and destination addresses, protocols, connection time, packet size, information on SYN, ACK, sequence numbers, etc.

The purpose is to classify the addresses that have initiated a SYN connection according to ACK responses or not. Indeed, the imbalance of flows can facilitate the detection of DDoS. In a normal packet flow, the number of incoming packets corresponds to the number of outgoing packets over a given period of time. For example, each packet in TCP connection is normally acknowledged. However, during the attack the number of incoming and outgoing packets is unbalanced. The second treatment makes it possible to update the addresses considered malicious in the Bloom filter according to Fig. 3.

IV. Results and Discussions

The experimental results produced the following Fig. 6 shows the curve representing the number of packets sent per second to the server.

Fig. 7 shows a point cloud of the number of packets received per second. We will use this graph to classify the data into three classes. One class of packets is considered normal, another is considered suspicious and the last one is considered malicious.

The Machine Learning method makes it possible to dissociate normal, suspicious and attacking IP addresses. On Fig. 8, Fig. 9 and Fig. 10 below, we can see the LDA, KNN and SVM methods resulting from the relationships between training data and test files. The SVM is the approach that achieves the best results.

In our approach, learning will be performed several times on 75% of the original classification data set. Training performance is given by train: mmce. For each iteration, the formed model will be tested on a subset of training (75%) and a subset of tests (25% of the original data set).

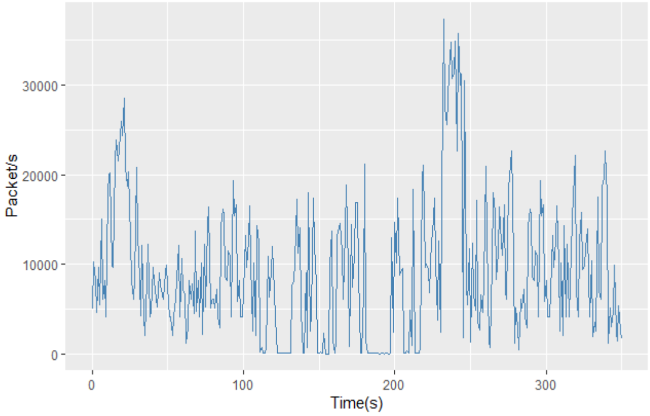


Fig. 6. Number of Source Packets / s.

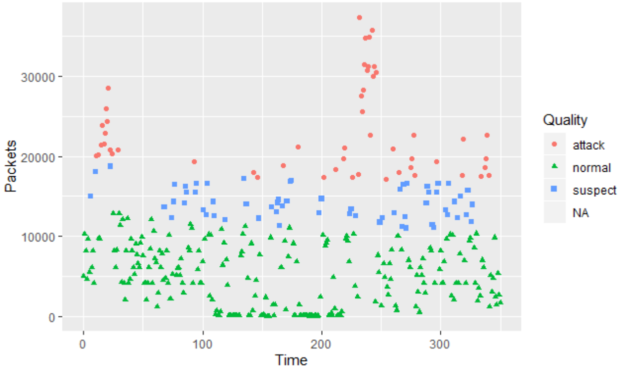


Fig. 7. Point Cloud of Source Packets per Second.

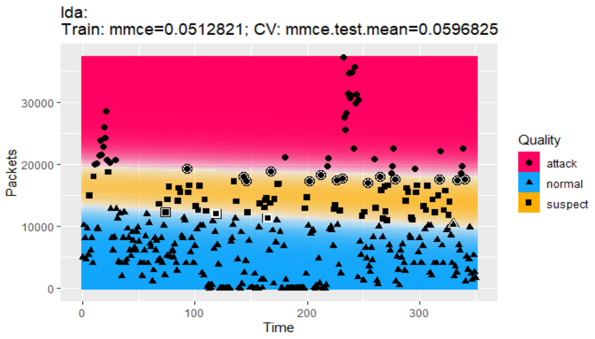


Fig. 8. LDA Classification.

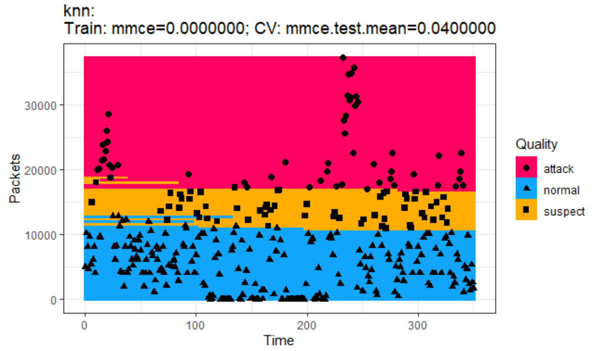


Fig. 9. KNN Classification.

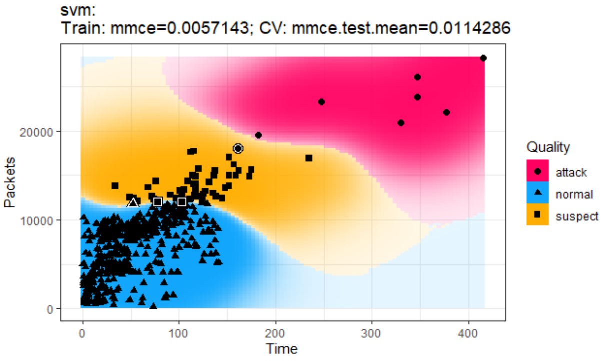


Fig. 10. SVM Classification.

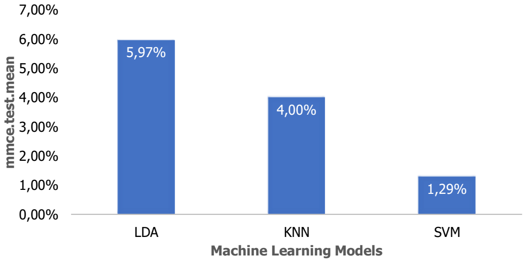


Fig. 11. Average Model Error.

The performance of the model will be measured by the average misclassification error (mmce.test.mean) of each model LDA, KNN and SVM, see Fig. 11.

Thus, the classification methods exposed made it possible to simulate the requests sent to a server. Our method manages to classify the addresses from the packets involved with an acceptable error rate. Indeed, according to Fig. 11, the error rate in classification is 1.29%. Thus, our proposal allows classifying the IP addresses of packets resulting from DDoS attacks and normal packets.

V. Conclusion and Future Work

In the age of large data, with the exponential growth of network traffic, network attacks are becoming more diversified and sophisticated. In this document, we use the SDN architecture and Bloom filter to ensure the computing power of Openflow controllers, storage and data access. The Machine Learning algorithm allowed the IDS to detect and suppress DDoS attack traffic. We focused on analyzing the data in order to avoid false positives as much as possible. Thus, the network application classifier based on the SVM learning model allows the expected objectives to be achieved with greater precision.

As part of our future work, we plan to extend our analysis to Machine Learning, in order to find an appropriate model to ensure a higher accuracy rate and eliminate false alarms.

Acknowledgment

The Publication of this research was supported by the Mathematics Research Institute.

References

[1] C. Kolias, G. Kambourakis, A. Stavrou, J. Voas, ‘‘DDoS in the IoT: Mirai and other botnets,’’ IEEE Computer, 50(7), (2017), p80-84. https://doi.org/10.1109/MC.2017.201.

[2] M. Roesch, ‘‘Snort-Lightweight Intrusion Detection for Networks,’’ In: Proceedings of the USENIX Systems Administration Conference (LISA November 1999), pp. 229–238.

[3] V. Paxson, ‘‘Bro: A System for Detecting Network Intruders in Real-Time,’’ International Journal of Computer and Telecommunication Networking 31(24), (1999), pp2435–2463.

[4] U. Dincalp, M. Serdar, O. Sevine, E. Bostanci, I. Askerzade, ‘‘Anomaly Based Distributed Denial of Service Attack Detection and Prevention with Machine Learning,’’ 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2018.

[5] P. Ombase, P. Scholar, S. Bagade, N. Kulkarni, A. haisgawali, ‘‘DoS Attack Mitigation Using Rule Based and Anomaly Based Techniques in Software Defined Networking,’’ In Proceedings of the 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 23–24 November 2017; pp. 469–475.

[6] K. Park, H. Lee, ‘‘On the effectiveness of route-based packet filtering for distributed DoS attack prevention in power-law Internets,’’ In: Proceedings of the ACM SIGCOMM 2001, Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, ACM Press, New York (2001), pp 15–26.

[7] T. Peng, C. Leckie, K. Ramamohanarao, ‘‘Protection from distributed denial of service attack using history-based IP filtering,’’ In: Proceedings of IEEE International Conference on Communications (ICC 2003), Anchorage, AL, vol. 1, pp. 482–486.

[8] E. Fenil, P. Mohan Kumar , ‘‘Survey on DDoS defense mechanisms,’’ wiley, wileyonlinelibrary.com/journal/, Décembre 2018.

[9] V. Chidri, V. Balasubramani, S. Sadath Ali, S. Shrikrishna Hegde, P. Sadanand, ‘‘A Survey on Distributed Denial-of-service Attacks and Defense Mechanisms,’’ JETIR1504089 Journal of Emerging Technologies and Innovative Research (JETIR) www.jetir.org, avril 2015.

[10] T. Jog, M. Natu, S. Shelke, ‘‘*Distributed capabilities-based DDoS defense,’’* 2015 International Conference on Pervasive Computing (ICPC), janvier 2015, pp. 1–6.

[11] C. Buragohain, M. Kalita Santosh Singh, D. Bhattacharyya, ‘‘Anomaly based DDoS Attack Detection,’’ Chaitanya Buragohain, Manash Jyoti Kalita Santosh Singh, Dhruba K.Bhattacharyya, International Journal of Computer Applications (0975 –8887) Volume 123–No.17, August, 2015, pp35-40.

[12] A. Cardigliano, L.Deri et T. Lundstrom, ‘‘*Commoditising DDoS mitigation*,’’ septembre 2016, p. 523–5282016 International Wireless Communications and Mobile Computing Conference (IWCMC).

[13] N. Lu, S. Su, M. Jing, and J. Han, ‘‘A router-based packet filtering scheme for defending against dos attacks. China Communications, 11(10, 2014), pp136–146.

[14] R. Koning, B. de Graaff, G. Polevoy, R. Meijer, C. de Laat, P. Grosso, ‘‘Measuring the efficiency of SDN mitigations against attacks on computer infrastructures,’’ Future Generation Computer Systems **91**(1), 144{156 (2019)., https://doi.org/10.1016/j.future.2018.08.011, https://doi.org/10.1016/j.future.2018.08.011.

[15] R. Patgiri, S. Nayak, and S. K. Borgohain, “Preventing DDoS using bloom filter: A survey,” ICST Transactions on Scalable Information Systems, vol. 5, no. 19, Article ID 155865, 2018.

[16] C. Tseung, K. Chow, and X. Zhang. ‘‘Anti-DDoS technique using self-learning bloom filter,’’ In Intelligence and Security Informatics (ISI), 2017 IEEE International Conference on, pages 204–204. IEEE, 2017.

[17] P. Cao, ‘‘Bloom filters-the math,’’ University of Wisconsin-Madisson, Madisson (1998). http://pages.cs.wisc.edu/~cao/papers/summary-cache/node8.html.

[18] Open Networking Foundation, ‘‘OpenFlow Switch Specification,’’ Version 1.5.1 (Protocol version 0x06), https://www.opennetworking.org /images//openflow-switch-v1.5.1.pdf.

Development of a Vehicle for Driving with Convolutional Neural Network

Arbnor Pajaziti[1], Xhevahir Bajrami*[2], Fatjon Beqa[3]
Dept. of Mechatronics
Faculty of Mechanical Engineering
University of Prishtina, Prishtina, Kosovo

Blendi Gashi[4]
Faculty of Computer Sciences
University of Prizren
Prizren, Kosovo

***Abstract*—The aim of this paper is the design, simulation, construction and programming of the autonomous vehicle, capable of obstacle avoidance, object tracking also image and video processing. The vehicle will use a built-in camera for evaluating and navigating the terrain, a six-axis accelerometer and gyro for calculating angular velocities and accelerations, Arduino for interfacing with motors as well as with Raspberry Pi which is the main on-board computer. The design of the vehicle is performed in Autodesk Fusion 360. Most of the mechanical parts have been 3D printed. In order to control the chassis of the vehicle through the microcontrollers, the development of the PCB was required. On top of this, a camera has been added to the vehicle, in order to achieve obstacle avoidance and perform object tracking. The video processing required to achieve these goals is done by using OpenCV and Convolutional Neural Network. Among other objectives of this paper is the detection of traffic signs. The application of the Convolutional Neural Network algorithm after some of the examinations made has shown greater precision in recognizing STOP traffic sign of different positions and occlusion ratios, and finding the path for the fastest time.**

Keywords—Image processing; traffic sign; object tracking; autonomous vehicle; convolutional neural network

I. Introduction

In recent years, autonomous vehicles have become of great interest to the research and industrial communities. The literature related to autonomous vehicles takes mainly two directions hardware and software developments. For the full functionality of autonomous vehicles, there is no need for more hardware. Software and testing is where much work needs to be done. Many researchers have been oriented in designing the small vehicle prototypes due to cost effectiveness and reduction of software testing with different sensors and algorithms.

The Raspberry Pi as a processing chip has been used to build a monocular vision autonomous car prototype. An HD camera along with an ultrasonic sensor has been used to provide necessary data from the real world to the car [1].

A method for autonomous control and decision making and reporting system, the types of mini robots contains self-neural schema framework for autonomous control has been proposed by [2].

The electronic design and motion planning of a robot based on decision making regarding its straight motion and precise turn using Artificial Neural Network (ANN) has been proposed in [3, 8]. The ANN helps in learning of robot so that it performs motion autonomously. The weights calculated are implemented in microcontroller [3].

Obstacle avoiding technique is very useful in real life, by changing the IR sensor by a kinetic sensor, which is on type of microwave sensor whose sensing range is very high and the output of this sensor vary in according to the object position changes [4].

The purpose of this paper is to build a vehicle that will be driven autonomously through decisions taken by Artificial Intelligence. In more detail, the objectives are:

- Design of the vehicle, and its construction;
- Design and development of the PCB board that will control the vehicle;
- Development of a communication method between the vehicle and the laptop;
- Development of the electronic circuits and program for serial communication between Arduino and Raspberry Pi;
- Creating a server for transmitting video and sensors data from Raspberry Pi to the laptop;
- Detection of traffic signs;
- Stopping the vehicle if there are any obstacles ahead;
- Building a Convolutional Neural Networks (CNN) model for predicting the movement.

Aim of this paper is to get closer in trend with today's companies that produce autonomous cars, one of which is well known, Tesla. But there are also many companies that are working to release autonomous cars on our roads.

The paper is organized in 8 sections, starting with mechanical design, electric circuits, controllers programming, training the models for the detection of traffic signs, programming and control with CNN, data processing method, and, concluding remarks with future work.

*Corresponding Author.

II. Mechanical Design and 3D Modelling by using the Autodesk Fusion 360

For modelling and designing of the parts, the Autodesk software, Fusion 360, has been used. Initially, motors and wheels have been designed by measuring the actual physical parts. Then, the assembly of the motors with wheels and the respective joints have been done. The chassis has been designed in seven different parts, because the printing area of 3D printer used was 200x200 mm, and the dimensions of the model exceed this area. Following this step, the motors and wheels have been assembled into chassis. Other parts needed for the implementation of this project have been designed, and are shown below on Fig. 1.

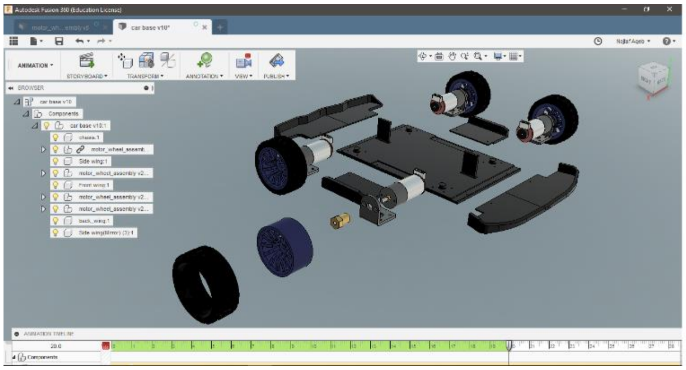


Fig. 1. A View of the Design Process by using the Fusion 360.

A. Assembly of Mechanical Parts

Initially, the motors and wheels were joint together. Then, the printed parts have been glued together and the chassis was completed. After completing the chassis, the next step was to mount the motors and wheel into chassis. This was done using M2.5x6 bolts, and also an adhesive for an additional safety factor.

III. Design of the Electric Circuits of the Vehicle

In order for the mechanic parts to be controlled through electronics, a PCB board was required. Also taking into account the numerous numbers of necessary components, PCB implementation was a good step in eliminating parasitic resistance, as well as the interconnectivity of the components being fixed so as to provide better performance.

Also, the development of PCB board significantly reduced the needed number of wires, compared to the development of the circuit in the breadboard, thus having smaller dimensions and better visual. PCB has been developed as a modular, in case if one of the components is damaged, for various reasons, it can be easily replaced by another.

For designing the circuit, Proteus 8 Professional software has been used. Since the software did not have the component libraries that were used, each component had to be manually measured and added to the program libraries, then the circuit was designed. In the PCB board, components have been placed and then soldered, Fig. 2.

The PCB board components are: 2x boost convertor-XL6009, 2xTP4056, switch, battery terminal, MPU6050, LED indicator, resistors, 3x 2N2222 transistors for RGB, boost-converter for Arduinos, 2x Arduino Pro Mini, buzzer, power amplifier for RGB lights, voltage divider, and distances for Raspberry Pi [5].

It's worth to mention that the L298N motor drivers are not placed on the board but are connected to it, due to their large dimensions.

A. Development of the Transmitter Box

For laptop and Arduino to communicate, the chosen method was through the Radio Modules, in this case through NRF24L01. For the implementation of the transmission box, Fig. 3, a 3D box has been designed and printed initially, and an Arduino Pro Mini, and a NRF24L01 radio module have been installed.

Since the Arduino Pro Mini could not be programmed directly from the laptop, it was necessary to use an FTDI.

The NRFL01+ module with antenna is capable for providing communication up to 1km in open terrain.

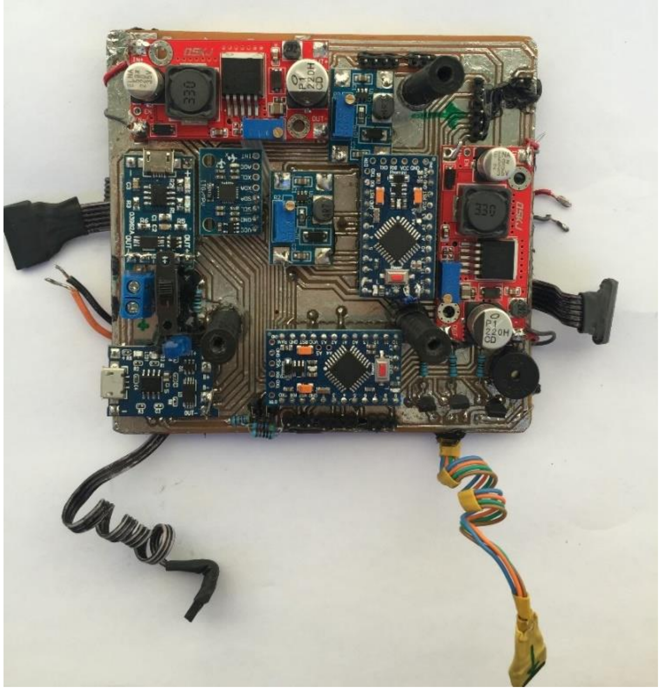


Fig. 2. View of the Completed PCB.

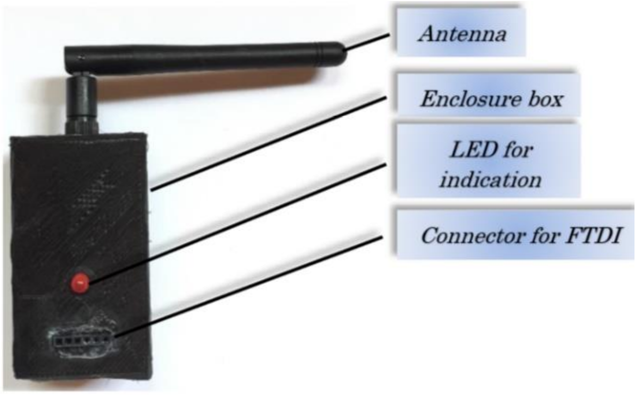


Fig. 3. View of the Transmitter Box.

B. Sensors

Similar to any autonomous vehicle, the number of sensors should be considerable so that the vehicle has sufficient information of the surrounding environment. In this project, a total of five sensors were used, two of which are ultrasonic and, on the output, provide continuous distance information, while the other three are infrared (IR) sensors and output digital information. Two sensors are mounted on the front of the vehicle, Fig. 4, while the other three on the rear, Fig. 5. The reason why there are fewer sensors on the front, is that in front there is also a camera that provides enough information even without sensors.

Also, on the PCB is a gyroscope and accelerometer (MPU6050) as well as each of the DC motors also have the encoder itself, but their information is not used in this project since there is enough information from the camera and the sensors. Using the encoders would have added unnecessary complexity to the project.

C. Batteries

Given the huge power draw from all four DC motors and the large number of electronic components, only two types of batteries were suitable for high discharge rates: Li-Po and Li-Ion. The latter are chosen since they are safer, lighter, and are cheaper and also have a smaller size. The type of batteries used is LGDBB31685, Table I. These batteries are recycled from laptop batteries. Table I shows the specifications of this type of batteries.

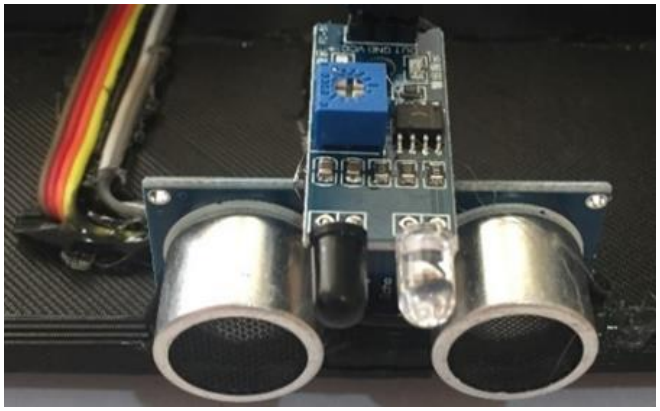


Fig. 4. Sensors on the Front Side of the Autonomous Vehicle.

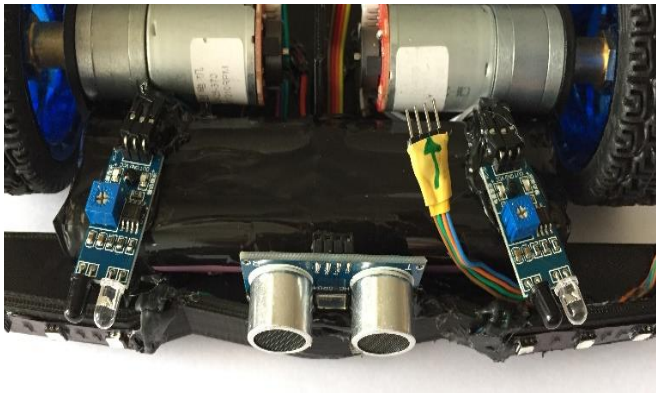


Fig. 5. Sensors on the Rear Side of the Autonomous Vehicle.

The configuration used is 4P1S, so four batteries are connected in parallel holding the nominal voltage at 3.7V, while increasing the capacity to 10400 mAh. Since batteries were recycled, their capacity is measured and the batteries have saved about 80% of their original capacity, which is more than enough for this model vehicle. Whereas, for the supply of Raspberry Pi two batteries are connected parallel with each other along with a voltage booster since Raspberry Pi works with a voltage of 5V, Fig. 6.

A final preview of the vehicle is shown in Fig. 7.

TABLE. I. BATTERY FEATURES

Capacity:	*2600 mAh* Rated
Voltage:	*3.7 V* Nominal
Charge:	*4.2V* Maximum *1250 mA* Standard *2500 mA* Maximum
Discharge:	*3.0 V* Cutoff *500 mA* Standard *3750 mA* Maximum

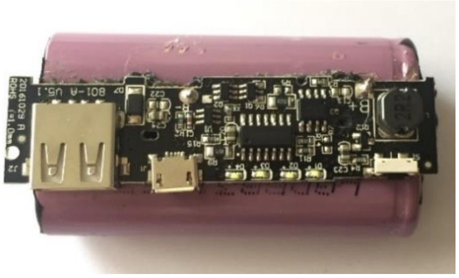


Fig. 6. Raspberry Pi Battery Power Supply.

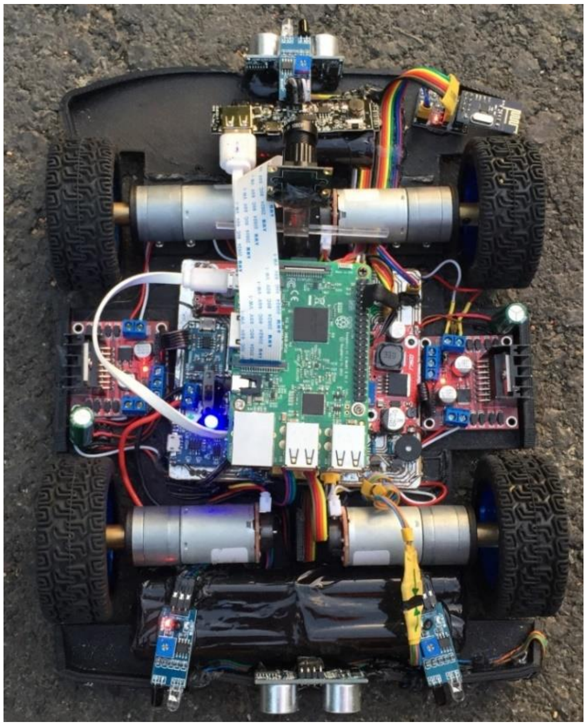


Fig. 7. Top View of the Autonomous Vehicle Model.

IV. Controllers Programming

Using only one microcontroller to process everything in the vehicle was not enough. The way this problem got resolved was by using three microcontrollers (Arduino Pro Mini) and one Raspberry Pi. Two of Arduino's are located on the vehicle's board and share the workload among themselves. First Arduino deals with the reception of radio signals and the controlling all four DC motors, while the other Arduino reads the sensor data, processes them, and transmits them to the Raspberry Pi, also it communicates with the preliminary Arduino through a single pin and controls the RGB lights at the back of the vehicle and even a buzzer. Third Arduino, is in the transmitter box and sends commands to the first Arduino for driving the vehicle. Thus, each Arduino was programmed separately and performs specific task different from one another. For programming the Arduinos, Arduino IDE software was used.

A. Arduino-Arduino Connection on PCB Board

As mentioned earlier, both Arduinos work "at their limits" to meet the requirements of this project. In order to communicate with one another, from the ready-to-use communication methods, SPI could not be used since it requires a lot of pins (such are SCLK, MISO, MOSI, and SS) and is used for communication with radio modules. The other method, the I^{2}C, could not be used because in the Arduino driving motors, pins A4 and A5 (SDA, SCL) were already used for directional control of the DC motors. Serial communication through the Tx/Rx pin was impossible because in the second Arduino these pins were used for communication between it and Raspberry Pi. Therefore, since there was only one free pin in the first Arduino, it was chosen to have this pin with PWM and a completely different approach was implemented, Fig. 8. Pin 3, from the first Arduino is connected to the second Arduino's A7 analogue pin. Since pin 3 is with PWM, this can output voltage in the 0-5V range, this voltage is now read by the analogue input of the second Arduino. By changing the voltage levels, various variables can be controlled. Four levels were needed, one for the front and back lights, the braking lights, and one for the buzzer. But if the pins are directly connected, this method will not work because the different voltage levels generated at pin 3 output are simply pulses with 0 and 5V values, and these will be detected by the other ADC as signals with the value 0 and 1023 (10 bits ADC) and not in-between values.

The way this problem was solved is by applying a RC filter shown in Fig. 9. Resistor and capacitor used values are: R=4.7kΩ and C=100nF on Arduino, Fig. 10.

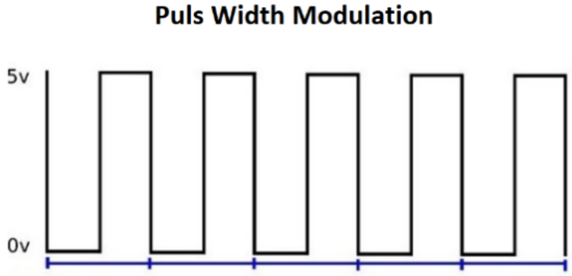


Fig. 8. Method of the PWM Function.

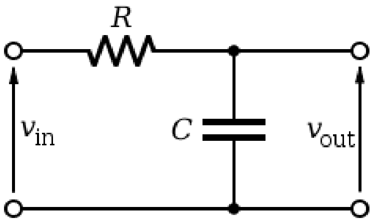


Fig. 9. Low-Pass RC Filter.

Fig. 10. RC Filter on Arduino.

From the figure it can be noticed that the signal initially comes to A1 because it was initially designed like that in circuit but later changes to A7. This filter "flattens" the PWM pulses and the output now is a constant voltage dependent on specified value of the PWM. Also, for greater precision a digital filter in the code was implemented. This filter work by taking the average of 10 consecutive measurements with a delay of 12 ms between each measurement.

B. Arduino-Raspberry Pi Connection

As noted earlier through the paper, Arduino reads the sensors and passes the data to Raspberry Pi via serial communication with the Tx/Rx pins. But connecting these two directly would damage Raspberry Pi as the latter works with 3.3V while Arduino with 5V. This problem has been solved by using a voltage divider to reduce the voltage level of Arduino from 5V to 3.3V. The pin connection method is Tx (Arduino) → Rx (Raspberry Pi), Rx (Arduino) →Tx (Raspberry Pi), Gnd →Gnd.

In Fig. 11 is shown the voltage divider scheme. The resistor value R1 is chosen arbitrarily, while the value of the resistor R2 is calculated from the following equation (1), given that the other variables are: Vin = 5V, Vout = 3.3V, R1 = 1.2 kΩ.

$$Vout = Vin * \frac{R1}{R1+R2} \qquad (1)$$

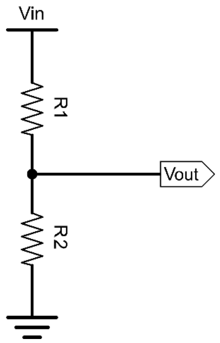


Fig. 11. Voltage Divider Scheme.

The Rx (Arduino) → Tx (Raspberry Pi) connection does not need a voltage divider, since the signal received at the Rx of Arduino is 3.3V, therefore within the working range of Arduino. Therefore, just one voltage divider is required.

C. Connection Arduino-Raspberry Pi

The Raspberry Pi-Laptop communication is implemented via a TCP server and socket. The Python library that enables this communication is called socket server. More detailed information on using this library and its implementation is available on the link in the references.

Both the Raspberry Pi and the laptop must be connected into the same Wi-Fi router to communicate. If possible, only these two without other devices for faster communication must be connected.

After obtaining devices IP addresses, these IPs are then written in Raspberry Pi as well as on the laptop:

```
h, p1, p2 = "192.168.0.100", 8000, 8002  # IP and ports h-->
host.
```

Through the above line, host, two ports are defined. It should be noted that the IP can vary depending on the devices. To start communication, the program on the laptop should be initiated first, to start the server, and then in the clients (programs in Raspberry Pi). Hence the laptop is the server in this case, while the client is Raspberry Pi. The client sends data to the server.

D. Communication Laptop-Arduino

Data processed by the laptop through the serial port (USB) are sent to Arduino in the transmitter box and then conveyed to the vehicle through radio waves. In order for Python to send this data to serial port, the pyserial library is utilized:

```
import serial as s

ser = s.Serial('COM6', 115200).
```

The above lines enable communication, where 'COM6' is the port Arduino is connected, while '115200' represents baud rate.

V. Training of Models for Detection of Traffic Signs

Among other objectives of this paper is the detection of traffic signs. It is known there are many traffic signs, but for this project only the STOP sign is selected.

The selected detection mode is based on Haar features.

In order to detect a particular object, a cascade classifier should firstly be trained [6, 7]. This is done by taking positive, Fig. 12, and negative images, Fig. 13. Positive images are those images that contain the object that has to be detected, while negative images may be anything as long as they don’t contain target detection object. As negative images are used the track, the road, i.e. mainly behind the scenes from moving the vehicle [9, 10], but without the traffic signs in those images. The greater the number of positive and negative images is, the longer training will take, but the classifier will be more accurate.

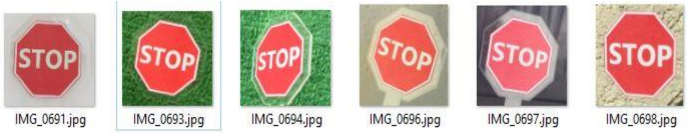


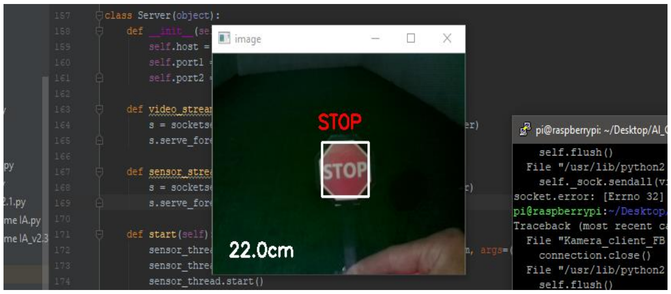
Fig. 12. Examples of Positive Images.

Fig. 13. Examples of Negative Images.

There are many ways to capture a large number of images, ranging from downloading from the internet to individual object photography; the method that is preferred in this paper is to capture a video with the target object for detection and then extracting images from the video frames. This speeds up the process of capturing images.

After capturing positive images, they should be cut in dimensions as close as possible to each other and should only contain the target for detection. After this step, comes the classifier training, whereby the trained classifier is stored as a xml file.

The trained classifier is then stored as *stop.xml* and used for detection. Below is given an example showing STOP sign begin detected as shown in Fig. 14.



```
class Server(object):
    def __init__(se
        self.host =
        self.port1 =
        self.port2 =

    def video_strea
        s = socketse                                                er)
        s.serve_fore

    def sensor_strea
        s = socketse                                                r)
        s.serve_fore

    def start(self):
        sensor_threa                                                n, args=(
        sensor_threa
        sensor_thread.start()
```

```
    self.flush()
  File "/usr/lib/python2
    self._sock.sendall(v
socket.error: [Errno 32]
pi@raspberrypi:~/Desktop/
Traceback (most recent ca
  File "Kamera_client_FB
    connection.close()
  File "/usr/lib/python2
    self.flush()
```

Fig. 14. Stop Sign being Detected after Training the Classifier.

VI. Programming and Control with Artificial Intelligence

This is the most challenging part of this paper. The method used for Artificial Intelligence training is CNN. CNNs are layers of perceptrons. The first being an input layer, then many hidden layers and an output layer. For example, with an image, each perceptron in the input layer would correspond to a pixel value. Then the next layer receives input from every perceptron in the previous layer, and if it passes, the value then it will be sent into the next hidden layer until the output layer is reached. The dataset was tested on different positions of the STOP traffic sign and calculated the accuracy and time required to identify those signs.

For this research paper, firstly you should drive the vehicle manually, where for each key pressed for driving the vehicle, the corresponding frame and the direction of movement are saved.

The frame obtained from the video is initially converted from RGB (colour format) to grayscale. In grayscale image formats, each of the individual pixels has a certain value that represents the brightness of that pixel.

Grayscale image is a matrix where each element of the matrix represents a pixel, and each pixel has a certain value that indicates how bright or dark that pixel is. This file, from a two-dimensional matrix, becomes a one-dimensional array of information for each pixel and is stored as a npz file.

In the input layer of the CNN as input is given the previously stored npz file (the string of pixels), while the output layer is feed with information about the button pressed for that frame.

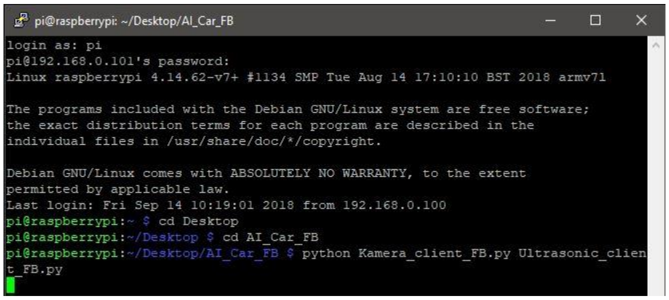
The computer then by comparing human input with the current frame calculates backwards weights of each joint separately to match the input. So, in a way, the output responds to the input while calculating the weights of the joints, so this method is called backpropagation, and that's how neural networks they learn in this case. Repeating the process of calculating the weights for each input frame, and each current input by the human for all training data, the computer is able to generate a CNN model which for any given frame responds with an output which in this case is: Front, right, left or back.

After completing weights calculation for each frame, the neural network model is stored as *.xml, then this file is used to predict the direction of car's movement.

A. Starting the Video Transmission and Sensor Data

Since the CNN model for vehicle motion prediction is located on a laptop, the camera and sensor data must be transmitted from Raspberry Pi to the laptop. In Raspberry Pi, there are two Python scripts that enable this stream, but firstly must be executed, Fig. 15. This can be done in several ways including the Raspberry Pi connection with HDMI, but the preferable method is by accessing the RPi via SSh (Secure Shell). The way the scripts are executed in Raspberry Pi directly from the laptop is as follows: Initially we connect to Raspberry Pi through hostname and password. Then, we need to know where the Python scripts are located. After navigating to the path containing the Python scripts, we run them. Here follows the procedure of navigating and running the scripts.

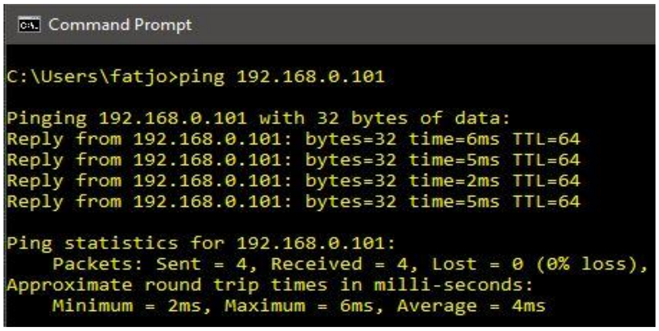
During this process, both the laptop and Raspberry Pi must be connected to the same network. The video transmitted by Raspberry Pi has been reduced to 240x320px for faster transmission, this way having less lag and transmission delays.



pi@raspberrypi: ~/Desktop/AI_Car_FB

```
login as: pi
pi@192.168.0.101's password:
Linux raspberrypi 4.14.62-v7+ #1134 SMP Tue Aug 14 17:10:10 BST 2018 armv7l

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Fri Sep 14 10:19:01 2018 from 192.168.0.100
pi@raspberrypi:~ $ cd Desktop
pi@raspberrypi:~/Desktop $ cd AI_Car_FB
pi@raspberrypi:~/Desktop/AI_Car_FB $ python Kamera_client_FB.py Ultrasonic_clien
t_FB.py
```

Fig. 15. Python Scripts Execution through Putty.



Command Prompt

```
C:\Users\fatjo>ping 192.168.0.101

Pinging 192.168.0.101 with 32 bytes of data:
Reply from 192.168.0.101: bytes=32 time=6ms TTL=64
Reply from 192.168.0.101: bytes=32 time=5ms TTL=64
Reply from 192.168.0.101: bytes=32 time=2ms TTL=64
Reply from 192.168.0.101: bytes=32 time=5ms TTL=64

Ping statistics for 192.168.0.101:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
Approximate round trip times in milli-seconds:
    Minimum = 2ms, Maximum = 6ms, Average = 4ms
```

Fig. 16. Pinging Raspberry Pi.

If one ping Raspberry Pi from the laptop, one can get the time it takes for Laptop - Raspberry Pi communication, Fig. 16. In this case an average round trip of 4 ms was achieved.

Further, the evaluation of learning performance for the STOP traffic sign recognition with the different slant angles and occlusion ratios based on CNN has been done.

Fig. 17 and Fig. 18 show the learning performance with slang angles normal and parallel to sign, respectively. Fig. 19 shows the learning performance with occlusion ratios.

B. Manual Drive for Data Collection

The vehicle is capable of moving in eight different directions: forward, backwards, forward-right, forward-left, backwards-right, backwards-left, and also spin in place clockwise and counter clockwise.

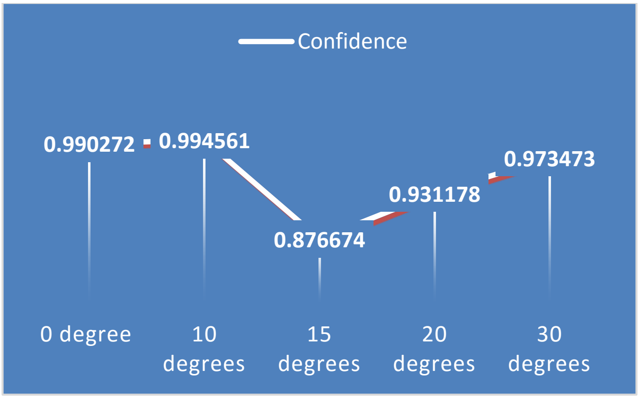


Fig. 17. Performance with Slang Angles Normal to Sign.

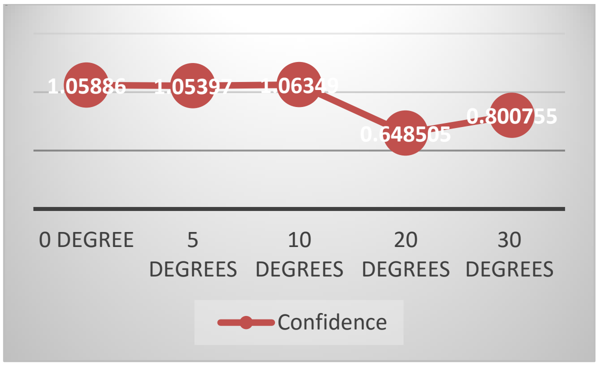


Fig. 18. Performance with Slang Angles Parallel to Sign.

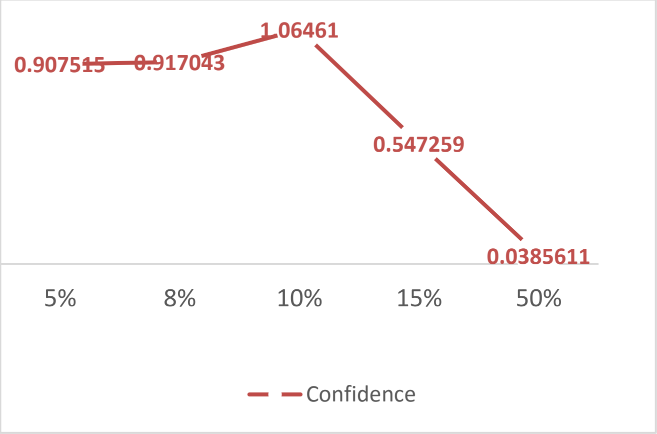
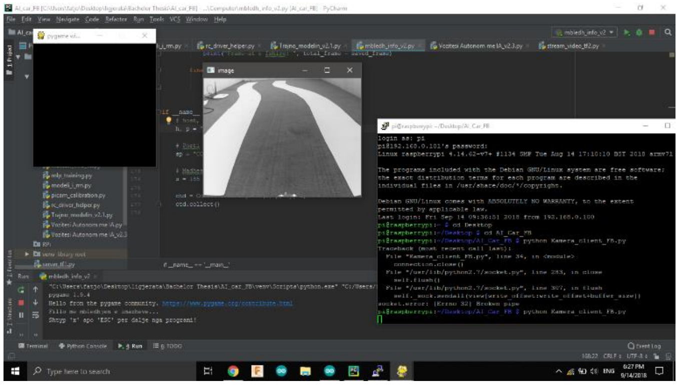


Fig. 19. Performance with Occlusion Ratios.



```
"C:\Users\fatjo\Desktop\Ligjerata\Bachelor Thesis\AI_car_FB\venv\Scripts\python.exe" "C:/Users/
pygame 1.9.4
Hello from the pygame community. https://www.pygame.org/contribute.html
Fillo me mbledhjen e imazheve...
Shtyp 'x' apo 'ESC' per dalje nga programi!
```

```
login as: pi
pi@192.168.0.101's password:
Linux raspberrypi 4.14.62-v7+ #1134 SMP Tue Aug 14 17:10:10 BST 2018 armv7l

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Fri Sep 14 09:36:51 2018 from 192.168.0.100
pi@raspberrypi:~ $ cd Desktop
pi@raspberrypi:~/Desktop $ cd AI_Car_FB
pi@raspberrypi:~/Desktop/AI_Car_FB $ python Kamera_client_FB.py
Traceback (most recent call last):
  File "Kamera_client_FB.py", line 34, in <module>
    connection.close()
  File "/usr/lib/python2.7/socket.py", line 283, in close
    self.flush()
  File "/usr/lib/python2.7/socket.py", line 307, in flush
    self._sock.sendall(view[write_offset:write_offset+buffer_size])
socket.error: [Errno 32] Broken pipe
pi@raspberrypi:~/Desktop/AI_Car_FB $ python Kamera_client_FB.py
```

Fig. 20. Data Collection Process for CNN Training.

The first step in CNN training is manually driving the vehicle. In the Python code, the vehicle is driven by the W, A, S and D keys: W for forward, S for reverse movement, D-right, A - left, WD - right front, WA - forward left, SD-backwards right as well as SA - backwards left. During manual data collection, Python scripts for video transmission from Raspberry Pi on the laptop should be executed at the same time, so that for each key pressed, the laptop will capture the frame and the corresponding key. These frames, as mentioned earlier from a 2D matrix, are transformed into a string and stored as a *.npz file. Data collection process is very important because the accuracy of the CNN training and its prediction depends directly on the data collection process. In this case, over 440 Mbytes of data for CNN training were collected as in Fig. 20.

VII. Data Processing Method

Throughout the paper is mentioned that the video and data are transmitted from the RPi to the laptop. Knowing the fact that RPi is a single board computer, the question why the processing isn't done directly on Raspberry Pi instead of laptop might arise. This all comes down to processing power. Below is given a Table II showing key differences between these two. From the table, it's clear that laptop's processing power is far "superior" compared to Raspberry Pi's.

The way how the entire hardware "talks" to each other is shown in Fig. 21.

TABLE. II. Laptop–Raspberry Pi Comparison

	Processor model	**Processor clock speed**	**RAM**	**Internal memory**
Laptop	Intel core i3	@2.53 2.53 GHz	6 GB	240 SSD +750 GB HDD
Raspberry Pi 3B	ARM cortex A53	@1.2 GHz	1 GB	8 Gb SD card

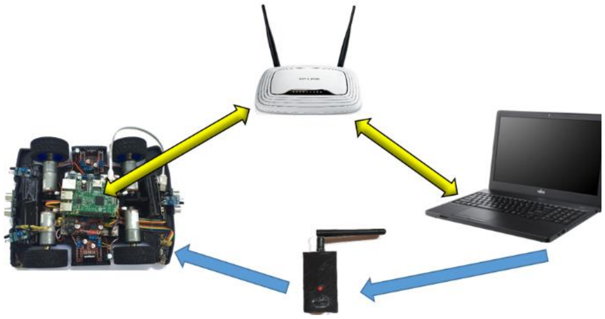


Fig. 21. Laptop–RPI–Laptop–Arduino Communication Routes.

VIII. Conclusions

For this research paper, mechanics was applied for designing the vehicle and mounting of the entire hardware, electronics was used to develop a circuit board, and the hardware was synchronized with the electrical circuits through the programming language. The fusion of all of these has resulted into a mechatronic project.

In conclusion, a vehicle capable of moving in all the necessary directions (forward, forward right, forward left, backwards, backwards right, backwards left as well as turn and rotate in place) was "born".

The vehicle has been equipped with sensors for gathering necessary information about the surrounding environment as well as camera to collect the pictures of the traffic signs.

Based on the CNN algorithm one can conclude that the influence due to occlusion is much larger than that of slant angle. Therefore, it can identify the stop sign even if the sign is inclined 30 degrees and with 15% occlusion.

The importance of this paper can be seen in the fact that some steps have already been made to navigate the autonomous car model that can be compared with the autonomous machines of the elite companies in this field.

There is also room for improvements that will be made in future models of this vehicle using other learning methods for traffic sing recognition.

References

[1] Pannu G. S., Ansari, M. D. & Gupta P. Design and Implementation of Autonomous Cars Using Raspberry Pi, International Journal of Computer Applications (0975 – 8887), pp.222-222, Volume 113, No.9, March 2015.

[2] M.Karthikeyan, Mr. G.Sreeram, M.Tech, (Ph.D). Intelligent Exploration and Surveillance Robot In Defense Environment. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering. Vol. 3, Special Issue 1, February 2014.

[3] G. N. Tripathi and V.Rihani. Motion Planning of an Autonomous Mobile Robot using Artificial Neural Network. Mody Institute of Technology and Science, Lakshamangarh, Sikar, Rajasthan.

[4] Rakesh Chandra Kumar et al. Obstacle Avoiding Robot – A promising One. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 2, Issue 4, April 2013.

[5] Beqa, F. Development of autonomous vehicle driven by Artificial Intelligence, University of Prishtina, 2018.

[6] Pajaziti, A., & Bajrami Xh. & Paliqi A. Path Control of Quadruped Robot through Convolutional Neural Networks, 18th IFAC Conference on Technology, Culture and International Stability, Sept 13-15, 2018, Baku, Azerbaidschan, 2018.

[7] Bajrami, X., Gashi, B., & Murturi, I. (2018). Face recognition performance using linear discriminant analysis and deep neural networks. International Journal of Applied Pattern Recognition, 5(3), 240-250.

[8] Hwu, T., Isbell, J., Oros, N., & Krichmar, J. (2017, May). A self-driving robot using deep convolutional neural networks on neuromorphic hardware. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 635-641). IEEE.

[9] Ishibushi, S., Taniguchi, A., Takano, T., Hagiwara, Y., & Taniguchi, T. (2015, November). Statistical localization exploiting convolutional neural network for an autonomous vehicle. In IECON 2015-41st Annual Conference of the IEEE Industrial Electronics Society (pp. 001369-001375). IEEE.

[10] Tai, L., Li, S., & Liu, M. (2017). Autonomous exploration of mobile robots through deep neural networks. International Journal of Advanced Robotic Systems, 14(4), 1729881417703571.

Hybrid Latin-Hyper-Cube-Hill-Climbing Method for Optimizing: Experimental Testing

Calista Elysia[1], Michelle Hartanto[2], Ditdit Nugeraha Utama[3]
Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University, Jakarta, Indonesia 11480

***Abstract*—A noticeable objective of this work is to experiment and test an optimization problem through comparing hill-climbing method with a hybrid method combining hill-climbing and Latin-hyper-cube. These two methods are going to be tested operating the same data-set in order to get the comparison result for both methods. The result shows that the hybrid model has a better performance than hill-climbing. Based on the number of global optimum value occurrence, the hybrid model outperformed 7.6% better than hill-climbing, and produced more stable average global optimum value. However, the model has a little longer running time due to a genuine characteristic of the model itself.**

Keywords*—*Hill-Climbing; Latin-Hyper-Cube; Optimization

I. Introduction

Optimization, also known as mathematical programming, is a process or action to obtain the highest achievable performance under the given constraints. The final goal of all optimization is to minimize the effort required or to maximize the desired benefit. Several methods for optimization have been developed are hill-climbing (HC) [1], simulated-annealing (SA) [2], generic algorithm [3], particle swarm optimization [4], ant colony optimization [5], and others.

Many researchers have performed various studies in the scope of optimization. Author in [6] conducted a study to improve the sampling algorithm in the Bayesian network by applying the Latin-hyper-cube (LHC) sample. This study showed a better result compared to applying simple random sampling. Author in [7] discussed the Quasi-Newton methods which are an iterative optimization method. There is also a study that used the hill-climbing optimization method which was found to effectively detect grids on microarray images taken from databases from GEO and Stanford genomic laboratories [8]. Also, [9] constructed an optimization model using simulated-annealing and hill-climbing then compared them. The result said that hill-climbing consumed the shortest running time. Author in [10] applied genetic algorithm to optimize the control (process) parameters. Finally, [11] explained the success of simulated-annealing method for global optimization problems by studying the ideal version of the algorithm.

The objective of this study is to scientifically experiment an optimization problem by comparing HC method with a hybrid method combining LHC and HC (L2HC). While HC is going to use a random starting point in any position, the hybrid method is divided starting point into several identified clusters. This two methods are going to be tested using the same data set in order to get the comparison result for both methods. Here, the proposed hybrid method conjoining two methods LHC and HC is a novel-configuration in optimizing. It is an optimization method by dividing search-area via several definitive-clusters, and conducting searching alternatives in each cluster.

The organization of this paper is as follows. Section II presents a literature overview about optimization, HC, and LHC. The related works and experimental method is going to be delivered in Section III and Section IV, respectively. In Section V, we present experimental result and analysis. Finally, Section VI concludes the paper.

II. Literature Overview

A. Optimization

Optimization derived from Latin words ‘optimus’ which has the ‘best’ meaning, it can be interpreted as a process of finding conditions that provide optimal value from an objective function [12]. The optimal value can be a minimum value or a maximum value of the objective function in accordance with the existing problems (Fig 1).

The combination method L2HC will be discussed in this experimental paper. The HC method itself is included in the method of mathematical programming or modern (non-traditional) optimization techniques that are very useful for finding the optimal value of an objective function of several variables that are in certain constraints. They are heuristic methods and are often faster than exact methods, especially in the case of non-linear objective functions with many variables. While the LHC method is included in the statistical sampling method which is a method for analyzing experimental data and developing empirical models to get the most accurate representation of the physical situation [13].

B. Hill Climbing

HC is a method that aims to find local maximum or minimum values through simple iterations that continue to move towards increasing values if looking for maximum values or decreasing values if looking for minimum values from an objective function to finding the nearest peak value or closest valley point [12]. This method is very simple and has been successfully applied to various optimization problems. The success of this method is due to the fact that choosing a heuristic that more accurately predicts the actual solution that produces more opportunities to obtain the optimal solution [1][8][14][15]. Pseudocode of HC method is displayed in Code 1. Searching for optimum value with HC method is greedy and the search starts from a random starting point. This causes the optimum results can be the local optimum, except for the random value of the luckiest starting point [16].

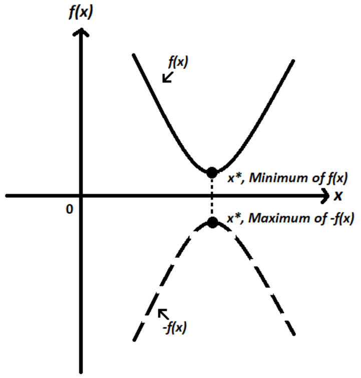


Fig. 1. Illustration of Optimization for Maximum and Minimum Value [13].

There are several parameters needed for the HC method, some of them are mandatory; such as, $cVal$, $bestNeighbor$, $bestVal$, which respectively represents the current value, neighbor's best value, and best value (local or global optimum). Based on the pseudocode, it can be comprehended that the iteration process is going to be terminated when the local optimum value has been obtained. That is when there is no value from the $bestNeighbor$ variable that is better than the value of the $cVal$ variable.

Code 1. Pseudocode of Hill Climbing Method [12]

```
Procedure HillClimbing()
Begin
   <...variables definition...>
   //randomizing new parameter combination
   cPos <-- random()
   cVal <-- objFunction(cPos)
   //looping until local optimum is found
   While(search is not terminated)
      //finding the best neighbor
      bestNeighbor <-- getBestNeighbor()
      //finding local optimum
      If(cVal>=bestNeighbor)
         Local optimum is found
         Search is terminated
      Else //move to next position
         cVal <-- bestNeighbor
         cPos <-- bestNPos
      End if
   End while
```

C. Latin-Hyper-Cube

LHC was proposed by [17], which is a method for selecting samples from populations. LHC has been proven to be able to reduce variance [18] and produce better analysis results compared to simple random sampling method [19]. In this method, how many intervals of equal probability so input variable value can be divided into several cubes according to the intervals are firstly determined. Each parameter combination is randomly selected and is determined how many combinations of parameters can be in the same group. After getting a combination of parameters randomly, check which group of parameters the combination is. If it exceeds the maximum number of parameter combinations in that group, the value must be randomized again [12].

Code 2 shows a pseudocode for the LHC method when implemented as an optimization method. Here, randomizing a position from the population is required, and the clusters checked then by using a function named $checkCluster()$. It will return a Boolean value and refers to equal status ($eqStatus$) variable, returned 'true' if it exceeds the maximum number in that group, and 'false' if it able. This random process will stop when $eqStatus$ is 'false'.

Code 2. Pseudocode of Latin-Hyper-Cube Method [12]

```
Procedure LatinHypercubeSampling()
Begin
   <...variables definition...>
   <...parameters cluster making...>
   bestVal <-- 0
   While(loopCount)
      //randomizing new parameter combination and
      //checking the equality
      eqStatus <-- true
      While(eqStatus=true)
         //randomizing new parameter
         cPos <-- random()
         //checking cluster equality status
         eqStatus<--checkCluster(cPos)
      End while
      cVal <-- objFunction(cPos)
      //checking the best value
      If(cVal > bestVal)
         bestVal <-- cVal
         bestPos <-- cPos
      End if
      loopCount++
   End while
End
```

III. Related Works

Numerous studies exploring about LHC or HC method have been already conducted by many researchers. Author in [20] adopted LHC sampling to design a sophisticated gas turbine. This sampling method is applied recursively to identify the most important input parameters. Author in [21] proposed a new method named 'IDLHCSA' for history matching to get reliable forecast. 'IDLHCSA' combines iterative discrete Latin-hyper-cube (IDLHC) to find good matched models with SA method.

Furthermore, [22] examined the performance of HC method for mesh router node placement in wireless mesh network. The result shows that connectivity and user coverage are achieved well. Author in [23] applied smart HC using the ideas of LHC sampling to find an optimal configuration for web application server. This proposed method can learn from previous searches and more efficient than traditional heuristic methods. Also, [24] operated HC method to prove about statement of 'optimization has a better solution when it closer to the local optimum value' is wrong. Besides the local optimum value, number of steps to reach the local optimum also needs to be considered.

Particular in hybrid model, [25], in discussing a communication behavior for social dynamical systems, studied a hybrid opinion network containing of continuous valued and discrete valued agents. The agents talked about were copier, voter, and averager agents. Here, the communication topologies were modeled. The study concluded that the voters' existence has dissimilar impact on the evolution and consensus value of negotiation process.

Additionally, [26] also investigated resilient consensus problem in hybrid multi-agent system. The hybrid multi agent system itself consists of continuous time and discrete time dynamical agents. Author in [26] successfully constructed a hybrid censoring strategy to reach resilient consensus. The consensus here defined as compromise between cooperative agents and Byzantine agents (as uncooperative agents).

IV. Research Methodology

This experiment involved five stages (Fig 2); i.e. preliminary study, construct model, data preparation, experiment, and evaluation. In the fisrt stage, optimization, HC, and LHC methods were learned deeply. Then, the model was constructed using class diagram. For the next two stages, python 3.6 is functioned methodically for generating the data set and test the optimization method. We performed this experiment by using 7th generation of Intel core i7 processor, 8GB RAM, and operating system Windows 10 64-bit as the hardware and software specifications. With the help of interpolate.lagrange from scipy library and linspace from numpy, data are generated (Fig. 3) and the equation is provided in □1). It is going to produce interpolated values randomly. The last stage is to perform an evaluation by analyzing the experiment result.

In addition, while doing the experiment stage, each method randomizes 100 starting point for one process. It means that HC will do 100 iterations while hybrid L2HC will do 20 iterations with 5 starting point for each iteration. This process is done 10 times in this work so the total iteration obtained is 1000 iterations.

$$f(x) = -6{,}317\mathrm{e} - 44x^{13} - 2{,}611\mathrm{e} - 39x^{12} + 2{,}727\mathrm{e} - 34x^{11} - 7{,}965\mathrm{e} - 30x^{10} + 1{,}241\mathrm{e} - 25x^{9} - 1{,}196\mathrm{e} - 21x^{8} + 7{,}563\mathrm{e} - 18x^{7} - 3{,}197\mathrm{e} - 14x^{6} + 8{,}987\mathrm{e} - 11x^{5} - 1{,}631\mathrm{e} - 7x^{4} + 0{,}000179x^{3} - 0{,}1038x^{2} + 24{,}48x + 923 \quad (1)$$

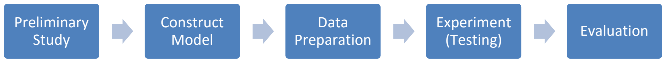


Fig. 2. Experiment Stages.

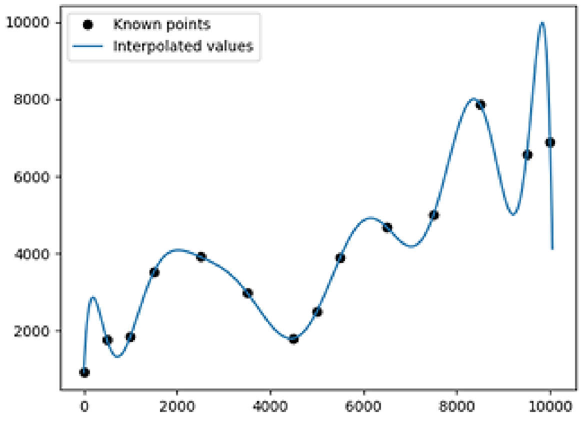


Fig. 3. Dataset Operated in Experiment.

V. Result and Analysis

Schematically, the constructed model is configured by Fig 4. It consists of three classes, where class L2HC (the class symbolizes the proposed hybrid model) is consisting two classes LHC and HC. All attributes defined in a main class, where they practically belong to and able to be equipped by both classes LHC and HC.

The pseudocode of L2HC is provided in Code 3. HC has seven parameters that are initiated and defined (section B). The first is *currentPosition* which defines the initial position that will be used in the optimization search and *currentValue* is the value of that *currentPosition*.

Then, bestNeighborPosition defines the position of the best neighbor around currentPosition and bestNeighborValue is the value of that bestNeighborPosition. LocalOptimum is found when there is no neighbor that have better value than current value, and terminatedLoop is a Boolean which indicates when the search must stop. There are also operations that are used here, they are definingOF() where the objective function for optimization is defined, randomizingPosition() for initial positioning, findingBestNeighbor() to find the best neighbour value, movingToNext(), and findingBest() to find the best value.

The idea of hybrid L2HC method is applying LHC method to HC method. This method is done by dividing the data into several clusters using clustering operation and running a HC method in each cluster. It must be ensured that the process is run only once on each cluster, this is checked by the *checkCluster*() function. Can be concluded here that the other parameters used in L2HC model are *clusterNumber*, *loopCount*, and *eqStatus*.

The challenging work in developing the proposed model is to merge the part of clustering algorithm with the part of HC searching. At this point, the searching part was inserted in the clustering looping fragment (see Code 3). It technically affects that the proposed model operates two types of looping block for both checking the clusters and searching the optimal value.

Practically, it was dissimilar with [25] and [26] for catching consensus rate, here we constructed a hybrid model via combining two types of methods for obtaining new optimization value for heuristic optimization problem.

After running 10 times process, the number of occurrence of HC method to get the global optimum value was 73 and hybrid L2HC got 85 times global optimum value from 1,000 iterations. From the obtained global optimum, 53.8% is produced by the hybrid method, outperforming HC method around 7.6%. Details of number occurrence for each process can be seen in Fig 5.

The average of local optimum obtained from each iteration is calculated and presented in Fig 6. We can see 8 out of 10 processes show hybrid L2HC method is better than HC. Also, the average obtained by hybrid L2HC method is much more stable. In other hands, HC produces an up and down average value. This can be happened because LHC will divide population into several clusters and randomize starting point for each cluster. So the starting point for each iteration in the process more or less will be in the similar position. Therefore, it will lead to a similar local optimum value.

Code 3. Pseudocode of the Hybrid L2HC Method

```
Procedure LHS_HC()
Begin
   initialize iteration
   initialize numCluster
   while(it<iteration)
      <...variables definition...>
      <...parameters cluster making...>
      While(loopCount<numCluster)
         eqStatus <-- true
         While(eqStatus = true)
            cPos <-- random()
            eqStatus <-- checkCluster(cPos)
         End While
         search <-- true
         cVal <-- objFunction(cPos)
         While(search is not terminated)
            bestNeighbor <--
getBestNeigh(cPos,
            totalData)
            If(cVal>=bestNeighbor)
               opt = cVal
               //Local optimum is found
               Search is terminated
            Else //move to next position
               cVal <-- bestNeighbor
               cPos <-- bestNPos
            End if
         End While
         loopCount++
      End while
      it++
   End while
```

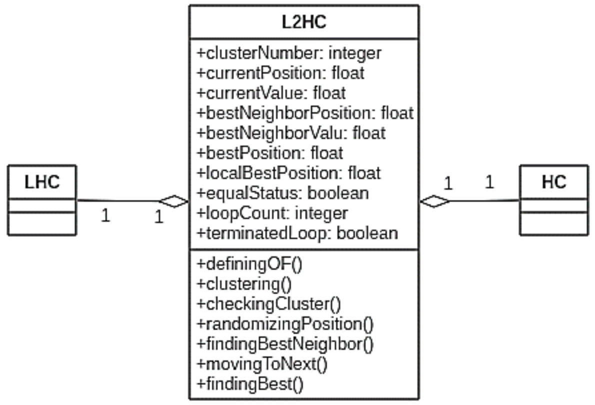


Fig. 4. Constructed Model.

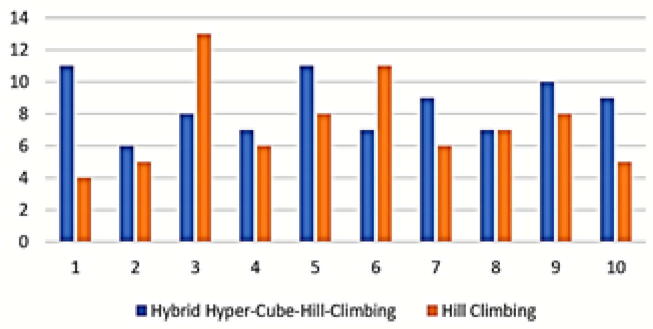


Fig. 5. Comparison of Global Optimum Occurrence.

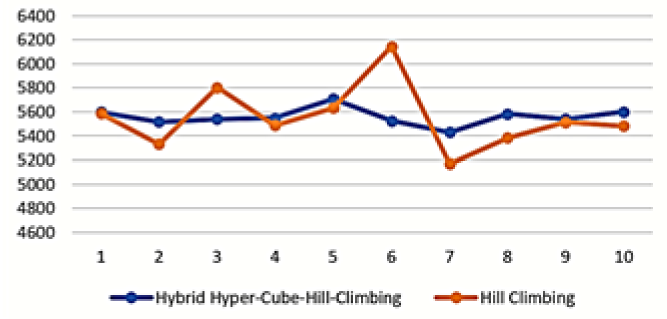


Fig. 6. Average Comparison of Optimum Value.

TABLE. I. AVERAGE OF RUNNING TIME FOR EACH PROCESS

	Hill Climbing	*Hyper-Cube-Hill-Climbing*
Running Time	0,001718695	0,002031107
	0,001718657	0,001718674
	0,002031164	0,002968359
	0,002031169	0,002031209
	0,001562223	0,002187619
	0,001562517	0,002187636
	0,002187026	0,001874907
	0,00203126	0,002030938
	0,001718521	0,001718628
	0,002031105	0,001562419
Average	**0,001859234**	**0,002031150**

Furthermore, average of running time is also calculated and displayed in Table I. Here, we got a fact about the average running time of 10 times process with hybrid L2HC method is more worse than HC. Hybrid L2HC need more time to randomizing the starting point. It will always randomize if starting point in one iteration has the same cluster.

VI. Conclusion

Experiments using two optimization model, namely, HC and hybrid L2HC, were done in this work. The conclusion revealed that hybrid L2HC has a better performance than HC itself. Based on the number of global optimum value occurrence, hybrid L2HC model outperformed 7.6% better than HC, and from the average comparison of the global optimum value, hybrid L2HC model is more stable and has a greater number. However, hybrid L2HC has a little longer running time. This happened because when randomizing the starting position, it is going to check whether the cluster where the position is located has been initialized or not, while hill climbing does not do the cluster checking first.

The future study for this work is regarding the effectiveness of the model. Although both models are good for optimization problems, it is found in this work that the effectiveness of both HC and hybrid L2HC are still need to be improved. Their effectiveness respectively are only 7.3% and 8.5%. Considering this value, hopefully further research can improve the model.

ACKNOWLEDGMENT

We would like to thank BINUS University who has supported our works, particularly BINUS graduate program, Master of Computer Science.

REFERENCES

[1] P. R. Norvig and S. A. Intelligence, A modern approach. Prentice Hall, 2002.

[2] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” J. Chem. Phys., vol. 21, no. 6, pp. 1087–1092, 1953.

[3] J. H. Holland, “Genetic algorithms,” Sci. Am., vol. 267, no. 1, pp. 66–73, 1992.

[4] R. Eberhart and J. Kennedy, “A new optimizer using particle swarm theory,” in MHS’95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 1995, pp. 39–43.

[5] A. Colorni, M. Dorigo, and V. Maniezzo, “Distributed optimization by ant colonies,” in Proceedings of the first European conference on artificial life, 1992, vol. 142, pp. 134–142.

[6] J. Cheng and M. J. Druzdzel, “Latin hypercube sampling in Bayesian networks.,” in FLAIRS Conference, 2000, pp. 287–292.

[7] K. Bryan, “Quasi-Newton Methods,” Rose-Hulman Inst. Technol., 2004.

[8] L. Rueda and V. Vidyadharan, “A hill-climbing approach for automatic gridding of cDNA microarray images,” IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 3, no. 1, p. 72, 2006.

[9] D. N. Utama, N. Ani, and M. M. Iqbal, “An optimal generic model for multi-parameters and big data optimizing: A laboratory experimental study,” in Journal of Physics: Conference Series, 2018, vol. 978, no. 1, p. 12045.

[10] R. Malhotra, N. Singh, and Y. Singh, “Genetic algorithms: Concepts, design for optimization of process controllers,” Comput. Inf. Sci., vol. 4, no. 2, p. 39, 2011.

[11] H. E. Romeijn and R. L. Smith, “Simulated annealing and adaptive search in global optimization,” Probab. Eng. Informational Sci., vol. 8, no. 4, pp. 571–590, 1994.

[12] D. N. Utama, “The Optimization of the 3-d Structure of Plants, Using Functional-Structural Plant Models. Case Study of Rice(Oryza sativa L.) in Indonesia,” Georg-August University, 2015.

[13] S. S. Rao, Engineering optimization: theory and practice. John Wiley & Sons, 2009.

[14] A. Cawsey, The essence of artificial intelligence. Prentice Hall PTR, 1997.

[15] K. A. Sullivan and S. H. Jacobson, “A convergence analysis of generalized hill climbing algorithms,” IEEE Trans. Automat. Contr., vol. 46, no. 8, pp. 1288–1293, 2001.

[16] J. Boyan and A. W. Moore, “Learning evaluation functions to improve optimization by local search,” J. Mach. Learn. Res., vol. 1, no. Nov, pp. 77–112, 2000.

[17] M. D. McKay, R. J. Beckman, and W. J. Conover, “Comparison of three methods for selecting values of input variables in the analysis of output from a computer code,” Technometrics, vol. 21, no. 2, pp. 239–245, 1979.

[18] M. Stein, “Large sample properties of simulations using latin hypercube sampling,” Technometrics, vol. 29, no. 2, pp. 143–151, 1987.

[19] J. C. Helton and F. J. Davis, “Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems,” Reliab. Eng. Syst. Saf., vol. 81, no. 1, pp. 23–69, 2003.

[20] A. M. Briones, D. L. Burrus, J. P. Sykes, B. A. Rankin, and A. W. Caswell, “Automated Design Optimization of a Small-Scale High-Swirl Cavity-Stabilized Combustor,” J. Eng. Gas Turbines Power, vol. 140, no. 12, p. 121509, 2018.

[21] C. Maschio and D. J. Schiozer, “A new methodology for history matching combining iterative discrete Latin Hypercube with multi-start simulated annealing,” J. Pet. Sci. Eng., vol. 169, pp. 560–577, 2018.

[22] A. Xhafa, E. Spaho, D. Elmazi, and M. Takizawa, “A Study on Performance of Hill Climbing for Router Placement in Wireless Mesh Networks,” in 2015 10th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA), 2015, pp. 460–465.

[23] B. Xi, Z. Liu, M. Raghavachari, C. H. Xia, and L. Zhang, “A smart hill-climbing algorithm for application server configuration,” in Proceedings of the 13th international conference on World Wide Web, 2004, pp. 287–296.

[24] L. Hernando, A. Mendiburu, and J. A. Lozano, “Hill-Climbing Algorithm: Let’s Go for a Walk Before Finding the Optimum,” in 2018 IEEE Congress on Evolutionary Computation (CEC), 2018, pp. 1–7.

[25] Y. Shang, "Hybrid consensus for averager-copier-voter networks with non-rational agents," Chaos, Solitons & Fractals, vol. 110, pp. 244-251, 2018.

[26] Y. Shang, "Consensus of hybrid multi-agent systems with malicious nodes," IEEE Transactions on Circuits and Systems II: Express Briefs, pp. 1-1, 2019.

Evaluation of Usability Dimensions of Smartphone Applications

Shabana Shareef[1], M.N.A. Khan[2]
Department of Computer Science
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology
Islamabad, Pakistan

***Abstract*—This study analyses different techniques used for evaluation of various usability dimensions of software applications (apps) being used on the smartphones. The scope of this study is to evaluate various aspects of the usability techniques employed in the domain of smartphone apps. Usability assessment methodologies are evaluated for different types of applications running on different operating systems like Android, Blackberry and iOS, etc. Usability evaluation techniques and methodologies with respect to usability heuristics like field experiments, laboratory experiments models and usability standards are discussed in detail. The issues for evaluation of usability of smartphone apps are identified by considering limitations and areas of improvement outlined in the contemporary literature. A conceptual framework for usability evaluation of smartphone apps is also designed which would be validated through experimentation in the thesis work. This study is particularly useful to comprehend usability issues and their likely remedies to produce high quality smartphone apps.**

Keywords—Usability; Jakob-Nielson usability heuristics; smartphone applications; ease of use; understandability; learning curve

I. Introduction

All Smartphones have become a daily use items and are popular among all the sections of society. Smartphones are now the most popular mobile technology. The statistical data show that one in each three citizens own a smartphones [1]. Smartphones are in fact modern mobile phones with an additional highly developed computing capability and connectivity. Their extensive modes for input are provided by a touch sensitive display. The most popular smartphones operating systems are Android, iOS, Windows phones and Blackberry. The quick and growing amount of smartphone apps on the Google play and Apple stores have facilitated and impelled the software experts to design applications of better quality in order to compete in the markets. There are numerous measureable aspects on the quality of software product and usability is one of the most significant aspects [1].

In the recent years, the introduction of numerous technologies has revolutionized our mode of communication, entertainment and completing daily routine tasks. Simultaneously, the method of digital convergence has resulted in the inventions of several devices like PDAs, smartphones, tablets etc. which are able to gather different forms of human-computer interaction (HCI) in an integrated way. The HCI research community recommends considering different requirements when evaluating those applications, such as quantitative data (metrics), subjective evaluation (users' impressions) and context data (e.g. environment and devices' conditions.).

Usability is known as a significant quality dimension to evaluate the quality and usefulness of smartphone applications [2]. Usability is a quality attribute that assesses how easy and simple user interfaces are to use. According to Nielson, usability can be defined as a method for improving the design process. Usability is assessed on the basis of six dimensions which include learnability, memorability, efficiency, effectiveness, error rate and user satisfaction. Similarly, IEEE Standard.610.12-1990 defines usability as “The ease with which a user can learn to operate, prepares inputs for, and interprets outputs of a system or component.” Usability dimensions affect four contextual factors which are users, technology, activity and environment. There are two classifications of usability each having different number of parameters. First classification includes effectiveness, efficiency and user satisfaction which are part of ISO 9241-11 standard. Second classification includes understandability, learnability, portability and attractiveness, and is known as ISO 9126-1 [3].

Table I illustrates six usability dimensions and their corresponding attributes. Brief description of these usability classification attributes are:

- Effectiveness: How accurately the user achieves the goals by using the app?
- Efficiency: How much resources are consumed to perform certain tasks?
- User Satisfaction: How do the users sense about their experience by using the applications?
- Learnability: Is the system easy to learn? Novice users should be able to complete basic tasks in a short period of time with minimum training.
- Memorability: The system is easy to remember/memorize. Users can return to it after a long period of time and complete tasks without retraining.
- Error Rate: If a user faces an errors while using the system, the system is capable to auto recovery.

TABLE. I. USABILITY DIMENSIONS AND THEIR ATTRIBUTES

Components of Nileson Usability Framework			
Usability Dimensions	***Contextual Factors***	***Threshold values for dimensions***	***Classification standards***
Learnability Memorability Effectiveness Efficiency User Satisfaction Error Rate	Users Technology Activity Environment	Time Error rate Number of apps Number of users	**1. ISO 9241-11** Effectiveness Efficiency User Satisfaction **2. ISO 9126-1** Understandability Learnability Memorability

There are numerous ways to access the usability of smartphone apps. The prominent methods include expert reviews, user testing, field experiment, laboratory experiments, system usability scale measurement and user surveys. The questionnaire consisted of four parts, i.e., demographic (This part of the questionnaire gathered the demographic facts of participants, which includes age, gender, schooling, and earnings, and requested contributors to document their level of enjoy with smartphones and methods to connect to the mobile Internet. Participants' revel in with smartphones changed into measured with the aid of utilization hours in step with day and the total utilization duration) statistics, customers' preferences for the layout capabilities of smartphones, customers' reputation of smartphones, and customers' utilization behavior of smartphones [4].

Applying a heuristic evaluation approach using SMASH turned into shown to be effective in figuring out a huge percentage of the usability issues the aged customers confronted at the same time as interacting with a smartphone. The usability troubles had been not most effective because of UI design, a number of the problems had been due to problems of the elderly in performing the gestures which carried out to the corresponding undertaking; particularly, the “drag and drop”, and “faucet and keep” gestures. Therefore, it is recommended that the usage of those gestures be removed or at least reduced.

Usability and user experience is a vital quality attribute for apps. The hedonic aspects such as fun, emotions and enjoyment are focused by user experience. Hedonic aspects meet up the universal needs, but they do not essentially have a utilization worth. User-ability is another aspect which is integrated with the user experience to determine whether the user felt pleasant or otherwise with the system during the inspection. An approach is used by the users in which they draw and write their hedonic aspects about the application [5]. This will leads to map the usability attributes on the usability heuristics as follows:

1) Visibility of the system (system should keep the user informed about all processes and state changes through feedback within a reasonable amount of time).

2) Match between system and real world (should speak users language instead of system oriented language).

3) User control and freedom (Allow users to undo and redo the previous tasks).

4) Consistency and standards (the user should be able to do things in a familiar standard and consistent way).

5) Error prevention (display appropriate mock up message).

6) Minimize the users memory load (interface should show the visible objects, actions and options).

7) Efficiency of use and performance (animated icons and transmissions should be displayed efficiently).

8) Aesthetic and minimalist design (avoid unwanted information, particularly the one that is out of context).

9) Help and documentation (easy to find different content and documentation help is available).

10) Help users identify, analyze and improve from errors (error should be displayed in user familiar language instead of system language). Customization and shortcuts (shortcuts are available for frequently used actions).

Physical interaction and ergonomics (should provide physical sense like buttons, position, etc.) [6].

If usability heuristics evaluations are conducted on a functioning/working product, the experts need to have some specific user tasks in mind to focus the inspection in the right direction [7].

Usability is an important factor as if an app is difficult to use then users would quit using that specific app. Usability testing in a right manner, at the right time and with the right observation would reduce the software risks of building the wrong product. For evaluation of smartphone applications, usability assessment is crucial so that developers can learn how to adopt them and consider the dynamicity of mobile scenarios.

Usability evaluation of smartphone applications is a potential research context that faces a number of challenges. These challenges emerge due to the unique restricted features of mobile phone such as limited bandwidth, varying environmental factors and unreliable network [4]. Additional challenges includes lack of usage-based testing and response, limited focus on interface architecture, navigation ignorance and connectivity, restricted resources and web connectivity issues. The technical capabilities of mobile apps and achieving high level user satisfaction are crucial for the success of mobile apps. Hence, usability testing of mobile apps is mandatory process to ensure that mobile apps are practical, effective and easy to use [8].

Because of the significance, a large number of usability guiding principles have been modeled to support the structure of usable applications. These guiding principles are specially proposed for web-based and desktop applications. Due to the mobility nature of smartphones devices, the smartphone apps are different in many ways from the conventional applications [9]. To date, the guiding principles for usable smartphone applications are isolated and limited. This adds difficulty and complexity to evaluate the usability of smartphone apps [10].

This aim of this study is to explore the various usability dimensions and the corresponding issues that need to be considered while designing and evaluating mobile apps. In this study, we evaluate different usability dimensions along with their testing parameters that are necessary to be adhered to ensure better quality of the mobile apps and their user-friendliness.

II. Literature Review

Moumane et al. [3] present empirical evaluation for the use of software quality standards ISO-9126 in mobile environments and highlight issues related to the software such as user guides, use of simple data entry procedure and existence of online help. The study performs hardware based evaluation such as display resolution, memory size and screen size by using ISO-9241 and ISO-5062 standards benchmark. The authors analyze two usability evaluation standards for mobile apps including ISO-9241and ISO-25062.ISO standard 9241is a base quality model and includes efficiency and satisfaction as the usability dimensions. The level of user satisfaction against three usability factors were evaluated as 62%,33% and 20% respectively. ISO standard 25062 includes two usability factors reliability and portability. The proposed framework was compared with ISO standard 9126 on the basis of three usability measures efficiency, effectiveness and satisfaction by conducting a survey.

Sorber and Kortum [11] addressed the subjective usability of a large number of mobile apps for both tablets and phones across Android and Mac operating systems and target the consistency measure in this context. The objective of the study is to propose a baseline for usability measures for mobile apps. The author describe the usability on a large number of mobile apps for both tablets and phones across Android and Mac operating systems; characterize these results on system usability scale and describe the usability measures for consistency. The proposed solution is used only on small scale. Future work could be examining the usability of mobile apps in a more formal laboratory based environment.

Lu and Wei [12] analyze the level of enjoyment, mobility use towards the persistence use of smartphone apps. The goal of study is to revise the IS (Information System) continuance model to highlight the role of enjoyment and mobility of user perception towards the continuance of user intention to use the smartphone application. The IS continuance model is based on the ECT(Expectation Theory of Continuance).ECT is mostly used in the literature for consumer behavior to analyze the consumer satisfaction, service marketing in general and post-purchase behavior(e.g., repurchase, complaining), (Anderson and Sullivan1993;The usability measure include satisfaction, performance expectancy, continuance intention, post usage attitude and effort expectancy. The study defines the context of smartphone apps broadly and data were analyzed on a smaller scale. Future work could be to validate the proposed model on large scale by using some other design methodology.

Baek and Yoo [14] evaluate the usability attributes for "branded mobile applications" to measure and conceptualize the underlying scope of usability. The research objective is to propose and examine the measurement tool to explore the perceived level of usability application. Study proposed the usability factors as user friendliness, omnipresence speed, fun and personalization, for app and proposed a valid and reliable exploration of the usability application that included user perceptions. The study develops the usability scale development model. Variable of interests in data gathering were explored and measured using self-report survey. Future study could be conducted to validate the usability scale for usability using a randomly selected model from other mobile user population based on different types of branded apps like native, hybrid and desktop apps.

Lorusso et al. [15] address the usability and learnability of NFC based application. This research aims to explore the user satisfaction, usability, learnability and quality of the interaction between the children who have language disorder and the system application. Autonomy level, feedback, satisfaction and learnability are usability variables. Limited numbers of activities are offered by the selected prototype/hypothesis. As a future dimension, a systematic research should be conducted in educational environment.

Hussain et al. [16] explore the usability on kindle app for smartphone platform. For this purpose usability attributes as visibility, efficiency, satisfaction and enjoyability. Study proposed a descriptive based statistical methodology to evaluate the usability attributes. In laboratory based experiment 15 participants were chosen randomly with different age groups ranges as 18-29, 30-39 and 40-49. Five tasks were performed by the participants. One minute is set for execution of task. The test session is recorded by video camera. The front screen recorded the emotions, time, error and navigation from one page to another page of the users. A post-test questioner is given to the users. There were three measurements such as time, error, and frequency. Quantitative data is gathered by the test results and analyzed with the SPSS. Descriptive statistics include Min, Max and Standard deviation were used to analyze and present the resulted data, these resulted values are mapped on Likert scale which has values as strongly agree, Disagree, Natural, Agree and Strongly disagree. Future studies could be conducted on large no of participants when the sample is projected to the large population.

Nascimento et al. [5] explore the usability by addressed the relationship between user experience and usability. Study proposed a technique “userability” to evaluate the usability for mobile applications. The userability is the integration of user experiences and usability which helps to the developer designer as well as non specialist in the domain of human computer interaction. Proposed methodology consists of two further steps as heuristic evaluation for usability and 3E method. The 3E method stands for Expressing Emotions and Experiences. Study define the ten aspects from heuristic evaluation and from 3E method the two questions are finalize for evaluation as "what users feel regarding ten heuristics aspects and "what are the improvements did users feel for this aspects. Satisfaction is the main attributes in 3E and ten heuristic aspects for this evaluation emojis of face expression is used. The satisfaction attribute is scaled on questioner as unsatisfied, moderately satisfied, little satisfied very satisfied and very satisfied. The proposed methodology phases include the steps as Training, Application scope, scope of the activities, qualitative analysis of inspection questioner Detection issues and Data analysis. Five applications are used measurement are time, no of errors, no of duplicated issues, suggestions and duplicated suggestions. Grounded Theory method is used for validation to perform data coding.

Salman et al. [6] evaluate the usability of user interface for smartphone applications used by elderly. Heuristic evaluation methodology is used for this purpose. Issues are highlighted during the expert testing sessions. Identified problems are classified into different categories as appearances, language, dialogue and information. These classifications are further divided into sub categories and proposed solution for user interfaces. From heuristics evaluation two heuristics are frequently violated as minimize the user memory load and match between system and real world. The test session is performed by the 5 experts who have different age groups and have different qualification in human interaction domain on smartphone model J7. By the end of the results, 27 problems are highlighted extracted from a checklist. In future studies, a think aloud session would be conducted with the elderly participants at the time of development and design a prototype after getting their relevance feedback.

Lee and Lee [13] evaluate the usability attributes for augmented reality mobile application. This research aims to develop a tool for creating user-based design interfaces in mobile augmented reality (MAR) education and check list for usability with factor analysis and reliability. Examine the usability attributes of multimedia AR and to develop a usability evaluation tool through concretization. This study examined the usability attributes like learnability, ease of memory, usage convenience and satisfaction. The evaluation items collected from existing research were used as basic data for developing the usability evaluation checklist survey was conducted with 122 experts, and after factor analysis and reliability analysis, the final checklist for each usability evaluation item was prepared. Affordance and presence are main measures of reliability with cognitive affordance, sensible affordance, and physical affordance .proposed the usability evolution tool with focus group interview, factor analysis and reliability on evolution question. Proposed evolution steps do not validate with the existing model. Future work could be to validate the proposed usability steps on large no of applications in formal laboratory experiment.

Liu et al. [17] address the usability aspects under acceptance and usage behavior of smartphone applications. Factors are analyzed regarding acceptance and behavior. Questioner was developed which is filled by 842 participants. Acceptance is measured as usefulness, ease of use, and intention of use. Nine factors are explored for acceptance as element for design interface, physical smartphone characteristics, feedback for touch screen, display screen, connectivity and application. A questioner is constructed comprising four parts demographic information (gender, age, education and income etc.), user preferences, features of smartphone (icon size, icon color, shape, font size, character spacing, etc.), users acceptance and their usage behavior (task based like shopping Skype chatting, etc.). Data were collected an online survey of smartphone users from a Chinese website. For data analysis EFA (exploratory factor analysis) is used to detect the critical design features factor and CFA (confirmatory factor analysis was) is used to check reliability and validity for measurement constructs.842 participants including 378 male and464 are female having at least 4 years' experience for using smartphone, age groups ranges as 20 to 51 years. Age attribute is divided into categories then find frequency of each age groups among male and female and result the mean and standard deviation. These studies result is generating by considering the android operating system on apple or any other operating system results may differ or better just because of hardware limitation as display resolution speed, etc.

III. Critical Analysis

In this study after reviewing the reviewing the literature the usability assessment methods for testing are briefly described and critically evaluate the literature to explore various usability dimensions as well as corresponding challenges that need to be considered while designing and evaluating the mobile applications. By considering the various usability dimensions, attributes, performance measures, contextual factors, testing parameters, proposed model and validation model as well.

The attribute of the critical table are Problem addressed, Usability dimension, Implementation method, platform used for application, standard model for comparison or validation purpose, mapped with Nielsen findings for user interface named as SMASH (smartphone usability heuristics) and limitation or future findings.

IV. Key Challenges

Since mobile technologies are used in every field, smartphone apps play a vital role for their success. So usability is a crucial factor to achieve the quality goal but usability testing of smartphone apps faces number of challenges like:

Connectivity [3,8]: The slow, unchangeable network association with small size bandwidth is an ordinary difficulty for smooth execution of smartphone apps. The difficulty mainly affects loading time and worth of stream media.

Small screen size [3,8]: The diversity of element ratio and pixel solidity can be massive.

Different display resolutions [4]: As different display resolution may produce different usability evaluation results, short resolution can disgrace the quality of information display on mobile devices.

Context of mobile [14]: It may illustrate as any statistics that differentiates a situation linked to conversation between user, apps and the encompassing historical past. It naturally consists of location identities of close by humans, gadgets and environmental basics that could divert person awareness.

Capability and Limited processing power [15]: Some smartphone apps want large quantity of remembrance to GUI assist which include three dimensional apps.

Lesser focus on navigation [14]: Interface structure play a critical position to get the consumer pride stage. On interface structure there's a number of unnecessarily links or buttons which burdened the consumer and frustrated pop-up messages on each second.

Lack of use testing and response [8,15]: The most important challenges in usability testing is the lack of user testing and their feedback during the design evaluation process. There is a need to get the acknowledgement from user to determine their needs, intentions, usability obstacles through a descriptive or statistical measure.

Not strengthening the engagement loop [3,15]: Designer should carefully design the app by getting the user experiences, their preferences for the application by performing the tasks and taking the feedback.

V. Conceptual Framework

Conceptual framework for usability evaluation of smartphone is described in Fig. 1.

Conceptual framework is the approach to represent a general concept, that guide people appreciate or reproduce the domain of that model. Conceptual models illustrate the relations between factors and the stream of data or processes. A conceptual model comprises the four fundamental characteristics as the potential reward to implement a theoretical model are numerous, but mostly depend on the own capability to invent a well-built model in the primary place. The key rewards of theoretical model include.

Description: The above conceptual model is designed for usability testing of Smartphone applications. The conceptual model comprises of six steps as selection of usability attribute, usability evaluation based criteria, fetch threshold values for usability attributes, select the application for usability testing, design test case generation on the basis of usability criteria, Test execution and check the criteria meet for validation. Three mini table which support to select the attributes as (learnability, efficiency, memorability, error, satisfaction and effectiveness) standards and models to follow for the basic initiation process as (ISO 9241-11 and ISO 9126-1) and measuring parameters as (time, error rate, number of users and number of applications laboratory or field experiment, and contextual factor as well) for usability testing as stored in a tabular form. All these steps should be followed before designing the smartphone application. Arrows show the flow of data, boxes indicate the main steps, diamond symbol show the decision in form of "Yes" or "No".

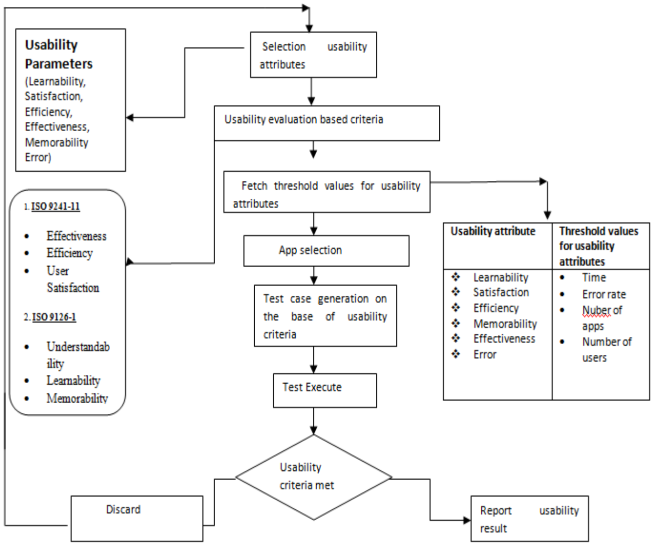


Fig. 1. Conceptual Framework for usability Evaluation.

Selection usability attributes: In the first step usability selection is the initiation process to usability of smartphone app. The user selects the usability dimensions as per requirement or his scope of the work. The usability attribute are stored as a list in the tabular form as learnability efficiency memorability, error, satisfaction and effectiveness, effectiveness, efficiency and user satisfaction are mostly and commonly used as per literature review. so selection the usability is a vital for further processing.

Usability evaluation based criteria: In this phase the usability evaluation based criteria is chosen to access which standards or models are the based for further processing. The criteria list for usability testing as standards and model like (ISO 9241-11 and ISO 9126-1) are stored in the tabular form with their different usability parameters.

Fetch threshold values for usability attributes: In the third step after selection of usability dimension and criteria the threshold values are selected. The threshold values must be defined to meet or to compare the mature results like volume of expected traffic, error rate, time etc. The threshold values may be quantitative or statistical measure like time, frequency, min, max, error rate, number of users and number of apps, etc. The threshold measurement values are fetched from a table in which these values are listed.

App selection: After set the usability dimension their testing parameters there is need to implement on an application. The application is selected on bases of the nature of the testing. The contextual factors as technology, type of users, activity and environment etc. play a vital role for selecting the application. Which platform, operating system is used for testing this will be set in this phase.

Test case generation on the base of usability criteria: In the 4th step when the application, platform operating system is selected then there's a need to design a prototype for testing. Prototype is usually used to estimate a brand new design to growth accuracy with the aid of device examination and consumer. A prototype is an premature pattern, reproduction or freed from a product built to check a belief or procedure or to carry out as a thing to be simulated or found out from. This can be questioner, laboratory, field experiment or in a controlled environment.

Test Execute: After selecting the platform, the operating system and the application the designed prototype is being executed by the users. The method of execution could be a pre or post questioner, advance techniques like eye tracing and facial recognition is used to get user experiences, and emotions "3E method" is used which explore the expressing emotions and experiences.

Usability criteria met: In the last step execution of the designed prototype and calculate the result after the processing. These results are validated with an existing model or prototype result. Results are noted and validate with the models that is chosen on the second step as criteria for usability evaluation. If the criteria meet the with the model or show better performance and then results are documented in the form of report But if the criteria is not meet then move back to the step one and repeat the process by changing the attributes or performance parameters for good results. evaluation results are met with the defined criteria or enhancement is occur while comparing then report is generated ,if the evaluation results are not met with the pre-defined criteria.

VI. Conclusion and Future Work

Usability is recognized as a significant quality dimension to determine the success of mobile applications. This study highlights the techniques which are being applied to evaluate the usability of smart phone applications. Usability assessment methodologies are evaluated for different types of applications running on different operating systems like Android, Blackberry and iOS etc. Assessment methods of usability testing are discussed in a great detail to explore various usability dimensions as well as corresponding challenges that need to be considered while designing and evaluating the mobile applications. Specifically, the study conducted a critical review of various usability dimensions, attributes, performance measures, contextual factors, testing parameters, proposed model and validation model as well. The prominent challenges identified in this study include: connectivity, small screen size, different display resolution, information input method, context of mobile, capability and limited processing power, navigation ignorance, no focus on interface architecture etc. A conceptual framework for usability evaluation of smartphone apps is also designed which would be validated through experimentation in the thesis work. This study is particularly useful to comprehend usability issues and their likely remedies to produce high quality smartphone apps. The study provides a conceptual framework for usability testing of smartphone applications. Future work could be conducted to validate this model in a formal manner.

References

[1] B.C. Zapata, J.L. Fernández-Alemán, A. Idri and A. Toval, "Empirical studies on usability of mhealth apps: a systematic literature review," Journal of medical systems, 39(2), 2015.

[2] R. Inostroza, C. Rusu, S. Roncagliolo, V. Rusu and C. Collazos, "Developing SMASH: a set of smartphone'susability heuristics," Computer Standards & Interfaces, 43 pp.40-52, 2016.

[3] K. Moumane, A. Idri and A. Abran, "Usability evaluation of mobile applications using ISO 9241 and ISO 25062," Springerplus, 5(1), 2016.

[4] H. Rahmat, H. Zulzalil and A. Abdulghani, "An approach towards development of evaluation framework for usability of smartphone applications," Malaysian Software Engineering Conference (MySEC), pp. 178-182, 2017.

[5] I. Nascimento,W.Silva, B. Gadelha and T.Conte, "Userbility: a technique for the evaluation of user experience and usability on mobile applications," Human-Computer Interaction, 37:3, 372-383,19 June 2016.

[6] H.M. Salman, W.F. Wan and S. Sulaiman, "Usability evaluation of the smartphone user interface in supporting elderly users from experts' perspective," IEEE Access, 6, pp. 22578-22591, 2018.

[7] D. Quiñones and C. Rusu, "How to develop usability heuristics: A systematic literature review," Computer Standards & Interfaces, 53(2017): 89-122, 2017.

[8] B.A. Kumar and P. Mohite, "Usability of mobile learning applications: a systematic literature review," Journal of Computers in Education, 5(1), pp. 11-17, 2018.

[9] R.Alturki, V.Gay and R. Alturki, "Usability attributes for mobile application: a systematic review," 7th International Conference on Computer Science, Engineering & Applications, 2017.

[10] D. Zhang and B. Adipat "Challenges, methodologies, and issues in the usability testing of mobile applications," International Journal of Human-Computer Interaction, 18(3), pp. 293-308, 2005.

[11] P. Kortum and M. Sorber, "Measuring the usability of mobile applications for phones and tablets," International Journal of Human-Computer Interaction 31(8), 518-529, 2015.

[12] J. Lu, C. Liu and J. Wei, "How important are enjoyment and mobility for mobile applications?," Journal of Computer Information Systems, 57 (1), pp. 1-12, 2017.

[13] W-H Lee and H-K Lee, "The usability attributes and evaluation measurements of mobile media AR (augmented reality)," Cogent Arts & Humanities, 3(1), 2016.

[14] T.H. Baek and C.Y. Yoo, "Branded app usability: conceptualization, measurement, and prediction of consumer loyalty," Journal of Advertising, 47(1), pp. 70-82, 2018.

[15] M.L. Lorusso, E. Biffi, M. Molteni and G. Reni, "Exploring the learnability and usability of a near field communication-based application for semantic enrichment in children with language disorders," Assistive Technology, 30(1), pp. 39-50, 2018.

[16] A. Hussain, E.O. Mkpojiogu, J.A. Musa and S. Mortada, "A user experience evaluation of Amazon kindle mobile application," In AIP Conference Proceedings (Vol. 1891, No. 1, p. 020060), AIP Publishing, 2017.

[17] N. Liu and R. Yu, "Identifying design feature factors critical to acceptance and usage behaviour of smartphones," Computers in Human Behaviour, 70(2017), pp.131-142, 2017.

A New Shoulder Surfing and Mobile Key-Logging Resistant Graphical Password Scheme for Smart-Held Devices

Sundas Hanif[1], Fahad Sohail[2], Shehrbano[3], Aneeqa Tariq[4], Muhammad Imran Babar[5]
Department of Computer Science and Software Engineering
Army Public College of Management and Sciences (UET, Taxila), Pakistan

***Abstract*—In globalization of information, internet has played a vital role by providing an easy and fast access of information and systems to remote users. However, with ease for authentic users, it has made information resources accessible to unauthorized users too. To authorize legitimate user for the access of information and systems, authentication mechanisms are applied. Many users use their credentials or private information at public places to access their accounts that are protected by passwords. These passwords are usually text-based passwords and their security and effectiveness can be compromised. An attacker can steal text-based passwords using different techniques like shoulder surfing and various key logger software, that are freely available over internet. To improve the security, numerous sophisticated and secure authentication systems have been proposed that employ various biometric authentication systems, token-based authentication system etc. But these solutions providing such high-level security, require special modification in the design and hence, imply additional cost. Textual passwords that are easy to use but vulnerable to attacks like shoulder surfing, various image based, and textual graphical password schemes are proposed. However, none of the existing textual graphical passwords are resistant to shoulder surfing and more importantly to mobile key-logging. In this paper, an improved and robust textual graphical password scheme is proposed that uses sectors and colors and introducing randomization as the primary function for the character display and selection. This property makes the proposed scheme resistant to shoulder surfing and more importantly to mobile key-logging. It can be useful for authentication process of any smart held device application.**

Keywords—Authentication; graphical password; shoulder surfing; mobile key-logging; security

I. Introduction

Access control mechanisms are widely used to protect user resources especially information asset. The legitimate user is required to authorize himself by passing the authentication technique employed on the system [1]. The conventional and widely used authentication method is login system protected with a textual password [2]. It is a variable length combination of alphabets, digits and special characters. Though it provides considerable security level, this approach has its shortcomings. To make a textual password robust against various password-based attacks, user has to select random characters, and some authentication systems require the user to change the password frequently. Users for their ease in remembering the passwords, tend to use either minimum length allowed for passwords or use common words, names or simply write them down in a notebook or a system file or use same password for multiple personal accounts, which makes the attacks like, dictionary attack, brute-force attack, hybrid attacks, social engineering attack, dumpster diving attack, shoulder surfing and key logging attacks possible [17, 19].

To overcome the limitations of Textual passwords, Graphical passwords are developed and used as an alternative method for authentication purpose [3]. As the name implies, this authentication method makes use of sequence of images or shapes instead of text, as the password.

Graphical password overcomes the drawbacks of textual password. Studies have shown that human brain can retain images more easily as compared to text [4, 5], and this property entitles graphical passwords as a more easily memorable method [6]. It is comparatively secure than textual password against dictionary, brute-force, social engineering and key-logging attacks [2] but vulnerable to shoulder-surfing attack [6, 20] where authentic user is observed while entering the password [7].

Graphical password systems can be classified as either recognition-based or recall-based approach; the latter of which is further divided into cued recall-based and pure recall-based approach [2, 7, 18]. In the recognition-based approach, the user selects a set of images from the available images in the registration phase which are recognized and reselected in the same sequence in the login phase. In the second approach, i.e., recall-based, the user recalls something that was selected in the registration phase. For this process there might be a clue given to the user–cued recall-based or no clue given at all during login phase–pure recall-based approach; the former of which is easy to use.

The existing graphical password schemes are divided into two types; image-based graphical passwords [2, 7, 8] and textual graphical passwords [1, 7]. In image-based, as name implies, images / symbols are used for the password. Whereas, the textual graphical password consists of a pie shape containing colors and sectors, which further contains different characters for selection as password. However, image-based graphical passwords are susceptible to shoulder-surfing attack as images can be retained easily in mind [4]. Many textual graphical password schemes are developed but they lack efficiency in terms of login time and robustness against shoulder-surfing and mobile key-logging attacks.

In this paper, we have proposed a textual graphical password scheme for smart held devices that is resistant to shoulder surfing and mobile-key logging attacks. This scheme is a combination of recognition-based and pure-recall based approach and incorporates randomization on every text character with a click. The rest of the paper is organized as follows. Section II reviews the related work, Section III explains the working of existing system. Section IV presents our proposed system and scheme. Section V gives an analysis of the proposed scheme and lastly, Section VI concludes the paper.

II. Related Work

S. Wiedenbeck et al. [9] proposed '*The Convex Hull Click (CHC)'* scheme which is a game like graphical authentication method where user without clicking on the images can select the graphical password in an unsecure environment. However, this scheme entails a longer authentication process. The login time consumption is reduced in a scheme presented by H. Gao et al. [10] where the author has proposed a graphical password scheme based on ColorLogin. In this scheme a group of chosen-color icons are displayed for the user to set as his pass-icons. The drawback of this scheme is a comparatively smaller password space and most importantly impracticality for colour blind users.

Prof Raut et al. [7] have presented another graphical password scheme that is also based on colours. This proposed scheme combines colours with textual characters in a pie chart. The user selects a colour and a sector as his password, however, only the characters fixed in a sector can be selected limiting the choice of characters for the user. A similar scheme is presented by Sumit H. et al. [1] which does not make use of colour in the pie chart. This scheme has the same limitation of fixed set of characters in any sector for a user to select. An image-based graphical password scheme is proposed by Pooja K. S. et al [8], for ATM systems where user has to select a sequence of images out of 16 images. The scheme provides shuffling of these images on every login but is prone to Hidden-camera attack.

Another image-based scheme is proposed by E. Darbanian et al. [11] in which a set of images are selected which are interpreted as characters at the back end. Each displayed image holds a value of a character that is translated by a pre-defined table. This scheme, however, is complex as user has to memorize the characters as well as the associated images. Mrs. Aakansha S. et al. [2] have presented an image-based graphical password scheme which is a combination of recognition and recall-based approach. In this scheme user is presented with a set of images and questions. The user has to select a number of images and three questions that are answered by clicking a specific point on the given images. This scheme provides a large password space but is inefficient in terms of time required for the login process. Another graphical password scheme is proposed by A. Ahmad et al. in [12] that comprises of textual characters shown in a grid. This scheme is tested to be robust against shoulder surfing attack but has high complexity and is not user-friendly.

L. Y. Por et al [13] presents a password scheme that is based on 'digraph substitution rules' that hides the activity performed to drive the password images. This scheme is resistant against shoulder surfing attacks as user clicks only on one of the pass images instead of both pass-images. However, this scheme requires the user to know the digraph substitution rules hence not very user friendly. Another graphical password scheme is presented in [14] by GC. Yang in 2017, and its improvement in 2018. This scheme is based on pass-position scheme which is similar to pass-point approach however in pass-position a relative value of the clicked location is also accepted rather than an exact value. This feature makes this approach user friendly but on the other hand prone to accidental logins, hence less secure.

A. Mishra et al. [6] have presented an image-based graphical password scheme which is based on falsification method. This scheme is resistant to graphical password attack but has poor security against key-logging attacks as user has to enter the credentials using a keyboard. Another image-based graphical password scheme is presented by K. Irfan et al. [15] which use both image-based and test-based approaches. The user selects few images on registration which are reselected on login. If these images match with the images stored in database, the user is asked to change his password by selecting new images. This approach makes this scheme less practical for changing the password on every login.

III. Existing System

The existing system [1] is a type of textual graphical password scheme that consists of pie that is divided into 6 sectors each having 12 randomly distributed characters in them. The divided sectors contain 72 characters in all; having 26 alphabets (26 upper case and 26 lowercase), 10 special characters and 10 decimal digits (0-9). The login page of the existing scheme is shown in Fig. 1. The user verification is done in two stages.

A. Registration Phase

The registration phase consists of the following steps;

1) The user selects one of the six sectors from the given pie shape, that is used as the pass-sector for further logins.

2) The selected sector has 12 of the 72 randomly placed characters that constitutes the textual password of the user.

3) The sector number and the textual password is encrypted and stored in the password table in the system.

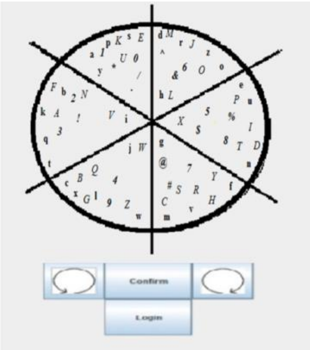


Fig. 1. Login Screen of the Existing Scheme [1].

B. *Login Phase*

The login phase consists of the following steps:

1) Upon login the system displays the pie shape having 6 sectors.

2) The system places the 72 characters equally divided among all the sectors.

3) The user has two buttons “Clockwise” and “Anti-clockwise” available.

4) The user by using these buttons in desired direction rotates the sectors in order to match the selected sector number and the characters of the set password.

5) When the desired sector coincides with the sector number, the user clicks the “Confirm” button and is allowed to successfully login into the system.

The existing system offers a fixed set of characters in each sector of pie for the registration or login phase. Thus, making the characters predictable, selected as password, as user has to select whole sector as the pass-sector. Secondly, a fixed sector containing the complete set of characters of password makes this approach vulnerable to mobile key-logging attack.

IV. Proposed Scheme

The proposed scheme is an improvement over the existing scheme and overcomes its shortcomings. It is a combination of recognition-based and recall-based graphical password scheme. Similar to the existing scheme, the proposed scheme is also based on text and sectors contained in a pie. The architecture of the proposed scheme is shown in the Fig. 2. It has been developed for Android based smart-held devices using Android studio, languages used are Java and Android.

The user selects the password from a range of 72 characters randomly distributed over 8 sectors during registration and login phase. The character set contains 26 upper-case and 26 lower-case alphabets, 0-9 digits and 10 special characters (@, !, #, $, %, *, &, ?, <, >). For securing the user credentials, the selected password along with other user-entered information, are hashed using hashing algorithm SHA-1 [16] and stored in the database, during both registration and login phase. Moreover, characters are always shuffled at the run time which will ensure security against shoulder surfing and mobile-key logging.

The proposed scheme comprises of a user registration and login phase that are described as follows:

A. *Registration Phase*

In Registration phase, user has to fill the registration form which includes First name, Last name and Email. Then user has to set the graphical password by first selecting the desired colour from the 8 available colours of the sectors. After that, the user will select different characters from the pie chart. When this data is submitted, the password is hashed using the hashing technique i-e SHA-1 (Secure Hash Algorithm). Fig. 3 shows the user registration process. SHA-1 is the cryptographic hash function which takes an input and produces a 160-bit (20 Bytes). Then this hashed password and all user details are stored into the database. The minimum requirement for the password is 8 characters that must contain 1 numeric, 1 uppercase letter, 1 lowercase letter and 1 special character.

The registration phase consists of the following steps:

1) On the registration page (Fig. 4(a)), user is asked to enter first name, last name and a valid email address. The user email will be used as username.

2) After filling these fields, the user selects the *‘Generate password’* button and a new screen appears containing the pie chart with 8 sectors each having a different colour that are selected separately (Fig. 4(b)).

3) There are four buttons below the pie chart for the character selection:

a) First button is of upper-case letters (A... Z), by touching this button, upper case letters will randomly display in the sectors of the pie chart.

b) Second button is of numbers (0… 9), by touching this button, numbers will randomly display in the sectors of the pie chart.

c) Third button is of special characters (#*$%), by touching this button, special characters will randomly display in the sectors of the pie chart.

d) Forth button is of lower-case letters (a… z), by touching this button, lower case letters will randomly display in the sectors of the pie chart.

4) User selects the desired color and then selects characters for his password meeting the minimum requirement set for the password. The colour and characters are selected by simply touching the screen on the desired point.

5) Upon submitting the selected graphical password, the user Email and password are hashed using SHA-1 and stored in database.

6) After the complete registration process, login page appears.

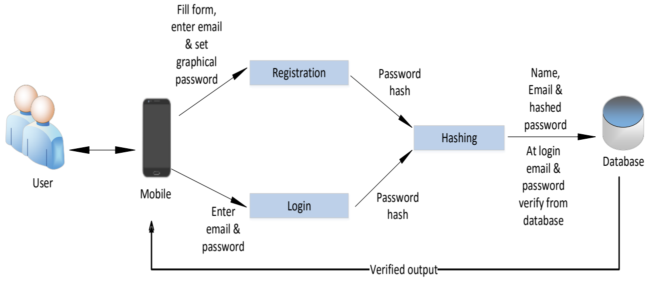

Fill form, enter email & set graphical password
Registration
Password hash
Name, Email & hashed password
User
Mobile
Hashing
Database
At login email & password verify from database
Enter email & password
Login
Password hash
Verified output

Fig. 2. Proposed Scheme Architecture.

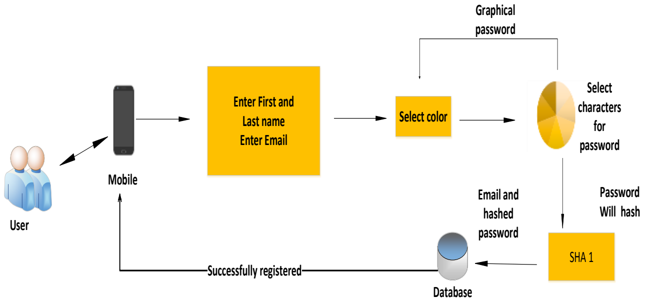
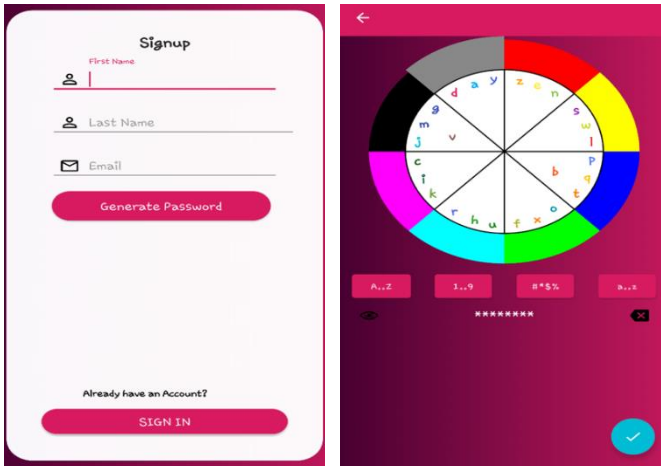


Fig. 3. User Registration.



(a) (b)

Fig. 4. (a). Example of a Signup Page. (b). Example of Password Generate Page.

B. *Login Phase*

Fig. 5 shows the user login process. When user submits the email and the graphical password, the entered password is hashed through the same hashing algorithm and compared with the already stored hashed credentials of the user in database. When both the hash values match, the user is verified and allowed to login to his account. If user forgets his password, it is sent to the user's linked email address.

The login phase consists of the following steps:

1) Step 1 is based on recognition-based approach in which user has to enter the username i.e., user email address. The application generates a message of incorrect email if user enters wrong username (email).

2) After entering the correct username, the user has to enter the graphical password by selecting the button shown on the login screen (Fig. 6(a)). It is based on pure recalled-based approach. The user enters the password that was set during registration repeating the same procedure as of password generation (Fig. 6(b)).

3) Let L = length of the set password (exclusive of selected colour) and I be the ith character of the password. Whichever of the four buttons of characters are selected, the characters in the sectors will randomly be permuted if I < L.

4) Once the password is entered, the user submits the password which is hashed and verified from the database. Forget password button takes user on the recover password page in case user forget his password.

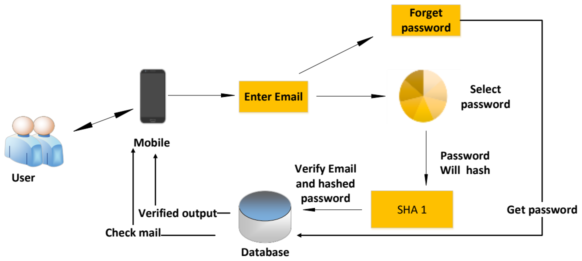
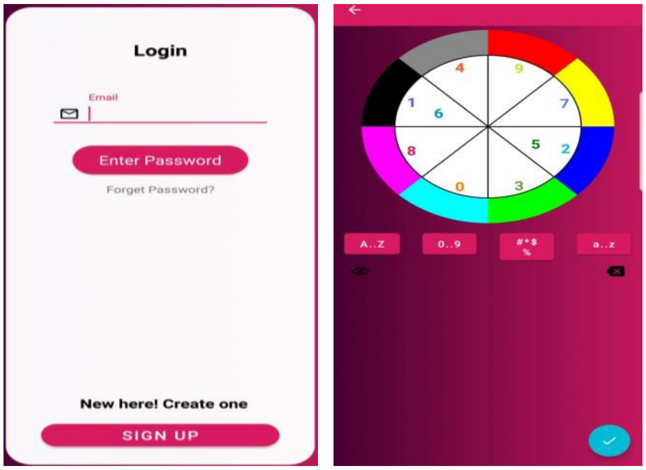


Fig. 5. User Login Process.



(a) (b)

Fig. 6. (a). Example of a Login Page. (b). Example of Enter Password Page.

V. Analysis of Proposed Scheme

In this section, the security analysis of the proposed scheme is made against shoulder surfing and mobile-key logging attacks and a comparison is done with the existing scheme.

A. *Large Password Space*

The proposed scheme is resilient against Brute force and Password guessing attacks. The total number of possible passwords that can be made with this scheme, make these attacks difficult. With the 72 characters, 8 sectors and considering the password length as '*L*', our scheme has a very large password space as compared with the existing scheme [1] which can be calculated as shown in (1):

$$Total\ no.of\ possible\ passwords = \sum_{L=8}^{15} 8 * 72^{L} = 5.88 * 10^{28} \quad (1)$$

B. *Resistant against Accidental Login*

The probability of correctly entering any character of the password is 1/72. The success probability of accidental login (P$_{al}$), with a password of length L, is calculated in (2)

$$P_{al(L)} = \left(\frac{1}{72}\right)^{L} \quad (2)$$

The comparison of password space and P_{al} of proposed scheme with existing scheme [1] is shown in Table I.

TABLE. I. Comparison of Proposed Scheme with Existing Scheme

Feature	Existing Scheme [1]	Proposed Scheme
Password Space	$4.346 * 10^{28}$	$5.88 * 10^{28}$
P$_{al}$(L)	$(\frac{1}{12})^{L}$	$(\frac{1}{72})^{L}$

C. *Robust against Shoulder Surfing and Mobile-Key logging Attack*

The proposed scheme is robust against shoulder surfing and mobile-key logging attacks. The existing schemes have a fixed set of characters displayed in each sector after the first display which shows them randomly but after that, they are fixed for the whole session of registration or login whereas in our proposed scheme, the randomization works on every single click till the length of the password to be selected during registration or login.

This randomization works independently on each character and places them in a different sector upon every click thus decreasing the probability of a sector being selected in the same sector with the same characters. This increases the security level of our scheme as character position is totally unpredictable.

Let probability of a character to be placed in any sector to be denoted by P(c) which can be calculated as shown in (3) and (4).

$$P(c) = \frac{1}{Total\ No.of\ characters} \quad (3)$$

$$P(c) = \frac{1}{72} = 0.014 \quad (4)$$

This factor plays an important role against Shoulder Surfing attack and mobile-key logging attack as the probability for the attacker to correctly guess any character on a given location is substantially low even if he tried to memorize the password characters and tries to login into the system.

Another important factor that makes our scheme strong against shoulder surfing attack is the ***rotation feature*** of the pie chart. The user can rotate colour rim and the sectors independently in clockwise or anti-clockwise direction. The selected colour and the sectors don't have to be aligned to enter the password.

This feature enhances the security of our scheme against shoulder surfing attack. An attacker cannot guess the associated colour with the characters as they need not be aligned with any particular sector for the password to work.

The proposed scheme provides strong authentication process in case of password recovery scenario. If a user forgets his password, the system sends the password to the linked email account of the user. This is very beneficial for the authentic user but any imposter trying to login into the system cannot access the system in any way.

Thus, our proposed graphical password scheme is highly secure and easy to use.

VI. Conclusion

User authentication plays a vital role in securing user accounts and confidential information. In this paper, a new graphical password user authentication scheme for smart-held devices is presented which is a combination of recognition and pure recall-based graphical password approach. With a combination of colour and alphanumeric characters, this scheme is viable for users comfortable with textual passwords.

The proposed scheme provides high security against Brute-force attack as it offers a very large password space as compared to the existing scheme. The randomization feature, incorporated with every click, adds robustness against shoulder surfing and mobile-key logging attacks. In case an authentic user forgets his password, the password is email to the user hence adding another layer of security and making an attacker unable to get hold of the password. The proposed authentication graphical password scheme is designed for smart held application and can be easily used as a secure gateway for any application.

The work can further be extended by increasing the number of characters and sectors in the pie, and by also increasing the number of colours of the sectors. To further enhance the robustness of this scheme, two factor authentications can also be incorporated.

References

[1] S. H. Wagh, A. G. Ambekar, "Shoulder Surfing Resistant Text-based Graphical Password Scheme", ICCT 2015, International Journal of Computer Applications (0975 – 8887).

[2] Mrs. A. S. Gokhalea, Prof. V. S. Waghmare, "The Shoulder Surfing Resistant Graphical Password Authentication Technique" 7th International Conference on Communication, Computing and Virtualization 2016, Procedia Computer Science 79 (2016) 490 – 498.

[3] X. Suo, Y. Zhu, G. Scott. Owen, "Graphical Passwords: A Survey", Department of Computer Science Georgia State University. 21st Annual Computer Security Applications Conference (ACSAC'05), IEEE

[4] Kirkpatrick. "An experimental study of memory", Psychological Review, 1:602-609, 1894.

[5] R. Shepard. "Recognition memory for words, sentences and pictures", Journal of Verbal Learning and Verbal Behavior, 6:156-163, 1967.

[6] A. Mishra, R. Jadhav, S. Patil, "A Shoulder-Surfing Resistant Graphical Password System", International Research Journal of Engineering and Technology (IRJET), Volume 5, March 2018.

[7] Prof Raut S.Y., J. B. Baviskar, K. Rahul S, S. Aditya N, S. Yogesh S, "Shoulder Surfing and Keylogger Resistant using Graphical Password Scheme", International Journal of Advanced Research in Computer Science, Volume 5, No. 8, Nov-Dec 2014.

[8] Pooja K S, P. V. Dhooli, Prathvi, Prof. Ashwini N, "Shoulder Surfing Resistance Using Graphical Password Authentication in Atm Systems", International Journal of Information Technology & Management Information System (IJITMIS), Volume 6, Issue 1, January - June (2015), pp.01-10.

[9] S. Wiedenbeck and J. Waters, L. Sobrado, J. C. Birget, "Design and Evaluation of a Shoulder-Surfing Resistant Graphical Password Scheme", AVI '06, May 23-26, 2006.

[10] H. Gao, X. Liu, R. Dai, S. Wang and H. Liu, "Design and Analysis of a Graphical Password Scheme", Fourth International Conference on Innovative Computing, Information and Control, 2009.

[11] E. Darbanian, Gh. D. Fard, "A Graphical Password Against Spyware and Shoulder-surfing Attacks", International Symposium on Computer Science and Software Engineering, IEEE, 18-19 Aug. 2015.

[12] A. Ahmad, M. Asif, M. Hanif, R. Talib, "Secure Graphical Password Techniques against Shoulder Surfing and Camera based Attacks", International Journal of Computer Network and Information Security · November 2016.

[13] L. Y. Por, C. S. Ku, A. Islam, T. F. Ang, "Graphical password: prevent shoulder-surfing attack using digraph substitution rules", Higher Education Press and Springer-Verlag Berlin Heidelberg, 2017.

[14] GC Yang, “PassPositions: A secure and user-friendly graphical password scheme”, 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), 8-10 Aug. 2017, IEEE.

[15] K. Irfan, A. Anas, S. Malik, S. Amir “Text based graphical password system to obscure shoulder surfing”, 15th International Conference on Applied Sciences and Technology (IBCAST), 2018, IEEE.

[16] D. Eastlake, P. Jones, “US Secure Hash Algorithm 1 (SHA1)”, RFC Editor, ACM, 2001.

[17] M. Raza, M. Iqbal, M. Sharif, W. Haider, “A Survey of Password Attacks and Comparative Analysis on Methods for Secure Authentication”, World Applied Sciences Journal 19 (4): 439-444, 2012

[18] GC Yang, H. Oh, “Implementation of a Graphical Password Authentication System ‘PassPositions’, Journal of Image and Graphics, Vol. 6, No. 2, December 2018.

[19] A. H. Lashkari, A. A. Manaf, M. Masrom, “A Secure Recognition Based Graphical Password by Watermarking” 11th International Conference on Computer and Information Technology, IEEE, 2011.

[20] Y. Higashiyama, N. Yanai, S. Okamura, T. Fujiwara, “Revisiting Authentication with Shoulder-Surfing Resistance for Smartphones”, Third International Symposium on Computing and Networking (CANDAR), IEEE, 2015.

Deep CNN-based Features for Hand-Drawn Sketch Recognition via Transfer Learning Approach

Shaukat Hayat[1], Kun She*[2], Yao Yu[4]
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China

Muhammad Mateen[3]
School of Big Data and Software Engineering
Chongqing University, Chongqing
P.R China

***Abstract*—Image-based object recognition is a well-studied topic in the field of computer vision. Features extraction for hand-drawn sketch recognition and retrieval become increasingly popular among the computer vision researchers. Increasing use of touchscreens and portable devices raised the challenge for computer vision community to access the sketches more efficiently and effectively. In this article, a novel deep convolutional neural network-based (DCNN) framework for hand-drawn sketch recognition, which is composed of three well-known pre-trained DCNN architectures in the context of transfer learning with global average pooling (GAP) strategy is proposed. First, an augmented-variants of natural images was generated and sum-up with TU-Berlin sketch images to all its corresponding 250 sketch object categories. Second, the features maps were extracted by three asymmetry DCNN architectures namely, Visual Geometric Group Network (VGGNet), Residual Networks (ResNet) and Inception-v3 from input images. Finally, the distinct features maps were concatenated and the features reductions were carried out under GAP layer. The resulting feature vector was fed into the softmax classifier for sketch classification results. The performance of proposed framework is comprehensively evaluated on augmented-variants TU-Berlin sketch dataset for sketch classification and retrieval task. Experimental outcomes reveal that the proposed framework brings substantial improvements over the state-of-the-art methods for sketch classification and retrieval.**

Keywords*—Deep convolutional neural network; sketch recognition; transfer learning; global average pooling

I. Introduction

In a human point of view, sketch analysis is not only considering a fundamental problem, but it has a prominent role in the field of human-computer interaction (HCI). Sketches can be seen everywhere and have a significant role in daily life activities, i.e., education sector, art, design, and entertainment, etc. All through human society progress, the sketch has been utilized as a fundamental tool for conveying feelings, thoughts, judgments, and opinions. Since the ancient time, the people express their views in the form of sketch related petroglyphs or cave paintings. Such kind of art examples can be easily seen today in pre-historic art caves throughout the world.

The Technological explosion makes the sketches easy and ubiquitous. There exist for several fascinating multimedia applications such as HCI [1] and some other relevant work [2-4]. With the fame of touchscreens and smart-phone devices encouraged the people to draw sketches digitally. Presently, an excessive utilization of advanced technological tools, the one needs to access query sketch more accurately and retrieve its relevant contents to be well-recognized through technological-based smart devices. However, to acknowledge the needs of the society and to balance with technological advancement, the researchers have been analyzed various novel tasks regarding sketch recognition [5, 6], and sketch-based image retrieval [7, 8], in a field of computer vision. The idea behind the sketch classification or recognition is to extract the information from the desired object class of labeled sketch-images among the pre-defined set of object-classes. Based on the extracted information, the label of the targeted instance can be correctly identified. Classification or recognition techniques usually rely on extracted features through instance training before making a recognition. For sketch recognition, the researchers borrowed handcrafted features approaches which have been successfully used to extract features from natural images. There exist, Scale-Invariant Features Transforms (SIFT) [9] Histogram of Oriented Gradients (HOG) [10], descriptors and the bag-of-features has been already utilized. In this regards, different handcrafted features techniques are followed to yield the global features for sketch recognition, i.e., GF-HOG [11], FV [12], SSIM [13] and Structure Tensor [14]. Usually, handcrafted feature representations are not considered robust and also due to their high dimensionality make them computationally expensive. Current approaches to object recognition make the necessary use of deep learning and machine learning techniques. However, the most existing work in sketch recognition is based on deep learning approaches using deep convolutional neural networks (DCNNs) and showed an impressive result than handcrafted approaches [6, 8, 15].

In the recent past, deep learning frameworks based on DCNNs shows a breakthrough in different areas of computer vision, including vision recognition on large-scale challenging dataset [16, 17]. Moreover, deep learning approaches also benefitting sketch-based recognition and can provide useful features representations by analyzing large-scale sketch dataset, such as TU-Berlin sketch benchmark [18, 19]. Deep learning is capable of generating more distinctive features from sketch images and can leverage the performance for sketch classification or recognition as compared to use hand-crafted features. Deep features for sketch recognition was first time utilized by [20] and design a specialized neural network model. As a result, the classification accuracy on sketch image dataset TU-Berlin [21] has been improved as compare to hand-crafted features. Similarly, two different well-known CNNs models,

*Corresponding Author.

namely LeNet [22] and AlexNet [16] are used to extract features from sketch images and show improvement in the recognition results [15]. On the other hand, some recent attempts utilized different layers of various CNNs architectures for features extraction for the purpose of sketch classification and retrieval [8, 18, 23].

Visual recognition or classifications by deep learning approaches are mostly rely on extracted features. Generally, Deep CNNs features are categorized into three basic levels, such as high, middle, and low-level features. Each level of extracted features are having their strengths and potential in producing results and accuracy [24-26]. In order to obtain a higher recognition or classification accuracy and reducing the computational efforts, the concept of transfer learning (TL) approach can be exploited to get more robust features by combining the learned knowledge from multiple DCNNs models [27, 28]. In TL approach, first, the Deep CNNs models are trained on the generic visual dataset, and then pre-trained models can be directly used to train on domain-specific datasets. The motivation behind the TL approach is to combine more comprehensive and relevant knowledge of input objects resulting from multiple CNN architectures and then pass them through a classifier for a final decision. We believe that by doing so, it can achieve more robust and higher recognition accuracy as compared to the one extracted through single deep CNN model.

The sketches are mostly handled through smart-phones and other portable devices for different purposes in daily life routines. In this regard, we attempt to facilitate such touch-screen environment to retrieve the query sketch contents with higher recognition rate. To overcome the existing deficiencies in the sketch recognition system and following the emerging trend of exploring deep learning for features extraction via transfer learning approach, we proposed three different well-known robust DCNNs architectures in the state-of-the-art visual recognition to the task for sketch recognition. The proposed DCNNs architectures includes Inception-v3 [29] , ResNet [30] and VGGNet [17]. All these architectures have achieved promising performance on various challenges. These networks are trained on large-scale image dataset ImageNet [31].

The main contributions in this manuscript can be presented in the following:

- A novel and efficient CNN-based framework for hand-drawn sketch recognition is proposed that exploits the strength of extracted features from the various pre-trained DCNNs via transfer learning with the utilization of global average pool (GAP) concept.
- An attempt to generate the augmented-variants of natural images paired with TU-Berlin sketch dataset for enhancing a sketch recognition performance.
- A performance analysis of three individual deep CNNs architectures compare with proposed framework in a context of transfer learning with GAP for sketch recognition. The proposed framework obtained state-of-the-art recognition accuracy on augmented-variants TU-Berlin sketch dataset and also assesses on TU-Berlin sketch dataset (without augmented-variants).
- An evaluation of the proposed framework for sketch retrieval task.

The rest of the manuscript is organized as follows: In Section 2, we briefly present related literature based on handcrafted features and deep features. Section 3 describes the overall details of the proposed approach including data preparation and augmentation variants used in this study, the concept of transfer learning, and different proposed pre-trained deep CNNs architectures utilized in the current research. Section 4 provides results, analysis and evaluations of the proposed methodology. We conclude the manuscript in Section 5 along with the future directions.

II. Related Work

We include a review work for sketch recognition utilizing handcrafted-feature methods. Further, we enclosed our review details about deep learning approaches which have been used for hand-drawn sketch recognition and retrieval task. To hold focus, we threw light on the review work entirely related to hand-drawn sketch recognition.

A. *Handcrafted Features*

Previous sketch recognition problem was handled about CAD and artistic drawings by [32-34]. After releasing a large-crowed source TU-Berlin hand-drawn sketch dataset in 2012 by [21]. This dataset gains popularity among the computer vision researchers to utilize it further for recognition tasks. Variety of traditional approaches was carried out to classify different categories of sketch dataset and was tried to achieve higher recognition accuracy. Some researchers employed hand-engineered features techniques to extract the features for sketch recognition such as scale-invariant features transforms (SIFT) [9], histogram of oriented gradients (HOG) [10] and the bag-of-features techniques [35, 36].

Although, a method proposed by [21], describe the inter-class similarities and intra-class variations in large crowed-source sketch dataset. Support Vector Machine (SVM) classifier was used to learn the sketch representation in various object categories. Original sketch benchmark proposed by [21] and was then modified by [12]. The modified work uses SIFT, Gaussian Mixture Model (GMM) based fisher vector encoding for sketch recognition and fed into SVM classifier. This approach enhances the recognition performance near to human (73.1%) [21] accuracy rate against the same sketch dataset. A star graph based ensemble matching strategy was employed by [37], it covers not only local feature, but global structures of sketches were also adopted to match them. Further, structure matching was encapsulated, and bag-of-features was learned to exploit in a single framework. Eitz et al. [25] demonstrated hand-drawn sketch classification through implementing local features vectors techniques, i.e., SIFT and other different descriptors such as spark feature, shape context, HOG, and SHOG are embedded in a bag of features model and evaluate the performance on large scale sketch-based image dataset through Sketch-Based Image Recognition (SBIR) system. In [38], the author threw light on the proposed method Symmetric-aware Flip Invariant Sketch Histogram (SYM-

FISH) for sketch image retrieval and classification. Another approach of multi-kernel features was demonstrated by [39], where different local features were extracted to analyze the sketch image, integrate them to improve the sketch recognition performance. Individually, every feature performance was calculated and found that HOG outperformed as compare to others.

Different researchers have made the efforts through handcrafted features for sketch recognition, among these, a Fisher vector spatial pooling (SV-SP) [12], sketch image representation approach raises the sketch recognition performance up-to 68.9% to come close to 73.1% [21] human accuracy on TU-Berlin sketch benchmark. Generally, handcrafted features are not considered robust, and one of the limitation is high dimensionality of these features make them computationally expensive.

B. Deep Features

Recently, deep neural networks (DNNs) are utilized for various kind of problems, which have shown immense performance in different applications, including image recognition [16, 40, 41]. Deep networks have changed the trend by replacing hand-engineered features to the learning strategy. Instead of this, a wide range of research has been conducted comprising natural image recognition. AlexNet [16] outperforms on image recognition in comparative with others, and handle the ImageNet challenge with more significant improvements. Moreover, the utilization of deep neural networks has been expended to other tasks with variant sizes of network structures and depth according to the nature of the problem.

The networks, VGGNet [17], and GoogleNet [42] with deeper structure and ability to handle the complexity limitations of neural networks were introduced. The emergence of these deeper networks laid the foundation of a vast neural network named ResNet [30] having the residual connection, to permit the network for identity mapping tasks between the layers of the network. These deep neural networks were chosen and exercised on natural images to overcome the problem. However, several deep learning approaches have been adopted for sketch recognition. For the first time, an effort has been made to specially design a deep convolutional neural network (DCNN) architecture named sketch-DNN by [20]. Another research [15] extracts sketch features from two famous pre-trained CNNs, namely, AlexNet [16] and modified version of LeNet [22] and yield little improvement in the recognition results. The major contribution has been presented in [5], a deep CNNs model namely Skatch-a-Net was introduced for sketch recognition and beats the human sketch recognition accuracy. Later on, the existing model was modified in [6] and the sketch recognition performance gap increased from 1.8% to 4.9% than human recognition accuracy. Five convolutional layers CNN was trained by [43] by taking sketch images mixed with natural images as augmented training dataset. Further, to enhance the discriminative ability of the network, the training was presented with multiple rotated version of the sketch edge map and predicted the results with the labels. Jamil et al. [8] attempts to recognize partially colored hand-drawn sketch images and implemented fine-tuned CNN on augmented TU-Berlin sketch dataset to retrieve query-based sketch images through proposed model. The author [18] applied a feature fusion approach for sketch-based recognition system to considered different layers of CNNs for features extraction from the TU-Berlin sketch dataset.

III. Proposed Method

This section describes the proposed framework based on DCNNs architectures for sketch recognition. In the proposed method, three well-known Deep CNN architectures, i.e., Inception-v3 [29], ResNet [30], and VGGNet [17] in the state of the art of visual recognition for sketch analysis are used. The weights of these architectures are available for modification. These pre-trained models downloaded from the webpage of keras [44]. The weights are loaded to all the corresponding architectures. The proposed architectures are trained on augmented-variants TU-Berlin sketch dataset. The block diagram of the proposed framework is presented in Fig. 1.

A. Data Preparation and Augmentation-Variants

To carry out this experiment, a hand-drawn sketch dataset TU-Berlin [21] is utilized. The learning performance of deep convolutional neural networks depends on the availability of a large amount of training data. Data transformation and deformation techniques are used for expending the training dataset as an additional data samples to the existing labeled one, to reduce the overfitting problem. An essential concept of the data augmentation is that; the labels of the instances remain unchanged after applying this operation. Data augmentation can improve the generalization and discriminative ability of the model [16].

The most advanced augmentation method is adopted by mixing natural images with different transformations of enhanced edge, edge maps corresponding to the sketch images in the training dataset through anisotropic diffusion approach [45]. Fig. 2 illustrates natural image, edge enhanced and edge maps of natural image transformations. This will enable the proposed framework to compare effectively natural images; its various transformations i.e., edge maps and sketch images to match for the sketch recognition task. Different transformations and mixing natural images have been used by [8, 43] and extracted the features from both type of images i.e., sketch, natural images for recognition and retrieval task. In our case, the addition of augmentation-variants to the corresponding sketch objects categories will enable the network to learn more discriminative features representations. It will also facilitate the end-user to query sketch image through sketch retrieval system.

It is stated that augmented variants with sketch images will enhance the generalization ability of trained CNN-based framework on unseen sketch images.

Most likely, the edge-map exists in the hand-drawn sketch objects. To make it easy for the CNN-based framework to handle the edge maps of the natural images, the enhanced edges of natural images are formed and model them with edge maps of natural images. Gaussian smoothing method is utilized on the edge maps of natural images to form the enhanced-edge images of natural images. Mathematically, it can be presented as:

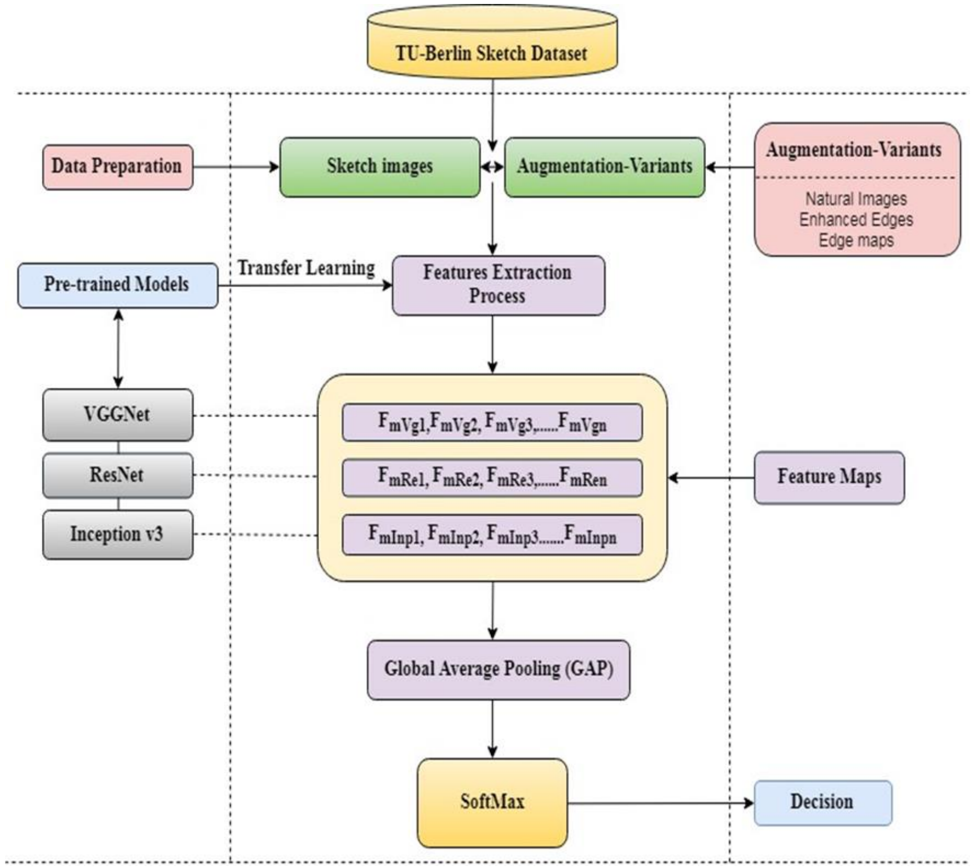


Fig. 1. Proposed Framework for Sketch Recognition.

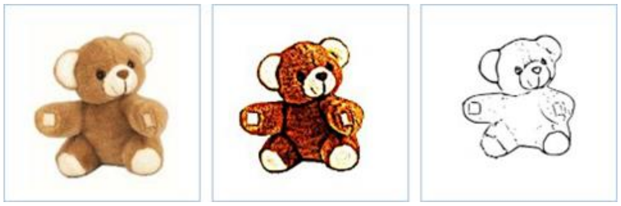


Fig. 2. Augmented-Variants of Natural Image for Training Proposed Framework, from Left to Right, the First Image is the Original (Natural Image), Second Sample Image is Enhanced Edge of Natural Image and the Last Represent the Edge Map of Natural Image.

$$I_{EE} = I(x,y) + \left[I(x,y) - \left[\frac{1}{2\pi\sigma^2} e^{\frac{-u^2+v^2}{2\sigma^2}} * I(x,y) \right] \right] \tag{1}$$

where, I represent input image, IEE indicate enhanced edge image, and x and y are the special coordinates for the image, σ = 0.5 represent standard deviation for filter and * is the convolution operation. The geometric transformation techniques are applied, i.e., flips and rotations on augmented-variants to rich the training samples and to avoid the overfitting problem.

To this end, the symbolic notations are assigned to represent the training data of sketch images, natural images, and other augmented-variants, i.e., edge maps and enhanced edges for the proposed framework. The sketch images are represented as:

$$S_{img_n} = \left\{ S_{img_1}, S_{img_2}, \ldots. S_{img_n} \right\} \in \square^{1\times n} \tag{2}$$

where, n denoted the number of training sketch images, similarly, for natural images;

$$N_{img_m} = \left\{ N_{img_1}, N_{img_2}, \ldots. N_{img_m} \right\} \in \square^{1\times m} \tag{3}$$

here, m shows the number of natural images for training. The labels assign to the natural images;

$$L = \left\{ l_1, l_2, \ldots. l_n \right\} \in \square^{c\times m} \tag{4}$$

and c shows the categories. Furthermore, the edge-maps and enhanced edge images generated from natural images and added to the relevant categories of natural images. Finally, natural images with augmented variants are sum-up to the corresponding sketch images within specified sketch object categories, i.e.

$$S_{img_n} N_{img_t} = \left\{ S_{img_1} . N_{img_1}, S_{img_2} . N_{img_2}, \ldots. S_{img_n} . N_{img_t} \right\} \in \square^{n\times t\times c} \tag{5}$$

where, n and t are the sketch and natural images with augmented-variants respectively and c represents the corresponding object category, for training the proposed framework.

B. Pre-Trained CNNs Architectures

Several state-of-the-art deep neural networks (DNNs) have been utilized for various kind of problems and gives outstanding performance in the field of computer vision application such as classification, recognition, etc [46, 47]. In the proposed methodology for sketch recognition, three different pre-trained deep CNNs models are adopted for features extraction via transfer learning, includes Inception-v3 [29], ResNet [30] and VGGNet [17]. Although the architectures of these networks are different from one another, each of the adapted model architecture is describe in the following:

1) Inception-v3: Inception-v3[29] is a deep convolutional neural network and the winner of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC)-2014. This architecture has been implemented as an updated version of the GoogleNet [42]. with a depth of 44-layers and 21 million learnable parameters. Inception module is illustrated in Fig. 3.

2) ResNet: ResNet CNN is a very deep residual network, proposed by He et al.[30] to addresses the training problems confronted by deep CNNs. This model received promising results on ImageNet. The complexity of this network is higher than other CNNs architectures due to the existence of its 152 layers. Shortcut connections are one of the critical innovation of ResNet CNN, which carries out the identity mapping, and their output is linked to stacked layers' output. ResNet CNN is illustrated in Fig. 4.

3) VGGNet: VGGNet is a CNN model, invented by visual geometry group (VGG) of Oxford University [17]. This model is the first runner-up of ILSVR-2014 for classification and the winner of localization task. The architecture of VGGNet is similar to AlexNet, and the only difference is the depth of the VGGNet. This architecture consists of 19 layers, including convolution, pooling, and three fully-connected layers. The network consists of small convolutional kernel 3x3 with stride 1. It performs better than AlexNet. The architecture of the VGGNet is shown in Fig. 5.

C. Transfer Learning

In the context of traditional machine learning algorithms, it is assumed that the characteristics of features space based on training and testing data are equal [48]. However, in a practical world, such a big amount of data is not cheap and also very hard to collect. Transfer learning is the reasonable solution to tackle such kind of problems and can provide an accurate result with less training samples.

Transfer Learning technique is widely used in the machine learning to utilize useful information from the set of source point to the set of target point [49]. The inspiration behind the adaptation of transfer learning is to solve a problem with improved results for the target domain. To be more specific, for example, the base model is first trained on relevant data instances for a specific task and then move to the target task trained by their data instances [48]. Transfer learning is the best choice for the case when the dataset of the source domain is bigger than the dataset of the target domain. If the size of the dataset for the target domain is smaller and similar to the dataset of source domain, then overfitting possibility is high. Alternatively, the chance of overfitting is reduced, and only fine-tuning of the pre-trained model is required if the size of data for the target domain is large and similar to the dataset of the source domain.

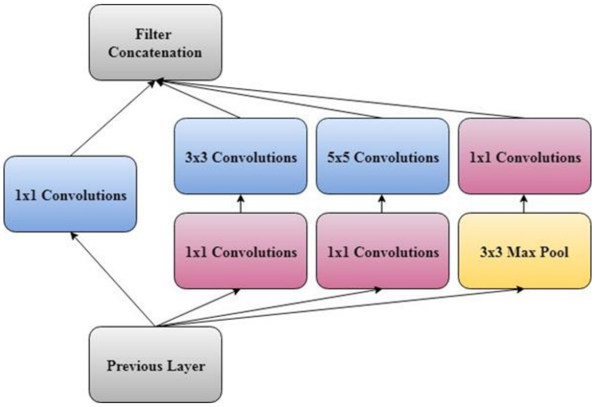


Fig. 3. Basic Inception Module.

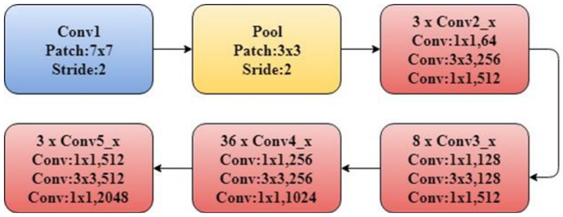


Fig. 4. Basic Architecture of ResNet.

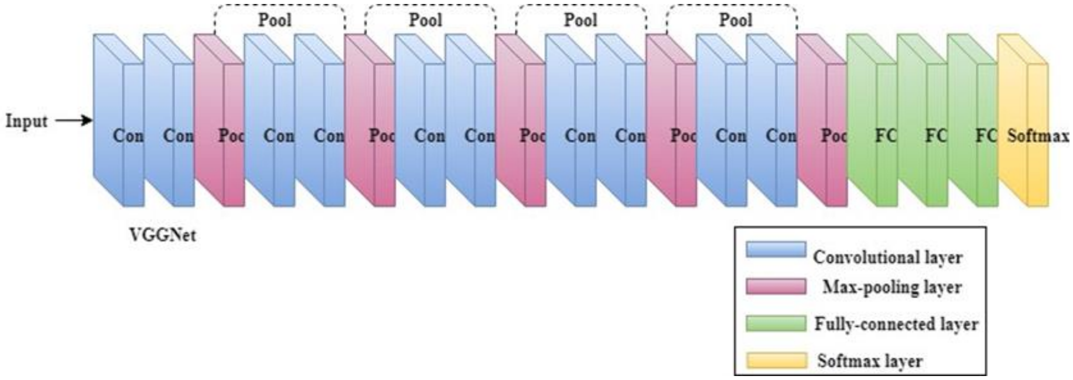


Fig. 5. The basic Architecture of VGGNet.

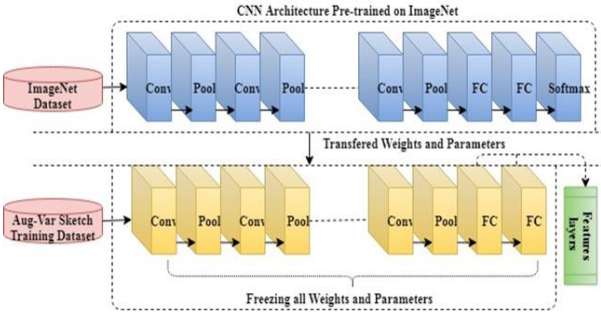

CNN Architecture Pre-trained on ImageNet
ImageNet Dataset
Conv
Pool
Conv
Pool
Conv
Pool
FC
FC
Softmax
Transfered Weights and Parameters
Aug-Var Sketch Training Dataset
Conv
Pool
Conv
Pool
Conv
Pool
FC
FC
Features layers
Freezing all Weights and Parameters

Fig. 6. Transfer Learning Concept.

In this study, three different deep CNNs models are utilized and pre-trained on ImageNet (Source domain) [50]. Based on trained knowledge, these models re-used for sketch dataset (Target domain) and fine-tuned. The concept of transfer learning is shown in Fig. 6. After that, the transfer learning approach is adopted. By doing so, it enables the models to learn more generic features from other datasets without the efforts to conduct new training. So, these architectures are trained on augmented-variants sketch dataset. The resulting feature maps from the proposed deep CNNs architectures are then concatenated into a global average pooling (GAP) layer to generate a feature vector for sketch classification.

D. Global Average Pooling

For sketch classification, proposed deep CNNs architectures extract the high-level features maps by taking advantage of convolutional layers as a features extractor. Since, as compared to single model utilization for feature maps generation, a different feature maps from the different architectures demonstrate diverse characteristics for input patterns. In this case, these distinct high-level feature maps are concatenated to maintain the discriminative knowledge about the input data. Concatenation of feature maps from different architectures can create a curse-of-dimensionality. To overcome this problem, a global average pooling (GAP) layer [51] is applied to replace all of the fully connected layers in proposed deep CNNs architectures on top of the feature maps. All the extracted feature maps are concatenated into GAP layer. This layer takes the average of each feature map and generates the features vector as the output of the GAP layer to directly fed into the softmax classifier for each corresponding sketch object category. One of the advantages of this layer is the summarization of spatial information. The resulting features vector of GAP is more robust to a spatial translation of the input images. It can reduce the total number of parameters in the network and perform dimensionality reduction to enhance the generalization ability of the networks. However, the overfitting problem can be reduced automatically without optimizing any parameter in the GAP layer.

IV. Results and Analysis

In this section, the proposed framework is evaluated for sketch recognition. The detailed description of the dataset is provided for the utilization and validation of the proposed method for sketch recognition problem. Further, the achieved results are provided and compared with state-of-the-art methods.

A. Dataset

To evaluate the performance of the proposed method for sketch recognition, a TU-Berlin sketch dataset [21] is used. This dataset consists of total 20,000 sketch images distributed over 250 object categories. Each object category is having 80 sketch objects. Total 1,350 non-experts sketch drawers take participation in the sketch generation event conducted by Amazon Mechanical Turk (AMT) and was generated with aim of hand-drawn sketch recognition and classification purposes. The human recognition performance on this dataset is 73.1%. The size of each sketch image is 1111x1111. Few of sketch image samples from different object classes are shown in Fig. 7.

B. Natural Images for Augmented-Variants

To get the color images for augmented variants, the natural photos are collected from the publically available full-colored image dataset and most of them collected from the web. These dataset covers some of the object categories of TU-Berlin dataset. These images were collected from Caltech 256 image dataset [52], while the remaining images were taken from the web and generates the augmented variants, i.e., enhanced edges and edge maps of natural images. Finally, Total 31,500 colored images, including other augmented-variants, are collected. These images are then added to the corresponding sketch object categories of TU-Berlin sketch dataset for training the proposed framework. Some of the sample images corresponding to sketch images are shown in Fig. 8. The evaluations are carried out with 3-fold cross validation to allow comparisons with baselines.

C. Environment Setting

The proposed framework is implemented in open source keras using python libraries. It consists of three different pre-trained deep CNN architectures which are trained separately to extract the features from the augmented-variants sketch dataset. The overall training is conducted on NVIDIA dual Xeon processor with 13GB RAM and GPU cards. Ubuntu 16.04 operating system with the 64bit environment is used to perform training operations.

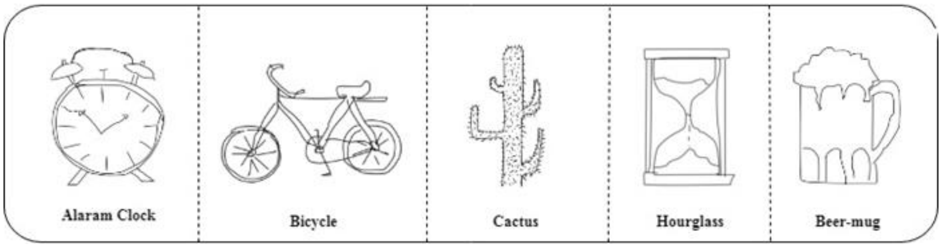


Fig. 7. TU-Berlin Sample Sketch Images.

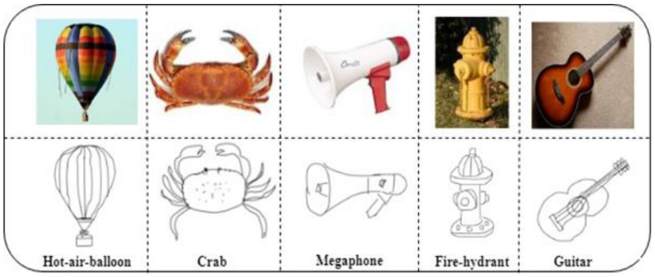


Fig. 8. Sample Colored Natural Images with Corresponding Sketch Image.

D. Results and Evaluation

In order to validate the performance of the proposed CNN-based framework for sketch recognition, the sketch classification accuracies are shown in Table I. According to tabulated outcomes, the performance of proposed individual CNN architectures, i.e., VGGNet, ResNet, and Inception-v3 obtained 78.93%, 89.61%, and 91.89% classification accuracy, respectively. However, the proposed framework achieves better performance with 94.57% sketch classification accuracy on augmented-variants TU-Berlin sketch dataset and beat the individuals, i.e., VGGNet, ResNet and Inception-v3 in performance gap with 15.64%, 4.96%, and 2.68%, respectively. It is evident from the tabulated results that proposed method outperforms in terms of sketch classification than the performance of other three individual architectures.

In this subsection, the performance of proposed framework is compared with state-of-the-art methods including sketch-based handcrafted features and deep features methods. The overall accuracy is shown in terms of percentage to demonstrate the results. Fig. 9 shows the hand-crafted features recognition performances on sketch images. The proposed method outperforms on HOG-SVM (recognition accuracy 56.0%), sketch recognition accuracy achieved through Ensemble method, Multi-kernel-SVM and Fisher Vector-SP were 61.5%, 65.8%, and 68.9%, respectively in the literature study.

HOG based features with SVM classifier has the lowest recognition rate. However, the best performance accuracy based on hand-crafted features is 68.9%, which is less than the human recognition accuracy (73.1%) on sketch data. The reason of lower performance of handcrafted features is that; mostly these methods have been designed to extract features from real photos and not suitable to covers the high variability of abstractions and appearance in sketch images.

Similarly, the results for deep features methods are summarized in Table II. It shows that deep networks perform better than hand-crafted features. Human recognition level accuracy on TU-Berlin dataset was first beats through deep network architecture [5] and enhance recognition rate with 1.8% higher than human recognition performance. Moreover, the performance gap grows from 1.8% to 4.2% when [43] implemented deep sketch model by mixing sketch images with colored images for sketch recognition. The sketch recognition accuracy 79.1% achieved by [8] using pre-trained VGGNet architecture through transfer learning approach.

Different reasons might cause variations in sketch recognition results while using DNNs model. It depends on the structure of the model, the depth of network architecture, different methods used for feature extractions, and tuning-up various parameters. Even though, to check the effect of every parameter for any model performance is also arduous and tedious task.

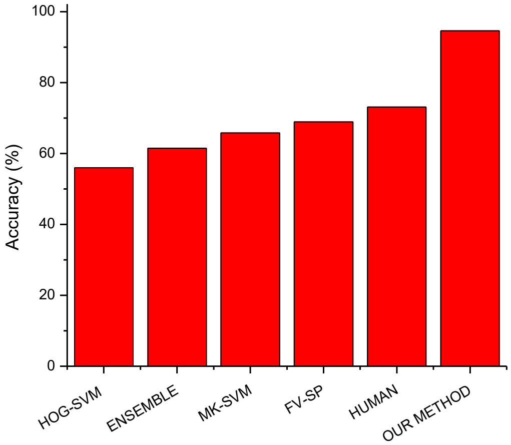


Fig. 9. Sketch Classification Accuracy based-on Hand-Crafted Features.

TABLE. I. COMPARATIVE SKETCH CLASSIFICATION ACCURACY OF THE PROPOSED FRAMEWORK WITH OTHER INDIVIDUAL DCNNS ARCHITECTURES

Index	Method	Accuracy
1	VGGNet	78.93%
2	ResNet	89.61%
3	Inception-v3	91.89%
4	**Proposed Method with GAP**	**94.57%**

Therefore, some conclusions can be made from the baseline results mentioned in the Fig. 9 and Table II. First of all, the proposed CNNs-based framework consistently outperforms in sketch classification on both handcrafted features methods and the sketch features analyzed through deep neural network models. This can show that the use of natural images plays a significant role in the evaluation of sketch images. Additionally, various augmentation-variants, specifically edge maps strengthen the proposed framework capability in sketch recognition accuracy. Secondly, the performance of individual deep CNNs becomes improved when it goes deeper. The best recognition performance through individual deep CNN architecture is achieved by Inception-v3. Thirdly, in a case of transfer learning approach, the distinct features of three deep CNNs architectures were combined, and the recognition performance is improved by employing GAP strategy. This outcome substantiates the experiments for augmented-variants TU-Berlin sketch dataset. In the proposed framework, it is declared that using transfer learning with GAP increases sketch recognition accuracy. This is also applicable for combining distinctly extracted features from multiple CNNs architectures as compared to individual CNN architecture.

TABLE. II. SKETCH CLASSIFICATION ACCURACY COMPARISON BASED ON DEEP FEATURES

Index	Methods	Accuracy
1	AlexNet-SVM [16]	67.1%
2	AlexNet-Sketch [16]	68.6%
3	LeNet [22]	55.2%
4	Human [21]	73.1%
5	Sketch-a-Net [5]	74.9%
6	Deepsketch [43]	77.3%
7	VGG-based Transfer Learning [8]	79.1%
8	**Proposed Method with GAP**	**94.57%**

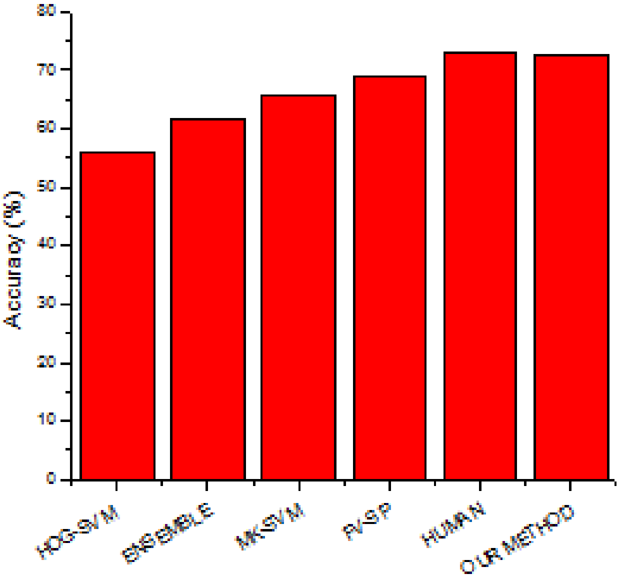


Fig. 10. Sketch Classification Results Comparison on TU-Berlin (without Augmented Variants).

On the other hand, we evaluate our proposed method on TU-Berlin sketch dataset (without augmented-variants), the experimental results are illustrated in Fig. 10. The proposed framework achieves a competitive performance of 72.82% classification accuracy as compared with 73.1% human recognition accuracy except those sketch based CNNs architectures [5, 7], which have been specifically designed for sketch classification.

E. Further Evaluation for Retrieval Task

The performance of the proposed deep framework is further evaluated on sketch retrieval task. For this test, the proposed deep CNNs-based frame work is used to extract features from both the sketch images and natural images. All the images are indexed with concerned features. For retrieval task, proposed framework is used to extract the features from the edge maps and query sketches separately and compared with all the retrieval candidate images in the database. Euclidean distance is computed to make the comparison between the query sketch images and the images in the database. The query images are randomly selected to retrieve the similar images from the image database.

The sketch object-based queries and top-9 retrieval results are shown in Fig. 11. The retrieval results are ranked with scored values. The lower scored value represents the higher rank similarity between the query sketch and the retrieved image. In most of the cases, the query images retrieved the most similar candidate images which show the enough discriminative features for retrieval task.

These images (i.e., teapot, beer-mug, guitar, etc.) are retrieved with high rank similarity. Interestingly, the retrieved image had very comparable edge maps which make the retrieval task in high ranks. Moreover, in some cases the system failed to retrieve the right candidate images, the reason might be the natural images having complex background leading the large difference between the sketch images and edge-maps.

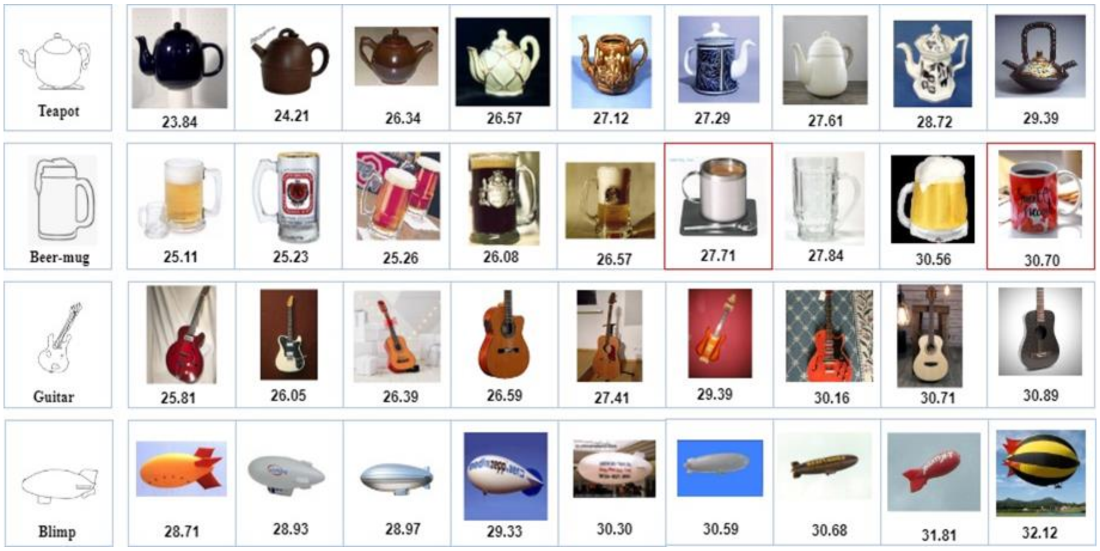


Fig. 11. Retrieval Performance of our SBIR, Top-9 Retrieval Outcomes of Four Sketch Queries by Proposed CNNs-based Framework. The First Column Indicates the Query Sketches, and the Most Similar Retrieved Candidate Images According to Ranked-Score from Four Different Object Categories are Shown From the Top to the Bottom-Row in the Sequence. Red-Box Images Indicate Incorrect Retrieved Images.

The capabilities of the proposed framework are extensively evaluated on a retrieval task. The lower rank similarity retrieval results with the corresponding scored values are illustrated in Fig. 12. Additionally, the retrieval performance of the proposed framework was also tested on sketches and natural images which was not a part of training or validation data. It is interesting to demonstrate that the query sketches retrieve the nearest candidate images with less ambiguity. However, the performance can be further improved by providing enough training data instances and also by reducing background complexity of the natural images. Results illustrated in Fig. 13 advocate that proposed CNN-based framework is capable to perform well on the variety of images which was not a part of training or validation data. These experiments validate the effectiveness of the proposed framework for sketch recognition.

F. Experimental Analysis

The proposed CNNs-based framework outperforms all baselines and achieves better performance on TU-Berlin (augmented-variants) sketch dataset which shows that augmented-variants are beneficial for sketch recognition. It is worth mentioning that handcrafted features performance on this dataset is worse than deep feature methods. In our case, the complex background of the natural images and generated edge maps from those images, make the proposed method more challenging and competitive. The proposed method performed better and retrieved the candidate images mostly in high ranked score of similarity. This demonstrate that proposed framework based on transfer learning with GAP is capable to extract the most discriminative features from both type of images i.e., sketch and natural images and could help to strengthen retrieval performance. But Fig. 12 and Fig. 13 represent the retrieval results of lower rank similarity and the results for images instead of using training and validation images respectively, where the incorrect retrieved images are outlined in red-boxes. Therefore, it is stated that the lower rank similarity performance and incorrect retrieved images might be the reasons of providing not enough training samples to the proposed framework as well as it might be not well-aligned with complex background of candidate retrieval images.

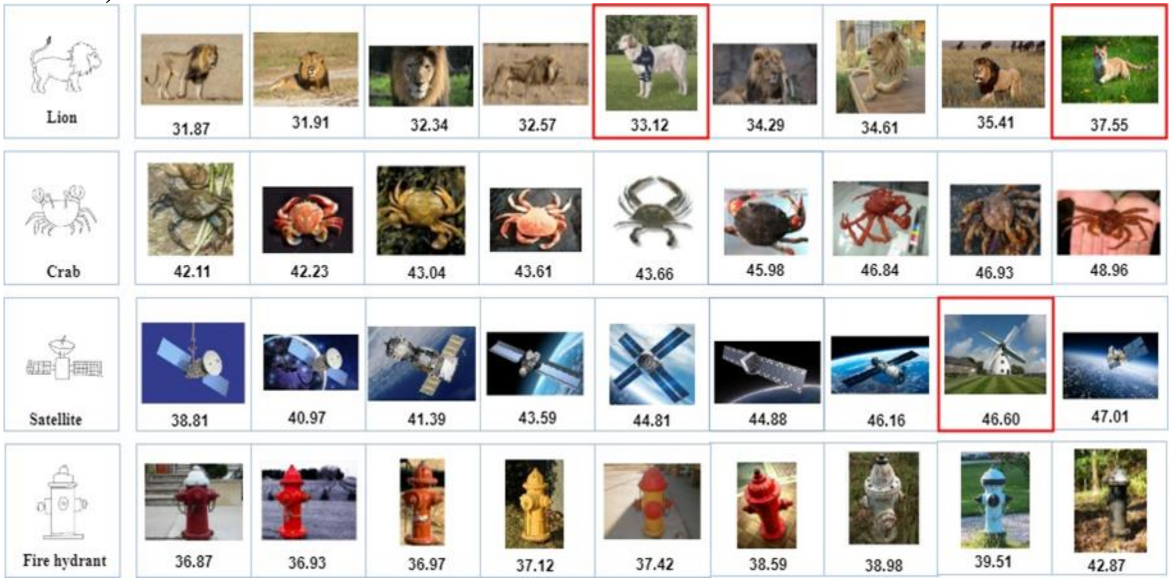


Fig. 12. Low-Rank Similarity Retrieval Performance of SBIR. The Red-Outlined Boxes Indicate Incorrect Retrieved Candidate Objects.

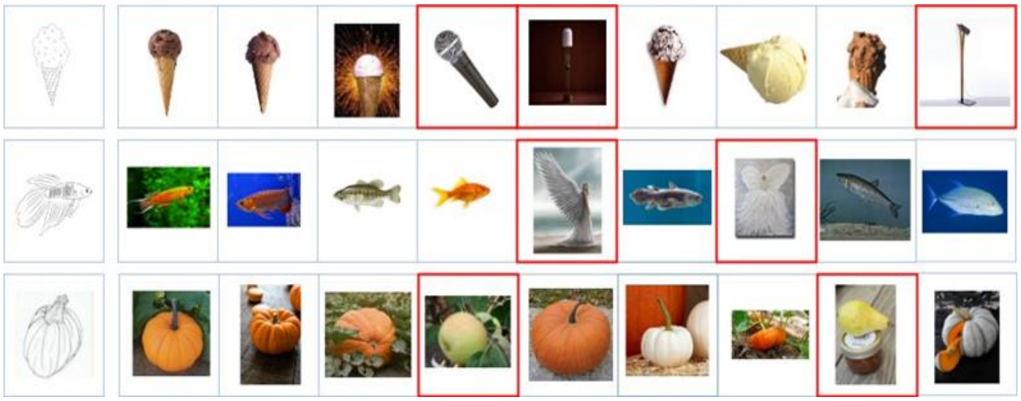


Fig. 13. Retrieval Performance based on Sketch-Query for other Object Categories.

V. Conclusions

In this manuscript, a deep CNN-based framework for sketch recognition via transfer learning with global average pooling strategy was proposed. The three different well-known pre-trained DCNNs models are analyzed in the proposed framework and fine-tuned on both with and without augmented-variants TU-Berlin sketch dataset. The sketch classification performance of the proposed framework was compared with different CNN architectures individually and also compared with other state-of-the-art methods. Considering the individual CNN architectures, it is observed from the experimental results based on augmented variants TU-Berlin sketch dataset that; the Inception-v3 showed higher accuracy than other two, i.e., ResNet and VGGNet architectures. However, VGGNet showed lowest classification accuracy among the other individual CNN architectures. The proposed framework outperformed the other existing methods in terms of sketch classification and retrieval task. The utilization of GAP not only reduces feature dimensionality but, also enhances the classification accuracy over the augmented-variants TU-Berlin sketch dataset. On the other hand, the proposed framework provides a competitive result as compare with human recognition accuracy on (without augmented-variants) TU-Berlin sketch dataset. In the future, deep learning approaches would be adopted for 3D shapes object retrieval task.

References

[1] D. Dixon, M. Prasad, and T. Hammond, “iCanDraw: using sketch recognition and corrective feedback to assist a user in drawing human faces,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA, 2010, pp. 897-906.

[2] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view Convolutional Neural Networks for 3D Shape Recognition." pp. 945-953.

[3] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep Human Parsing with Active Template Regression," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2015arXiv150302391L, [March 01, 2015, 2015].

[4] M. Eitz, R. Richter, K. Hildebrand, T. Boubekeur, and M. Alexa, “Photosketcher: Interactive Sketch-Based Image Synthesis,” IEEE Computer Graphics and Applications, vol. 31, no. 6, pp. 56-66, 2011.

[5] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales, "Sketch-a-Net that Beats Humans," arXiv e-prints, https://ui.adsabs.harvard.edu /abs/2015arXiv150107873Y, [January 01, 2015, 2015].

[6] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, “Sketch-a-Net: A Deep Neural Network that Beats Humans,” International Journal of Computer Vision, vol. 122, no. 3, pp. 411-425, May 01, 2017.

[7] O. Seddati, S. Dupont, and S. Mahmoudi, “DeepSketch 3,” Multimedia Tools and Applications, vol. 76, no. 21, pp. 22333-22359, November 01, 2017.

[8] J. Ahmad, K. Muhammad, and S. W. Baik, “Data augmentation-assisted deep learning of hand-drawn partially colored sketches for visual search,” PloS one, vol. 12, no. 8, pp. e0183838, 2017.

[9] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, November 01, 2004.

[10] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection." pp. 886-893 vol. 1.

[11] R. Hu, S. James, T. Wang, and J. Collomosse, “Markov random fields for sketch based video retrieval,” in Proceedings of the 3rd ACM conference on International conference on multimedia retrieval, Dallas, Texas, USA, 2013, pp. 279-286.

[12] R. G. Schneider, and T. Tuytelaars, “Sketch classification and classification-driven analysis using fisher vectors,” ACM Transactions on Graphics (TOG), vol. 33, no. 6, pp. 174, 2014.

[13] E. Shechtman, and M. Irani, "Matching Local Self-Similarities across Images and Videos." pp. 1-8.

[14] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, “A descriptor for large scale image retrieval based on sketched feature lines,” in Proceedings of the 6th Eurographics Symposium on Sketch-Based Interfaces and Modeling, New Orleans, Louisiana, 2009, pp. 29-36.

[15] R. Kiran Sarvadevabhatla, and R. Venkatesh Babu, "Freehand Sketch Recognition Using Deep Features," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2015arXiv150200254K, [February 01, 2015, 2015].

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." pp. 1097-1105.

[17] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2014arXiv1409.1556S, [September 01, 2014, 2014].

[18] E. Boyaci, and M. Sert, "Feature-level fusion of deep convolutional neural networks for sketch recognition on smartphones." pp. 466-467.

[19] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via siamese convolutional neural network." pp. 2460-2464.

[20] Y. Yang, and T. Hospedales, Deep Neural Networks for Sketch Recognition, 2015.

[21] M. Eitz, J. Hays, and M. Alexa, “How do humans sketch objects?,” ACM Trans. Graph., vol. 31, no. 4, pp. 1-10, 2012.

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, 1998.

[23] J. Bai, M. Wang, and D. Kong, “Deep Common Semantic Space Embedding for Sketch-Based 3D Model Retrieval,” Entropy, vol. 21, no. 4, pp. 369, 2019.

[24] N. Upadhyaya, and M. Dixit, A Review: Relating Low Level Features to High Level Semantics in CBIR, 2016.

[25] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, “Sketch-based image retrieval: Benchmark and bag-of-features descriptors,” IEEE transactions on visualization and computer graphics, vol. 17, no. 11, pp. 1624-1636, 2010.

[26] J. Yang, S. Li, and W. Xu, “Active Learning for Visual Image Classification Method Based on Transfer Learning,” IEEE Access, vol. 6, pp. 187-198, 2018.

[27] J. T. Zhou, S. J. Pan, and I. W. Tsang, “A deep learning framework for Hybrid Heterogeneous Transfer Learning,” Artificial Intelligence, 2019/06/06/, 2019.

[28] L. Zhang, D. Wang, C. Bao, Y. Wang, and K. Xu, “Large-Scale Whale-Call Classification by Transfer Learning on Multi-Scale Waveforms and Time-Frequency Features,” Applied Sciences, vol. 9, no. 5, pp. 1020, 2019.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. B. Wojna, Rethinking the Inception Architecture for Computer Vision, 2016.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/ 2015arXiv151203385H, [December 01, 2015, 2015].

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database." pp. 248-255.

[32] M. Rahim, N. Othman, and Z. Jupri, “A Comparative Study on Extraction and Recognition Method of CAD Data from CAD Drawings, Information Management and Engineering,” ICIME, vol. 9, pp. 709-713.

[33] C. L. Zitnick, and D. Parikh, "Bringing semantics into focus using visual abstraction." pp. 3009-3016.

[34] P. Sousa, and M. J. Fonseca, “Geometric matching for clip-art drawing retrieval,” Journal of Visual Communication and Image Representation, vol. 20, no. 2, pp. 71-83, 2009.

[35] A. McCallum, and K. Nigam, "A comparison of event models for naive bayes text classification." pp. 41-48.

[36] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features." pp. 137-142.

[37] Y. Li, Y.-Z. Song, and S. Gong, "Sketch Recognition by Ensemble Matching of Structured Features." p. 2.

[38] X. Cao, H. Zhang, S. Liu, X. Guo, and L. Lin, "Sym-fish: A symmetry-aware flip invariant sketch histogram shape descriptor." pp. 313-320.

[39] Y. Li, T. M. Hospedales, Y.-Z. Song, and S. Gong, "Free-hand sketch recognition by multi-kernel feature learning," Computer Vision and Image Understanding, vol. 137, pp. 1-11, 2015.

[40] N. Mboga, S. Georganos, T. Grippa, M. Lennert, S. Vanhuysse, and E. Wolff, "Fully Convolutional Networks and Geographic Object-Based Image Analysis for the Classification of VHR Imagery," Remote Sensing, vol. 11, no. 5, pp. 597, 2019.

[41] D. Mao, and Z. Hao, "A Novel Sketch-Based Three-Dimensional Shape Retrieval Method Using Multi-View Convolutional Neural Network," Symmetry, vol. 11, no. 5, pp. 703, 2019.

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2014arXiv1409.4842S, [September 01, 2014, 2014].

[43] X. Wang, X. Duan, and X. Bai, "Deep sketch feature for cross-domain image retrieval," Neurocomput., vol. 207, no. C, pp. 387-397, 2016.

[44] "https://keras.io/applications/."

[45] P. Perona, and J. Malik, "Scale-space and edge detection using anisotropic diffusion," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 7, pp. 629-639, 1990.

[46] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep Pyramidal Residual Networks for Spectral–Spatial Hyperspectral Image Classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 2, pp. 740-754, 2019.

[47] J. Kim, B. Kim, P. P. Roy, and D. Jeong, "Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure," IEEE Access, vol. 7, pp. 41273-41285, 2019.

[48] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," Journal of Big Data, vol. 3, no. 1, pp. 9, May 28, 2016.

[49] S. J. Pan, and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, 2010.

[50] 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, December 01, 2015.

[51] M. Lin, Q. Chen, and S. Yan, "Network In Network," arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2013arXiv1312.4400L, [December 01, 2013, 2013].

[52] "https://www.kaggle.com/jessicali9530/caltech256 ".

A Distributed Approach based on Transition Graph for Resolving Multimodal Urban Transportation Problem

Mohamed El Moufid[1], Younes Nadir[2], Khalid Boukhdir[3], Siham Benhadou[4], Hicham Medromi[5]
LRI, National High School of Electricity and Mechanics, Casablanca, Morocco
Foundation of Research, Development and Innovation in Sciences and Engineering, Casablanca, Morocco

Abstract—**All over the world, many research studies focus on developing and enhancing real-time communications between various transport stakeholders in urban environments. Such motivation can be justified by the growing importance of pollution caused by the transport sector in urban areas. In this work, we propose an approach of assistance for displacement in urban environment taking advantages of multimodal urban transportation means, where several modes of public transports are available. In addition, we also consider the possibility of using both private modes of transport and cities parking. The proposed distributed approach described in this paper is based on an abstraction of a city multimodal graph according to the available modes of public transport and road traffic and transition graph approach to move from a mode to the other mode. Numerical results are developed to justify the effectiveness of our approach.**

Keywords—Multimodal transport; distributed approach; transition graph

I. Introduction

Displacements in urban environments are one of the major problems facing most supercities around the world. In addition to the pollution and time waste generated by traffic jams expectations while looking-up for parking spots to park private vehicles [1], urban mobility favors certain health disturbances among cities citizens. In recent year authorities have encouraged the use of a diversity of transport modes in order to reduce the impact of this problem [2].

Several solutions have then been developed to help users to choose the optimal path to follow to reach their destinations. Early solutions concerned each a particular mode of public transport and each operator proposed a system that manages its own network. Subsequently, other solutions made it possible to propose the best optimal paths by combining two or more modes of public transports [3] [4].

Currently, researches are moving towards the proposal of an urban mobility management system by considering both all available cities public modes of transports, traffic road states and the availability of city car parks. Thus, the optimal route for a user would be to drive his private vehicle to a parking lot, then take a series of public transport lines.

In order to be able to ally with the development of the city's urban transport infrastructure, urban travel assistance services must be able to consider the growing complexity of public transport networks, road traffic density and daily requests for parking space.

Several studies examine the possibility of establishing both multimodal transport networks models, road states and the availability of car parks in order to propose approaches that calculate and determine the most appropriate paths for a user’s request. These approaches must be robust enough to keep their efficiency and effectiveness as the complexity of the multimodal network and the amount of data to be processed increases.

The major issues in the development of these approaches can then be considered in two aspects:

1) Management of heterogeneous data: Different modes correspond to a set of navigation datasets which are acquired, stored and managed by different public or private organizations. Likewise, car parks management systems make it possible to retrieve data concerning the availability of car parks at a given moment and to predict the occupancy rate for a later date. By the same token, road traffic supervision systems provide real time data not only from traffic roads states but also a full history of occupancy rate related to road traffic arcs [5].

2) Multimodal network and optimal path finding approach: Different types of data should be considered in determining the optimal path in a multimodal network. Various works then seek to elaborate the problem modeling approach which allows an adequate structuration and an easy accessibility to the information which is necessary for different computations [6]. Likewise, the resolution approach must also be sufficiently powerful and flexible to be able to maintain its performance when increasing the size of the multimodal network [7].

In the present paper we are going to present a new model of the multimodal network based on the graph theory and a parallel distributed approach to propose optimal paths between two points in a city.

The rest of the paper will be developed as follows: in Section II we present definitions of the concepts necessary for the description of the problem and the proposed solution. These definitions concern the modeling of the multimodal transport network using graph theory. Then we proceed in the third part to the mathematical modeling of the problems in order to propose the approaches of resolution allowing at first to reduce the complexity of the problems and then to propose the optimal solutions. We conclude our work with a conclusion and perspectives.

II. Problem Formulation

In this section, we present some necessary definitions for the formulation of the optimal path problem in a multimodal transport network using graph theory approach. The particularity of the problem is that the multimodal transport network contains several subnetworks, each subnetwork refers to either a particular mode of transport or the road traffic network of the city.

The model should propose an optimal structuring of the data in such a way that the system could have rapid access to each arc at any subnetwork of the multimodal network. Then it would be able to react in real time on following any change that may happen on an arc of the multimodal network. Different models have been proposed to consider the dynamic particularity mentioned above, generally based on graph theory: hypergraph theory [8] [9] [10], hierarchical graph [11], colored graph approach [12] and transfer graph approach [13]. However, most models fail to consider the following three components of a city's multimodal network simultaneously: public transport modes, road traffic network and city car parks [14].

In our study, we consider a global multimodal network composed of: a set of monomodal public transport networks, a set of parking spaces and a traffic road network. Each component of the multimodal network has its own specific parameters.

So, in regard to the resolution approach, the procedure of calculating the shortest path should be able to keep its efficiency and computing power when changing the size of the multimodal network; In most studies, the optimal path resolution approach in a multimodal network is considered as a whole. Thus, to perform calculations, the approach must traverse all nodes and arcs of the said network, which slows down the computation process and reduces the system performances.

Wherefore, in the approach, the calculations are done in a parallel and distributed way for each component of the global network. Thus, to calculate the optimal path between two points of a multimodal network, our approach consists on making intermediate calculations for each sub graph before considering the whole multimodal network. This need to reduce the size of the graph is justified by the complexity of the problem. Thus, with a graph of n nodes, e edges and m modes of transport, the corresponding graph generated will have a complexity in at least of $O(ne^3)$ [15] [16].

In the following, we define some concepts that will allow to present our model and our approach of resolution:

Let $G = (N, E, M, P)$ denotes a multimodal transport graph, where $N_j = \{n_{j1}, \dots, n_{jl}\}$ is a set of nodes belonging to the mode j, $\mathrm{M} = \{m_1, \dots, m_k\}$ is a set of modes, $P = \{p_1, \dots, p_s\}$ is a set of packing spaces.

An edge defined by $\mathrm{e_i} = (\mathrm{n_{jq}}, \mathrm{n_{jr}})_{\mathrm{m_j}}$ expresses that it is possible to go from node $\mathrm{n_{jq}}$ to node $\mathrm{n_{jr}}$ using transport mode $\mathrm{m_j}$. A value ${f^r}_{e_i}(t_k)$ is associated to each edge e_i indicating the cost of the edge e_i at departure time t_k according to the criterion r (e.g., distance or duration).

A. Définition1: Multimodal Path

Given a multimodal transport graph $G = (N, E, M, P)$. A multimodal path $p_{n_{io},n_{jd}}$ is a sequence of possible edges to go from the node n_{io} to the node n_{jd}, $p_{n_{io},n_{jd}} = \left((n_i, n_2)_{m_o}, (n_2, n_3)_{m_p}, \dots, (n_{j-1}, n_j)_{m_d}\right)$.

B. Définition2: Cost Function

The vector-valued function: $f_r(p, t_0,)$: $P * T \rightarrow R$ represents the cost of the path p departing at time t_0 according to the criteria r. R is a set of vectors, where each vector represents a criterion.

C. Definition 3: Optimal Path Problem

Given a graph $G = (N, E, M, P)$, the optimal path problem according to the criteria r consists in calculating a path $p_{n_{io},n_{jd}}$ from node o to node d departing at t_0 where $f_r(p, t_0)$ is minimal. This is called the optimal path (OP).

D. Transition Graph

Given a multimodal graph $G = (N, E, M, P)$, the transition graph is defined as $T_g = (C, T_r)$ where $C = \{C_1, C_2, \dots, C_k\}$ is the set of monomodal graphs, and T_r is the set of virtual transition edges which interconnect them. Each component $C_i = (N_i, E_i, M_i, P_i)$ is such that $N = \bigcup_{i \in \{1,\dots,k\}} N_i$, $E = \bigcup_{i \in \{1,\dots,k\}} E_i$, $P = \bigcup_{i \in \{1,\dots,k\}} P_i$.

The transition graph model consists in abstracting different modes of transport on the same map; we can distinguish in a transfer graph T_g between two groups of paths: intra-components paths and inter-components paths. An inter-component path refers to a path that connects two nodes belonging to distinct modes of transport [17], while an intra-component path defines a path that links two nodes belonging to the same mode of transport

Figure 1 illustrates an example of a transfer graph, where Mode 1, Mode 2 and Mode 3 represent three modes of public transport and Road Graph represent the road network. distinguishes between two categories of monomodal graphs can be described as follow:

- Mode Graph represent public modes (Mode 1, Mode 2, Mode 3): nodes represent stations of the mode and edges represent the paths of the transport mode (Tramway, bus, subway ...). Parameters of the graphs (departure times at a node, estimated duration to travel an arc ...) depend on the mode of transport and are given by the public transport operator. These parameters can be updated by a user or an operator once a disturbance occurs on a line of the mode.

- Road Graph represents the network of the road traffic: local nodes represent the intersections of the paths and transfer nodes represent car parks.

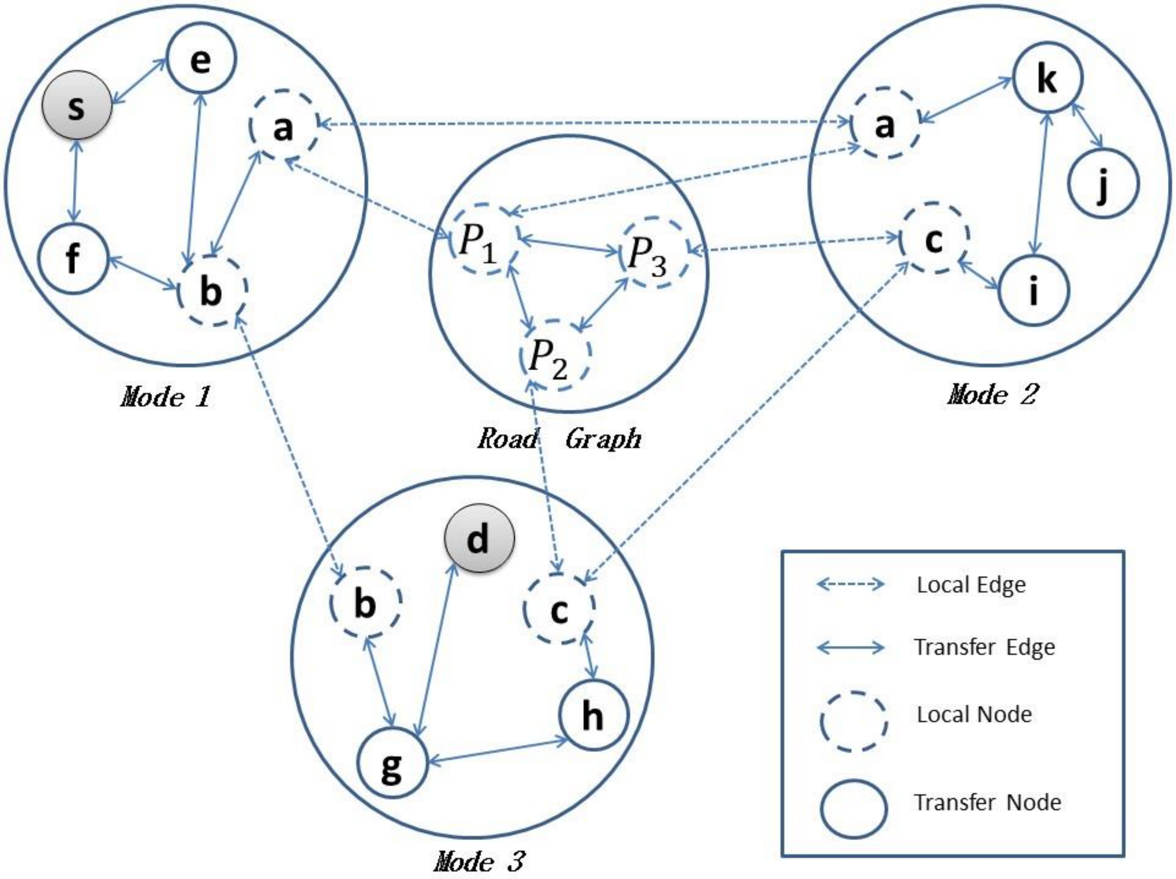


Fig. 1. Transition Graph Illustrative Case.

A citizen can drive his own car from an initial point to a second point, then park his car in a parking lot and finish his travel with a public mode of transport. Different parameters related to the state of an arc of a road network are updated by users who browse it in real time.

In the current study, transfer node represents walking movements either to go from a parking space to a public transport mode (if the latest mode of transport is a private vehicle), or to go from a station of a mode of transport to a station of another mode of transport.

III. Problem Resolution and Results

The objective function is a multi-objective function. Indeed, by varying the multimodal path between two nodes of the multimodal graph, various parameters can vary (duration, cost, level of comfort ...).

Let $p_{n_{io},n_{jd}}$ be a multimodal path from node n_{io} to node n_{jd}. The optimal path between a node i belonging to a mode m to a node j belonging to a mode n according to the criteria k can be defined as follow:

$$F_k = \min \sum_{i,j,m,n} x_{i_m j_n} * f_k^{t_0}\left(x_{i_m j_n}\right) \quad (1)$$

Where:

$x_{i_m j_n}$: logical variable defines whether the $e_{i_m j_n}$ edge linking a node i belonging to a mode m to a node j belonging to a mode n is used or not.

$f_k^{t_0}\left(x_{i_m j_n}\right)$: cost function of the direct arc linking a node i belonging to a mode m to a node j belonging to a mode n departing at time t_0.

The optimal global function should consider all possible criterions and proposes the set of optimal paths according to all criterions.

In our study, for problem simplification reasons, we consider that the optimal global function is a direct weighting of optimal paths according to each criterion [18].

The minimization of the global function can be written as follow:

$$F = \min \sum_k w_k * F_k \quad (2)$$

Where w_k are weighting coefficients and refers to a user's preference according to the criteria k.

The optimization of the global function is subject to the following constraints:

$$\sum_{j_n} x_{i_m j_n} - \sum_{i_m} x_{i_m j_n} = \begin{cases} -1 & if \quad j = o \\ 0 & \forall j \in V,\ j \neq o, d \\ 1 & if \quad i = d \end{cases} \tag{3}$$

$$\sum_{m,n} x_{i_m j_n} \leq 1, \quad \forall (i,j) \in E, \quad (m,n) \in M \tag{4}$$

$$\sum_{l} x_{l_v i_v} \geq x_{i_v j_v} \quad \forall (i_v, j_v) \in M_v \tag{5}$$

Where v refers to a private mode.

The equation (3) ensures that each node of the proposed paths is visited once at maximum.

The constraint (4) ensures that each arc is traveled by one mode of transport at maximum and the constraints (5) ensures that if a user leaves his car in a city's parking lot, the rest of the proposed path must not propose a return to his private vehicle.

The problem of the optimal path according to r criterions in a multimodal transport network composed of n nodes and e is a multiobjective function with complexity of $\mathcal{O}(\text{rn}e^3)$ [19]. In the literature, the common difficulty resides in maintaining performance of the approach when the considered network becomes larger and more complex [20] [21].

In the proposed work, we propose an approach allowing to reduce the complexity of the problem based particularly on The Depths-First-Search algorithm described in Algorithm 2 which is at a difficulty of $\mathcal{O}(|V| + |E|)$. The approach is based on the following ideas:

1) Reduce the size of the whole multimodal graph and assign for each mode of public transport/ road traffic network the corresponding sub-graph.

2) Look for the possible paths between the nodes and then evaluate them according to each desired criterion (duration, cost…), which makes it possible to go from a complexity of $\mathcal{O}(\text{rn}e^3)$ to a $\mathcal{O}(|V| + |E|)$ problem complexity.

3) Evaluate the possible paths according to each classified parameters of the user then assign weighting coefficients. this is the fastest approach for multi-objective evaluation [16].

4) In order not to repeat similar calculations we propose at first to check if the system has already responded to a similar request. If the case, it exploits them. Otherwise, it proceeds normally to the calculations.

In order to validate the performance of our proposed approach, and evaluate the win of each of the proposed ideas, we performed a number of experimental tests. We test and compare our approach in terms of execution time according to the size of the network considered.

We consider 10 main networks varying from 50 nodes to 10 000 nodes. We consider three main approaches:

1) The whole graph is considered, and the Algorithm 1 is considered from the step 3.

2) The transition graph is considered, but the second step is skipped in order to do all necessary calculations.

3) The whole algorithm is considered and previous requests and proposed multimodal paths are taken into consideration in order to reduce calculations.

Thus, following a user's request, an optimal path should be calculated as described in Algorithm 1.

Algorithm 1: Resolution Approach Algorithm

```
Require: (Multimodal-Graph, Start-Node, Destination-Node,
          Levels of preferences)

1. Create the simplified transition associated graph based
   on the multimodal-Graph.
2. Search in the Database for similar queries.
3. Determine possible multimodal paths from the starting
   point to the desired destination.
   The used method for in this step is defined in Algorithm
   2.
4. Evaluate possible multimodal paths according to the
   levels of preferences defined by the user.
5. Classify the set of evaluated paths according to user
   preferences.
6. Propose the optimal paths to user request.
7. Validate the chosen path and save the experience.
```

The second step of the Algorithm 1 is based on the DFS approach defined in Algorithm 2

Algorithm 2:Adapted DFS Algorithm

```
Require: (Transition-Graph, Start-Node, Destination-Node)
Put anyone of the graph's vertices on top of a stack.
Repeat {
    While stack is not empty {
            Take the top item of the stack and add it to the
            visited list.
            Create a list of that vertex's adjacent nodes. Add
            the ones which aren't in the visited list to the top
            of stack.}
    Until stack is empty}
```

Results of simulations of the programs under MATLAB student-use software using a configuration computer: Intel Core i3-370M, 2.40GHz and RAM 4 GB, show at Figure 2 that from a network of 1000 nodes, the last two approaches show a high efficiency compared to the first approach. Indeed, in the second approach, the global graph is abstracted into sub-graphs, this distributed approach makes it then possible to simplify the problem and consequently to reduce the execution time.

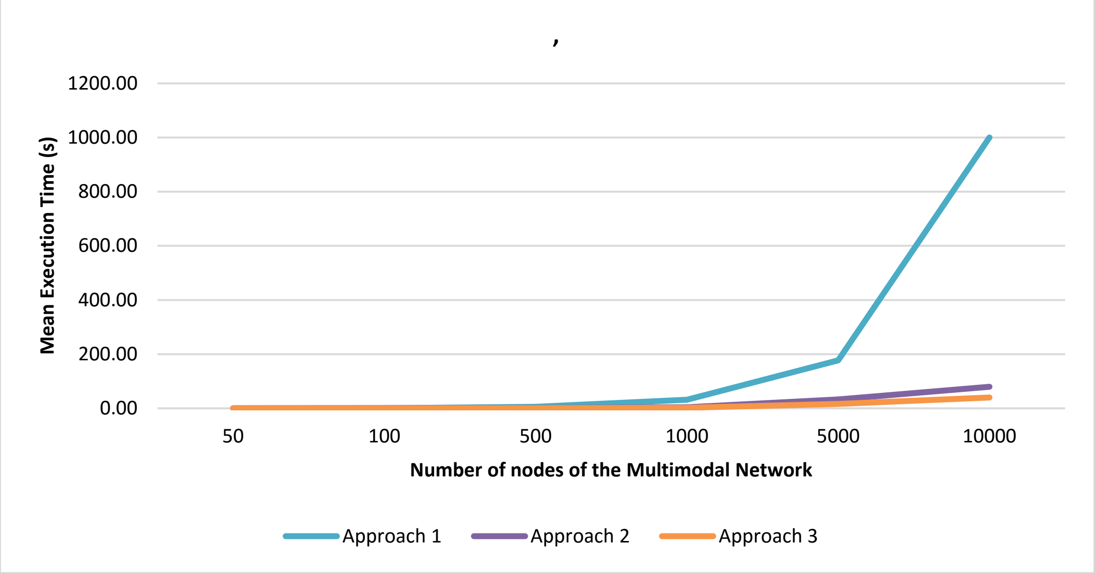


Fig. 2. Results of Comparisons.

The third approach, in addition to the advantages of the distributed aspect of its execution, it makes it possible to benefit from the previous experiences recorded in the database of the system. It consequently eliminates some calculation operations and subsequently reduces the execution time.

IV. Conclusion and Perspectives

The proposed work is useful to find the optimal multimodal paths in a multimodal network of transportation composed of a set of monomodal public transportation networks (Tramway, train, metro…), of road traffic network and a set of available parking in a city.

The global multimodal network is abstracted into sub-graphs where each graph refers either to a monomodal public transportation network or a traffic road network. The distribution of the problem makes it easier to solve, and allows acting on a sub-graph without affecting other monomodal sub-graphs.

The proposed resolution approach takes advantage of the simplification of the problem obtained by its abstraction and shows good results compared to the case where the whole network is considered in calculations.

In the third approach, artificial intelligence is integrated to identify whether the system has already responded to a similar request. In this, calculations aren't to be redone which impact positively the execution time and improve the efficiency of the system.

As perspective, different possible approaches for artificial intelligence could be compared in order to identify similarities and choose the most appropriate for the problem.

References

[1] T. Yogesh and P. Mr.M.D, "Advance Prediction of Parking Space Availability and other facilities for Car parks in Smart Cities," International Research Journal of Engineering and Technology (IRJET) , vol. 3, no. 5, pp. 2225-2228, 2016.

[2] W. R.C.P, S. W.Y., Y. Linchuan, L. Y.C. and W. S.C., "Public transport policy measures for improving elderly mobility," Transport Policy, vol. 63, pp. 73-79, 2018.

[3] L. David and L. Angélica, "Techniques in Multimodal Shortest Path in Public Transport Systems," Transportation Research Procedia, vol. 3, pp. 886-894, 2014.

[4] C. Lin, C. Peng and P. Jianping, "Delay Minimization for Data Dissemination in Large-Scale VANETs with Buses and Taxis," IEEE Transactions on Mobile Computing, vol. 15, no. 8, pp. 1939 - 1950, 2016.

[5] E. F. Nour-Eddin, L. Henry and k. Ajeesh, "Data fusion in intelligent transportation systems: Progress and challenges – A survey," Information Fusion, vol. 12, no. 1, pp. 4-10, 2011.

[6] L. Chao-Lin, P. Tun-Wen, C. Chun-Tien and H. Chang-Ming, "Path-planning algorithms for public transportation systems," in IEEE Intelligent Transportation Systems. P, Oakland, CA, 2001.

[7] L. Liu, "Data Model and Algorithms for Multimodal Route Planning," München, 2010.

[8] S. Pallottino and S. Nguyen, "Equilibrium traffic assignment for large scale transit networks," European Journal of Operational Research, vol. 37, no. 2, pp. 176-186, 1988.

[9] A. D. Febbraro and S. Sacone, "An online information system to balance traffic flows in urban areas," in Proceedings of the 36th IEEE Conference on Decision and Control, San Diego, CA, 1997.

[10] A. Lozano and G. Storchi, "Shortest viable path algorithm in multimodal networks," Transportation Research Part A: Policy and Practice, Elsevier, vol. 35, no. 3, pp. 225-241, 2001.

[11] M. Bielli, A. Boulmakoul and H. Mouncif, "Object modeling and path computation for multimodal travel systems," European Journal of Operational Research, vol. 175, no. 3, pp. 1705-1730, 2006.

[12] A. Ensor and F. Lillo, "Colored-Edge Graph Approach for the Modeling of Multimodal Transportation Systems," Asia-Pacific Journal of Operational Research, vol. 33, no. 1, pp. 1-21, 2016.

[13] H. Ayed, C. Galvez-Fernandez, Z. Habbas and D. Khadraoui, "Solving time-dependent multimodal transport problems using a transfer graph model," Computers & Industrial Engineering, vol. 61, no. 2, pp. 391-401, 2011.

[14] D. M. Sergio and R. Silvia, "An architecture for a Mobility Recommander System," Procedia Computer Sscience, vol. 98, pp. 425-430, 2016.

[15] Z. Athanasios and W. Whitney , "An intermodal optimum path algorithm for multimodal networks with dynamic arc travel times and switching delays," European Journal of Operational Research, vol. 125, no. 3, pp. 486-502, 2000.

[16] T. Gräbener, "Calcul d'itinéraire multimodal et multiobjectif en milieu urbain: Modélisation et simulation," Université des Sciences Sociales, Toulouse I, France, 2010.

[17] Z. Jianwei, L. Feixiong, A. Theo and T. Harry, "A multimodal transport network model for advanced traveler information systems," in Procedia - Social and Behavioral Sciences, 2011.

[18] D. Kalyanmoy, "Multi-objective Optimization," in Burke E., Kendall G. (eds) Search Methodologies. , Boston, Springer, 2013, pp. 403-449.

[19] F. Reza Zanjirani, M. Elnaz, S. W.Y and . R. Hannaneh, "A review of urban transportation network design problems," European Journal of Operational Research, vol. 229, no. 1, pp. 281-302, 2013.

[20] S. Juliana Verga , Y. Akebo , C. S. Ricardo and S. Wesley Vagner Inês , "Urban Transport and Traffic Systems: An Approach to the Shortest Path Problem and Network Flow Through Colored Graphs," in Nazário Coelho V., Machado Coelho I., A.Oliveira T., Ochi L. (eds) Smart and Digital Cities. Urban Computing. Springer,, Cham, 2019.

[21] C. Xinyuan and K. Inhi, "Modelling Rail-Based Park and Ride with Environmental Constraints in a Multimodal Transport Network," Journal of Advanced Transportation , vol. 2018, pp. 15 pages,, 2018.

An Intelligent Semi-Latin Square Construct for Measuring Human Capital Intelligence in Recruitment

Emmanuel C.Ukekwe[1], Francis S. Bakpo[2], Mathew C.Okoronkwo[3], Gregory E.Anichebe[4]
Department of Computer Science
University of Nigeria, Nsukka, Enugu, Nigeria

***Abstract*—Processing speed and memory recall ability are two major Human Capital Intelligence attributes required for recruitment. Matzel identified five domains of Intelligence. Unfortunately, there were no stated means for measuring them. This paper presents a framework for measuring Processing speed and Memory intelligence domains using Sternberg and Posner paradigms of short memory scanning test. A Semi-Latin square was constructed and used as a competitive platform for n= 20 student-applicant contestants. The Cumulative Grade Point Average rankings of 20 randomly selected final year student-applicants were used for the test. Results show that the CGPA performance ranking of the student-applicants differ from that of the HCI using the framework. A Wilcoxon Signed-Ranks Test was used to determine if the disparity in performance ranking was significant. Results show that there is indeed a significant difference in the performance ranking of the student-applicants using both approaches. The automated Construct was implemented using PHP and Mysql and deployed at (hcipredictor.eu3.org).**

Keywords—Memory recall ability; processing speed; Sternberg paradigm; Posner paradigm; human capital intelligence; Semi-Latin square

I. Introduction

Human capital intelligence (HCI) is an embodiment of knowledge, creativity, talents, habits, social and personality attributes, inherent in a person which could be measured in terms of economic value. HCI is a major factor to be considered especially for labour recruitment, leadership positions and managerial posts. Choosing the appropriate labour force for a business venture, leadership ability or managerial position has always been an uphill task especially in developing countries. This is so because the emphasis and major criteria for choosing the labour force is based on paper qualification and certification of the applicants. This leads to erroneous judgment in the quality of labour employed. This is evidently true because the quality of education in such countries is questionable. The education system is characterized by untested facilitators, examination malpractice, inconsistent education policies and poor funding and a lot more. A Nigerian education critic decries the condition of her educational system which gets bedeviled by the day as people are no longer judged by the latent ability in them but the certificates they have gathered by whatever means [1]. As a result of these, education system of developing countries continues to produce half-baked graduates with low HCI value. The implications are for more reaching. A lot of small scale business ventures and organizations had packed up due to incompetent labour force. Experts are being sought from developed countries for staff retraining and subsequent salvaging of ventures. Output from companies continues to dwindle in quality and number due to incompetent labour force. Business firms, institutions and the society at large continue to lack in competent and skilled labour force. The existing recruitment process is not producing the desired competent labour that is needed to contend with the fast increasing economic challenges and production requirements because it solely depends on certificates obtained from sick education system. The existing recruitment attempts at ascertaining proficiency relies on human resource tools which includes aptitude tests (either written or computer based) and basically oral interviews. These methods come with their short falls. A person's score in an aptitude test is known to be a function of the examiner, the subject matter focus and the educational background. These short falls raise doubts on issues like possibility of test questions being revealed by the examiners before the test, possibility of favoritism during oral interviews and other related issues. These challenges continue to make it difficult to select competent labour force with sound human capital Intelligence value especially in developing countries. One way to improve the quality of labour force is to focus on means of measuring Intelligence as a human capital value. Intelligence has been identified as a major index in recruitment. Intelligence is not only related to the extent of knowledge gained or acquired by the individuals. It reveals the capability to yield from proper training, reason conceptually, think and solve problems [2]. Intelligence varies among individuals hence the need for Intelligence tests. Intelligence tests are known to specifically measure abilities of a person while cognitive tests measure a person's learning in a specific subject area. Intelligence tests are known to produce desired results and high predictive values and when it is combined with well-structured interview it could have the highest predictive value of all the methods of selection [3]. The aim of this paper therefore is to proffer a standard method of measuring human capital intelligence of applicants for recruitment into establishments and industries rather than depending on their over-rated certificates. The sole objective is to ensure that the measurement is done under an equitable and competitive platform. The measurement is achieved using a short memory test which has speed and memory recall abilities as its yardstick.

*Corresponding Authors.

The major aim of the research therefore is to present a framework for measurement of processing speed and memory intelligence domains as HCI parameter which was lacking in [4]. The specific objective is to test the developed framework using 20 graduating student-applicants of the department of Computer science, University of Nigeria, Nsukka. The research question is interested in finding out if the results from the existing approach are the same as that of the proposed system under the hypothesis that:

H0: The results from both approaches are the same

Vs

H1: The results from both approaches differ

The significance of the work is to use the developed framework to especially aid human resource and recruiting agencies in recruiting more qualified labour that will enhance production.

II. Research Basis

According to [5], Intelligence could be measured under three basic abilities: creative, analytical, and practical abilities. Defining Information processing in terms of creative Intelligence, then we will see it as the ability to convert latent information into manifest information [6]. If we also consider that in real time, information processing requires instant decisions within limited time, we will also see speed and perception as creative and analytical intelligence respectively. According to [4], five domains of Intelligence has been identified as Reasoning domain, Processing speed domain, Memory domain, Comprehension domain and an Unknown domain which may be in existence. These identified domains reduce to specific ability tests such as reasoning tests, speed tests, memory tests and spatial tests. The challenge has always been how to proficiently measure these abilities distinctly. For instance, attempts at measuring the memory recall ability has been and is still ongoing. According to the authors in [7], [8], [9] and [10], several attempts had been made at measuring memory using mathematical models such as SAM, REM, MINERVA 2 model, EEG analysis and ERP. Although foundational research on memory argues that short term memory differ from long term memory which presumes that their method of measurement should also differ. However, recent unified attempt at measuring memory ability is seen in [11] and [12] where EEG, FMRI studies and Serial Recall Paradigms. It is not therefore the difference in memory ability that matters but the mode of measurement. Most of the existing measuring procedures attempt to measure these distinct domains using a generalization approach which arguably does not address the peculiar nature of these domains. For instance, Human intelligence and memory recall ability has been measured in literature using the Sternberg paradigm [13]. Sternberg information processing is an information processing paradigm that tests an individual recalling ability. It thrives to ascertain intelligence ability of humans to scan the memory in high speed for information retrieval. The Sternberg experiment involves different trials of experiment in which a random series of say from one to six different digits are displayed at fixed point on the screen for 1.2 seconds delay time. Also given a test digit after a 2 seconds delay time, subjects are to judge whether the test digit is contained in a short memorized sequence of digits previously displayed. In this manner, high speed scanning ability in human memory could be determined per individual and could be used as a measure of intelligence. Similarly, human intelligence and memory recall ability has been measured using Posner approach. The task presents participants with pairs of uppercase, lowercase, and mixed-case letters (drawn from the set A, a, B, b) side by side, 0.5 cm apart on the screen, and these participants were asked to determine, as quickly as possible, whether the letters were the same or different according to a particular rule. Participants indicated that the letters were the same by pressing the M key on a standard keyboard and different by pressing the Z key [14]. These two methods of measuring human intelligence have proved successful and had been the basis of research for many years. It has also been applied to animals in successfully measure of psychometric intelligence and reaction times in pigeons [15]. Unfortunately, these paradigms were applied solely for either Speed or memory recall ability. In other words, the measurement does not take into cognisance other domains of Intelligence. Secondly, the test is a onetime effort which may not really reflect the true ability of the participant in question. There is also no active interactive competition among the participants. Semi-Latin square presents a perfect platform for competition among participants whose human capital intelligence value is to be measured. Semi-Latin squares have found application in many areas of life. In Agricultural experiments for instance, the work of [16] gives credence to the use of a special group of Semi-Latin square known as *Trojan squares* as an experimental design.

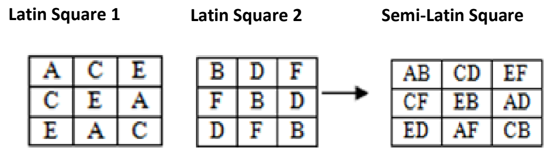
III. Theoretical Background

A. Semi-Latin Square for Equitable and Competitive Platform

According to [17], an (n × n)/k Semi-Latin square is an n × n array containing *nk* letters in such a way that each row-column intersection contains *k* letters and each letter occurs once in each row and once in each column. It suffices to say that no letter occurs more than once in each row and in each column where they are found. There exists a special type of Semi-Latin square called *Trojan Square*. A *Trojan square* is simply an arrangement obtained by superposition of *k* mutually orthogonal n × n Latin square (where such square exists), involving *k* disjoint sets of n varieties so that the resulting square has *kn* varieties, each occurring in n experimental units, *n* rows and *n* columns, with each row intersecting each column in a block of *k* experimental units. *Trojan squares* are constructed by superposition of two mutually orthogonal Latin squares. *Trojan squares* are known to be A-, D- and E-optimal among all binary incomplete-block designs of the same size [17]. The optimality feature of *Trojan squares* gives credence for using them in developing the competitive platform for assessing contestants.

B. Construction of Semi-Latin Squares

A Semi-Latin square is constructed by superposition of the Latin squares for instance, given two Latin squares 1 and 2, a semi-Latin square is obtained by superimposition of Latin squares as shown in Fig. 1.



A	C	E
C	E	A
E	A	C

B	D	F
F	B	D
D	F	B

AB	CD	EF
CF	EB	AD
ED	AF	CB

Fig. 1. A (3 × 3)/2 Semi-Latin Square.

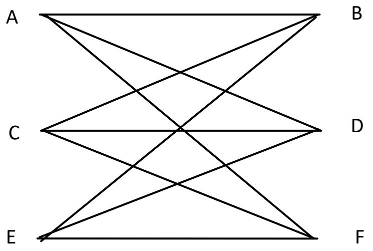


Fig. 2. (3×3)/2 Bipartite Variety Concurrence Graph.

The bipartite variety concurrence graph of the semi-Latin square is also shown in Fig. 2.

C. *Modified Sternberg and Posner Paradigm for Analytical and Creative Intelligence*

The Sternberg paradigm for analytical test of human intelligence tests for the ability and speed to recall and give answers to analytical problems. The HCI uses a modified version of the paradigm to test for analytical intelligence. The Sternberg paradigm in [12] was modified by displaying computer simulated arithmetic expressions requiring each contestant to complete under 5 seconds e.g. 5 + [ ] = -2. Similarly, a selection of pictures ranging from household items, fruits, human body parts and animals were randomly displayed in an inverted mode for contestants to quickly identify in less than 5 seconds. Similarly, the Posner paradigm was also modified to test for creative intelligence. The Posner paradigm was also modified by asking the contestants to identify missing vowels in certain words and displaying some reversed words for the contestants to identify the word within 5 seconds. The words used in this expert system are less than or equal to 5 in length and are obtained from advanced learners dictionary.

IV. METHODOLOGY

An experiment was carried out using the CGPA (Cumulative Grade Point Average) of twenty (20) graduating students of the Department of Computer Science, University of Nigeria, Nsukka. The students were randomly selected from the 2017/2018 Departmental Board approved list of batch A graduating students. The 20 student-applicants are the contestants competing to be employed into establishments and companies based on their final CGPA and class of honors. Their final CGPA represents their actual performance in school. After obtaining their CGPA, the students were then subjected to a (5×5)/4 Trojan Semi-Latin square competitive platform where they were tested for creative and analytical intelligence using Sternberg and Posner paradigms.

A. *Construction of Semi-Latin Square for the Competitive Platform*

The twenty students were first divided into four groups of five (5) students each. Using only their user names and designated symbols, the layout is shown in Table I and Table II.

The four groups gave rise to four sets of mutually orthogonal Latin squares which were constructed using the following equation [18].

$$a_{ij}^{h}(i \pm hj) mod\ 5 \qquad (1)$$

The four Latin squares are then superimposed together to get the (5×5)/4 Trojan square in Table III.

TABLE. I. LAYOUT OF THE STUDENTS' GROUPS

OFFOR (1)	GIFT (A)	NGO (a)	OKPA (@)
UGWU (2)	KACH (B)	EZE (b)	CHI (#)
EMMA(3)	VIN (C)	OBI (c)	UCHE ($)
JOHN(4)	UWA (D)	ROSE (D)	KALU (%)
IKE(5)	AGHA (E)	OKO (e)	JOEL (&)

TABLE. II. LAYOUT OF THE COMPETITIVE PLATFORM

Latin 1						Latin 2				
1	2	3	4	5		A	B	C	D	E
2	3	4	5	1		D	E	A	B	C
3	4	5	1	2		B	C	D	E	A
4	5	1	2	3		E	A	B	C	D
5	1	2	3	4		C	D	E	A	B
Latin 3						Latin 3				
a	b	c	d	e		@	#	$	%	&
c	d	e	a	b		&	@	#	$	%
e	a	b	c	d		%	&	@	#	$
b	c	d	e	a		$	%	&	@	#
d	e	a	b	c		#	$	%	&	@

TABLE. III. SEMI-LATIN SQUARE FOR TESTING TWENTY APPLICANTS

Sess	Grp1	Grp2	Grp3	Grp4	Grp5
1	1,A,a,@	2,B,b,#	3,C,c,$	4,D,d,%	5,E,e,&
2	2,D, c,&	3,E, d,@	4,A, e,#	5,B, a,$	1,C, b,%
3	3,B, e,%	4,C, a,&	5,D, b,@	1,E, c,#	2,A, d,$
4	4,E, b,$	5,A, c,%	1,B, d,&	2,C, e,@	3,D, a,#
5	5,C, d,#	1,D, e,$	2,E,a,%	3,A, b,&	4,B, c,@

B. *Experimental Procedure*

The constructed (5×5)/4 Trojan square ensures that each contesting candidate competes with others especially from opposite class for n=5 times. The experiment was carried out in 5 sessions each having 5 groups of 4 students. The students competed among themselves at the same time in different groups and session. This means that for a given session, the Sternberg and Posner tests run concurrently for each group in that session. The Trojan square also ensures that no contestant will be in more than one group in a session because completion for every group in a given session goes on simultaneously with others. The creative and analytical intelligence tests contestants on the four major domains of intelligence which are Reasoning domain, Processing speed domain, Memory domain, Comprehension domain as identified in [4].

As each group competes, the position ranking of each student is taken. The final score obtained is dependent on both the position ranking and the total score for all the groups. The questions were classified into: Posner reversed word test, Posner missing word vowel test, Sternberg Arithmetic test, Sternberg inverted alphabet test and Sternberg inverted picture test.

C. *Score Sheet*

The Experiment was carried out under the following control constraints:

a) Questions are automatically generated by the system in real time ensuring that students or administrators do not have prior knowledge of the questions before hand.

b) Equal time duration for each question displayed.

c) Each group in a session answers the same questions.

d) Each student answered a total of hundred (100) questions coming from Posner and Sternberg modified short memory tests.

e) Each contestant logs into the application and completes the test.

In general, for any Semi-Latin square (n×n)/k platform, the total number of questions per contestant is given as:

TQ = 20 *((n*k)/k) (2)

If n = 3 and k = 2 (i.e (3×3/2) = 6) then total no of Questions per candidate = 20 * 3 = 60, If n = 5 and k = 4 (i.e (5×5/4) = 10) then total no of questions = 20 * 5 = 100

D. *Score Inference Engine*

If letter ***A*** represents the students answer for a question and letter ***B*** represents the correct answer to a question, then using the set of logical values as Boolean algebra, the score inference engine is represented thus:

(Posner Reversed word test) =

$$\begin{cases} T, & \text{If } A = B \\ F, & \text{Otherwise} \end{cases}$$

(Posner Missing Vowel test) =

$$\begin{cases} T, & \text{If } len(A) = len(B) \text{ AND } B \in D \text{ (Dictionary)} \\ & F, \text{ Otherwise} \end{cases}$$

(Sternberg Arithmetic test) =

$$\begin{cases} T, & \text{If } A = B \\ F, & \text{Otherwise} \end{cases}$$

(Sternberg Inverted Alphabet test) =

$$\begin{cases} T, & \text{If } A = B \\ F, & \text{Otherwise} \end{cases}$$

(Sternberg Inverted Pictures test) =

$$\begin{cases} T, & \text{If } A = B \\ F, & \text{Otherwise} \end{cases}$$

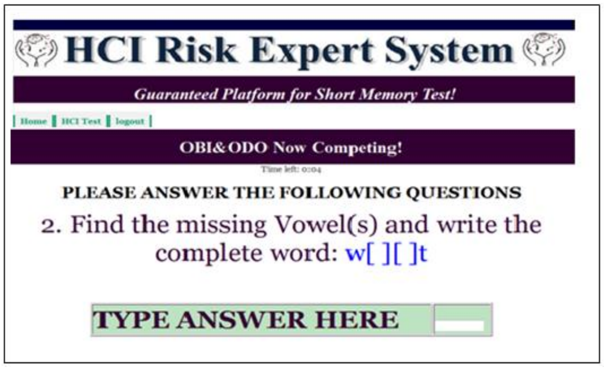
The score inference engine layout is summarized in Table IV.

E. *Applicatn Screen Shots*

Some of the screen shots from the application showing examples of test questions are shown in Fig. 3 and Fig. 4.

TABLE. IV. SCORE INFERENCE ENGINE

Test	Contestant answer (A)	Correct answer (B)	Rule	Score
Posner reversed word test	A	B	If A=B then	True
Posner missing vowels test	A	B	If len(A) = len(B) and A (found in dictionary) then	True
Sternberg Arithmetic test	A	B	If A=B then	True
Sternberg inverted alphabet test	A	B	If A=B then	True
Sternberg inverted pictures test	A	B	If A=B then	True



HCI Risk Expert System

Guaranteed Platform for Short Memory Test!

| Home | HCI Test | logout |

OBI&ODO Now Competing!

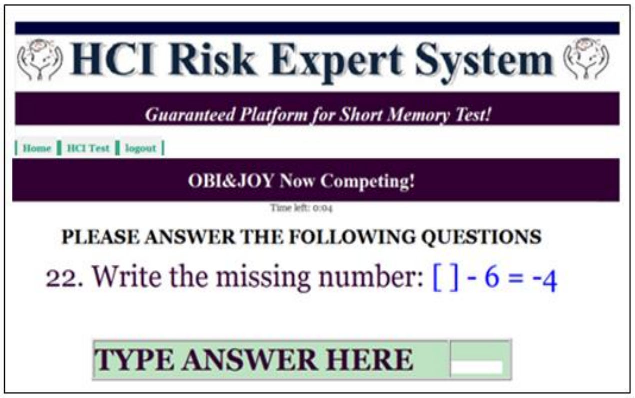
Time left: 0:04

PLEASE ANSWER THE FOLLOWING QUESTIONS

2. Find the missing Vowel(s) and write the complete word: w[ ][ ]t

TYPE ANSWER HERE

Fig. 3. Posner Missing Vowel Test for Creative Intelligence.



HCI Risk Expert System

Guaranteed Platform for Short Memory Test!

OBI&JOY Now Competing!

Time left: 0:04

PLEASE ANSWER THE FOLLOWING QUESTIONS

22. Write the missing number: [ ] - 6 = -4

Fig. 4. Sternberg Arithmetic Test for Analytical Intelligence.

V. Analysis and Results

A. Analysis

Using the competitive Semi-Latin square platform show in Table III, the Posner and Sternberg test for memory and speed processing was carried out by the developed application. Each scheduled group of four (4) competing applicants is meant to answer twenty (20) questions. The performance ranking of each applicant is noted for each competition. In this case, there were five (5) competition schedules which gave five (5) different performance ranking and another five (5) different sub total score for each applicant. The scaled CGPA score represents the contestant's score based on the Senate approved CGPA. The scaled CGPA is computed using the following equation:

$$CGPA\ score = \frac{CGPA}{5.0} \times \frac{100}{1} \qquad (3)$$

The HCI score is obtained by taking into cognizance the total score for each student in all the sessions and their respective total rank score (TR). The result is then scaled to 100 percent. The HCI score is obtained using:

$$HCI\ score = \left(\frac{CGPA\_score}{TQ} \times \frac{n}{TR} \times \frac{100}{1}\right) \qquad (4)$$

Where TQ = total questions, *TR* represents the total performance ranking of each student based on their position after every successful group competition.

For instance, IKE has a CGPA score = 56.6, TR = 14, TQ = 100, n = 5. The HCI score will thus be calculated as:

$$HCI\ score = \left(\frac{56.6}{100} \times \frac{5}{14} \times \frac{100}{1}\right) = 20.21$$

The CGPA ranking is obtained by taking the position of each contesting student based on the CGPA. Similarly, the IQ ranking is obtained by taking the position of each contesting student based on their IQ score.

B. Results Obtained

The system successfully measured the HCI value of the twenty (20) students. The results are compared to their CGPA's as shown in Table V.

A bar chart showing the deviation between the CGPA and HCI ranking is shown in Fig. 5.

The results obtained showed that there was a disparity between the performance rankings of the two approaches. In order to ascertain whether the deviation was really significant, the Wilcoxon Signed-Ranks Test was carried out using the students CGPA score and the obtained HCI score. Firstly, a normality test on the difference between the two scores from both methods was carried out. The results are shown in Table VI and Fig. 6, respectively.

TABLE. V. CGPA and HCI Scores

ID	User name	CGPA	CGPA score	HCI Score	TR	CGPA Ranking	HCI Ranking	Deviation
1	IKE	2.83	56.6	20.21	14	**13**	**11**	2
2	OFOR	4.65	93.0	93.00	5	**1**	**2**	1
3	UGWU	2.52	50.4	15.75	16	**16**	**13**	3
4	EMMA	4.59	91.8	91.80	5	**2**	**1**	1
5	JOHN	3.30	66.0	23.57	14	**12**	**17**	5
6	AGHA	2.77	55.4	15.11	18	**14**	**16**	2
7	VIN	1.69	33.8	9.70	20	**20**	**18**	2
8	GIFT	2.29	45.8	12.05	19	**17**	**10**	3
9	UWAH	3.91	78.2	32.58	12	**9**	**9**	0
10	KACH	2.08	41.6	10.40	20	**19**	**19**	0
11	OKO	4.03	80.6	57.57	7	**4**	**5**	1
12	EZE	2.18	43.6	12.11	18	**18**	**15**	3
13	ROSE	4.01	80.2	57.29	7	**5**	**4**	1
14	NGO	4.31	86.2	61.57	7	**3**	**3**	0
15	OBI	3.45	69.0	28.75	12	**11**	**14**	2
16	JOEL	2.74	54.8	19.57	14	**15**	**20**	5
17	KALU	3.49	69.8	31.73	11	**10**	**12**	2
18	UCHE	3.93	78.6	39.30	10	**7**	**7**	0
19	CHI	4.01	80.2	40.10	10	**6**	**6**	0
20	OKPA	3.68	73.6	33.45	11	**8**	**8**	0

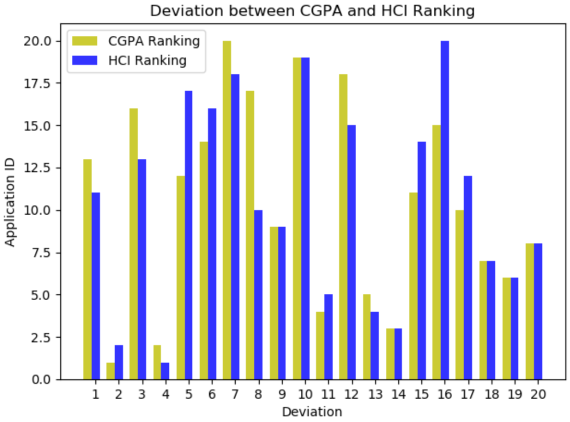


Fig. 5. Deviation between CGPA and HCI Ranking.

TABLE. VI. TESTS OF NORMALITY

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Diff_HCI_CGPA	.201	20	.034	.813	20	.001

a.Lilliefors Significance

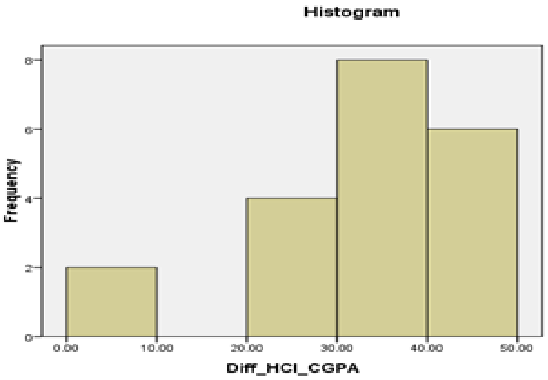


Fig. 6. Normality Test on Data Set.

The reported *P* value (0.034 and 0.001) are all less than 0.05, and the histogram was right skewed, hence we conclude that the data was not normally distributed. As a result of the failed normality test, an alternative non-parametric test was used. The Wilcoxon Signed-Ranks Test was used to determine whether there was actual difference in the results obtained from both methods.

The null hypothesis which says that the two methods are the same was tested against the alternative which says that the two methods vary.

H_0 : CGPA_Score = HCI_Score

Vs

H_1: CGPA_Score ≠ HCI_Score

The result of the Wilcoxon Signed-Ranks test is shown in Table VII.

The Wilcoxon Signed-Ranks Test Result shows that the Asymp. Sig. (2-tailed) value was 0.00 and less than 0.05. As a result, the null hypothesis was rejected while the alternative was accepted. The conclusion therefore is that the two methods produce different results contrary to expectation.

TABLE. VII. WILCOXON SIGNED-RANKS TEST RESULT

Test Statistics[b]	
	HCI_Score-CGPA_Score
Z Asymp. Sig. (2-tailed)	-3.724[a] 0.000

a. Based on Positive ranks
b. Wilcoxon Signed Ranks Test

VI. CONCLUSION

The framework was successful in measuring the processing speed and memory intelligence domains. The results obtained show that the existing approach to recruitment which emphasizes academic performance differs from the proposed framework which on the other hand lays more emphasis on processing speed and memory recall ability. Results show that the position ranking of the student-applicants based on their final CGPA results differ from their HCI ranking. The Wilcoxon Signed-Ranks Test which was significant at 0.00 validates the disparity. One would expect that their CGPA performance should tally with their HCI result under the developed construct but it wasn't so. The best candidate in the CGPA ranking is not necessarily the best in the HCI ranking. The Semi-Latin square construct could be said to be a true representation of the students' intelligence virtue because:

a) The 20 students had equal opportunity and constraint to compete with one another.

b) There was more than one session of test thereby reducing the issue of unfamiliarity of the system.

c) The questions from the construct were simulated and are not known beforehand.

d) The questions are not limited to a discipline but only test for creative and analytical intelligence using what the students could reason out within a short time.

VII. PROSPECTS FOR FUTURE RESEARCH

Future research interest should therefore focus on using creative and analytical intelligence as a major criteria for recruitment rather than placing so much emphasis on applicant's certificates. Interest should also focus on increasing the number of applicants from 20 to a higher number in order to fit into practical situations.

REFERENCES

[1] S. Obanubi, "The Love of Certificate:Where to Nigeria?". January 15, 2015. Retrieved from :http://unitenigeria.com/nigeria-and-the-love-of-certificate/.

[2] H.G. Uzma and H. Tajammal. "Comparative Study of Intelligence Quotient and Emotional Intelligence: Effect on Employees' Performance". Asian Journal of Business Management January 15, 2013, Vol 5(1): Pg 153-162, Retreived from: https://doi.org/10.19026/ajbm.5.5824.

[3] F.L. Schmidt and J.E. Hunter "The validity and utility of selection methods in personnelpsychology: practical and theoretical implications of 85 years of research findings. Psychological Bulletin, 1998, 124 (2), pp 262–74,Retreived from: https://doi.org/10.1037/0033-2909.124.2.262.

[4] L.. Matzel and B. Sauce "IQ. In: Vonk J.Shackelford T. (eds) Encyclopedia of Animal Cognition and Behavior". Springer, Cham, 2017, https://doi.org/10.1007/978-3-319-47829-6_1080-1axwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[5] D. McGonigle and K. Mastrian " Introduction to information, information science, and information systems. Jones & Bartlett, 2011. Retrieved from: http://samples.jbpub.com/9781449631741/92367_CH02_017_032.pdf

[6] G.W Jeroen and M.S. Richard, “Models of Memory”in Stevens”, Handbook of Experimental Psychology, Third Edition, 1988, Volume 2: Memory and Cognitive Processes. (Pp.43-76). New York: JohnWiley&Sons,Inc.

[7] H.U. Amin, A.S Malik., N. Badruddin and,W.T Chooi “ Brain Behavior in Learning and Memory Recall Process: A High-Resolution EEG Analysis. In: Goh J. (eds) The 15th International Conference on Biomedical Engineering. IFMBE Proceedings, vol 43. Springer, 2014.

[8] H.U Amin, A.S. Malik, S. Aamir, N. Kamel, W.T Chooi and H.Muhammad. “P300 correlates with learning & memory abilities and fluid intelligence”, Journal of NeuroEngineering and Rehabilitation, Vol 12, Article number: 87, 23rd September, 2015.

[9] S. Hanouneh, H.U. Amin, N.M. Saad and A.S Malik. "The correlation between EEG asymmetry and memory performance during semantic memory recall," 2016 6th International Conference on Intelligent and Advanced Systems (ICIAS), Kuala Lumpur, 2016, pp. 1-4. doi: 10.1109/ICIAS.2016.7824041.

[10] Mc Hafeezullah Amin, and A.S. Malik “Human memory retention and recall processes A review of EEG and fMRI studies”, Neurosciences 2013; Vol. 18 (4).

[11] M.S. Ahmed and S.A. Yasir “Examining the Effect of Interference on Short-term Memory Recall of Arabic Abstract and Concrete Words Using Free, Cued, and Serial Recall Paradigms”, Advances in Language and Literary Studies, December 2015, Vol. 6 No. 6, ISSN: 2203-4714.

[12] S. Sternberg. “High-Speed Scanning in Human Memory”, Science, New Series, Vol.153, No. 3736. Pp. 652-654, August, 1966. Made Available by JSTOR on September, 2005 Retrieved from:http://www.jstor.org/.

[13] A. Douglas and F. Bert “Age, Speed of Information Processing, Recall, and Fluid Intelligence”. University of Toronto, Scarborough Campus, Ontario, Canada., pg. 229-248, 1995.echnical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[14] C. N. Aljoscha, R. Rainer, M. Ralf and A. Alois “Intelligece and Reaction Times in the Hick, Sternberg and Posner Paradigms”. Person. individ. Diff. 1997, Vol. 22, No. 6, pp. 885-894ouglas and F. Bert “Age, Speed of Information Processing, Recall, and Fluid Intelligence”. University of Toronto, Scarborough Campus, Ontario, Canada., pg. 229-248, 1995.echnical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[15] D.A. Preece and G.H. Freeman, “Semi-Latin squares and related designs. J.Roy.Statist.Soc. Ser. B45, 267-277, 1983, retrieved from: https://doi.org/10.1111/j.2517-6161.1983.tb01250.x.

[16] R.A. Bailey and P.E. Chigbu, Enumeration of Semi-Latin Squares, Discrete math, 1997, 167/168, Pg 73-84, Retrieved from: https://doi.org/10.1016/S0012-365X(96)00217-8.

[17] R.A. Bailey. “Efficient Semi-Latin squares”. Statistica Sinica 1992, Vol.2 (413-437).

[18] R.N. Mohan, M. Ho Lee, and S.S. Pokhrel, On Orthogonality of Latin Squares, J. Comb. Infor. System Sci., 2005, Vol.30(1-4), Pg151- 179. Retrieved from: https://www.researchgate.net/publication/1959255_On_Orthogonality_of_Latin_Squares

Human Gait Feature Extraction based-on Silhouette and Center of Mass

Miftahul Jannah[1], Sarifuddin Madenda[2], Tubagus Maulana Kusuma[3], Hustinawaty[4]
Faculty of Computer Science and Information Technology
Gunadarma University, Jakarta, Indonesia

***Abstract*—When someone walks, there is a repetitive movement or coordinated cycle that forms a gait. Gait is different, unique and difficult to imitate. This characteristic makes gait one of the biometrics to find out one's identity. Gait analysis is needed in the development of biometric technology, such as in the field of security surveillance and the health sector to monitor gait abnormalities. The center of mass is the unique point of every object that has a role in the study of humans walking. Each person has a different center of mass. In this research, through a series of processes in image processing such as video acquisition, segmentation, silhouette formation, and feature extraction, the center of mass of the human body can be identified using a webcam with the resolution of 640 x 480 pixels and the frame rate of 30 frames/second. The results obtained from this research were gait frames of 510 frames from 17 pedestrian videos. Segmentation process using background subtraction separates the pedestrian object image from the background. Silhouette gait was produced from a series of image enhancement processes to eliminate noise that interferes the image quality. Based on the silhouette, feature extraction provides the center of mass to distinguish each individual's gait. The sequence of center of mass can be further processed for characterizing human gait cycle for various purposes.**

Keywords—Human gait; center of mass; silhouette; feature extraction; gait cycle; people identification

I. Introduction

Walking is a movement that allows one to move from one place to another by moving the foot forward in the correct position alternately [1]. Repeated movements or coordinated cycles form a gait. Every individual's gait is different, which makes it unique and difficult to imitate. These characteristics then make gait one of the biometrics to find out one's identity.

Biometrics is a technology of recognition and identification based on physiological or behavioral characteristics possessed by humans such as gait, face, voice, iris, fingerprint, etc. [2]. However, biometric recognition with face, sound, iris, and fingerprint cannot be done remotely and requires interaction with the subject to be observed. In contrary, gait biometric not require direct contact with the subject to be observed making the image acquisition of gait can be performed easily in public places as well as captured remotely. Gait is difficult to be hidden and engineered. This characteristic is very important in the surveillance system.

Gait identification is needed in various applications, such as in the health sector where identification is intended to identify the type of disease from abnormal gait motion. Gait identification also has an important role in video surveillance and access control systems for supervision and security, for example in security sensitive environments such as airports, banks and certain spaces. Other information can be obtained through identification of gait such as age, race, and gender.

In order to identify human gait, image processing is needed, which consists of several stages, such as capturing gait videos using a camera. Captured videos that consist of a set of image frames are extracted so that they can be processed frame by frame to produce a silhouette image. Feature extraction of silhouette images is the key step in identifying gait. Silhouette images represent binary maps of human walking, forming strong features to represent gait because they capture the movements of most parts of the human body [3]. Feature extraction has been carried out in existing studies such as extraction of the entire human body [4] or some limbs such as the waist, hips, feet [5].

A number of research related to gait extraction features have been carried out. As proposed in [6], gait extraction based on features of distance and angle between the two legs using Hough Transform was proposed. This study produced a silhouette and skeleton gait. Different approach was shown in [5], where gait was analyzed using DGait database. Feature extraction from 2D and 3D body silhouettes for gait identification was performed. Support Vector Machine (SVM) kernels were used for classification. This research successfully compared the two features (2D and 3D). In [7], Multi-scale Principal Component Analysis (MSPCA) was proposed, which performed gait recognition based on modelling limbs using a spline curve. The feature extraction uses the CASIA-B Gait Database silhouette dataset. For classification Neuro-Fuzzy and K-Nearest Neighbors (KNN) was used. Another approach was introduced in [8], where Kinect camera sensor was used for acquisition. The feature extraction process used static features and dynamic features such as wrists, ankles, body, knees, shoulders, arms and thighs. K-Nearest Neighbors (KNN) was used for gait classification. The research output is a database containing 20 pedestrians walking from right to left.

It was shown that image frames and silhouettes used in the previous research were not captured and processed in real time but based on provided datasets. Feature extraction has been done on several members of the human body but has not used the Center of Mass (CoM) that has a role in the study of humans walking. Therefore, real-time process of acquisition, silhouette generation, and feature extraction based on CoM are proposed in this research. The importance and findings of this research is the extraction of CoM sequence from human walking cycle, which can be further used for classification purpose.

II. Methodology

The feature extraction method consists of a number of stages, namely video acquisition, image segmentation, and silhouette generation/forming as shown in Fig. 1. The output of feature extraction process is the CoM and its location in the human body image.

A. Video Acquisition

The video acquisition was performed in real time when someone walks in front of the camera. Webcam was positioned 50 cm above the ground with the distance to the object of 3 meters. The viewing angle of the webcam is 90^o. Gait analysis of the video files can only be done frame-by-frame. The video resolution of the gait frame extraction process is 640 x 480 pixels.

The total number of video used in this research is 17 videos, where each video contains human pedestrian. In the process of frame extraction, all frames in the video were extracted. Afterward, using the extracted frames, background images were captured. The specification of the video acquisition is shown in Table I.

Fig. 2 shows the stages in identifying the CoM is the gait video acquisition process which consists of camera calibration, video recording, and video frame extraction. The camera calibration process is used to adjust the camera position, distance, capturing angle, and adequate lighting in order to obtain biometric information of the gait cycle and the configuration of the device used for the gait video recording process. The recording process starts with recording the background and recording the person walking in front of the background. Recorded videos are saved in .avi video file format. Once the video is recorded, frame extraction is taken place to produce a series of image frames to be able to do the next process as shown in Table II.

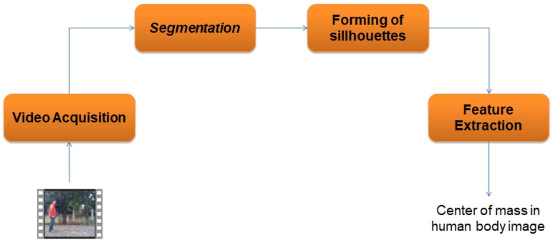


Fig. 1. Research Methodology.

TABLE. I. The Specification of the Video Acquisition

Number	Specification	Description
1	Camera	Logitech C170 Webcam
2	File Format	.avi
3	Duration	5 seconds
4	Video Resolution	640 x 480 pixels
5	Frame Rate	30 frames/second
6	Bit Depth	24-bit
7	Sensor Resolution	True 2 Megapixels

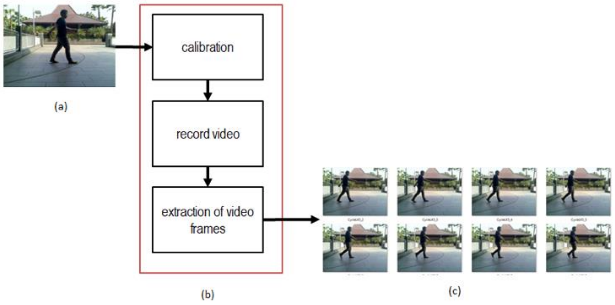


Fig. 2. Video Acquisition Process.

TABLE. II. Example Results of Frame Extraction

Frame background 	**Frame 1** 	**Frame 10**
Frame 20 	**Frame 25** 	**Frame 30**

Output frames from the video acquisition process are stored in the database and are used at the segmentation stage.

B. Segmentation

Segmentation was used to separate object of interest from its background. In this process, the background subtraction is carried out. Prior to the subtraction process, the color space of the image frame was converted into a grayscale image, as shown in (1) [9]. The process of background subtraction was performed by subtracting each pixel of the background image with each pixel of the gait image, foreground image $F(x, y)$ is obtained by subtracting the complete image $G(x, y)$ with background image $B(x, y)$ as shown in (2).

$$Gray = 0.299 * R + 0.587 * G + 0.114 * B \tag{1}$$

$$F(x, y) = | G(x, y) - B(x, y) | \tag{2}$$

The flowchart of segmentation process is shown in Fig. 3.

The results of image segmentation process are shown in Table III.

C. Forming of Silhouettes

Silhouette formation is a very important stage in gait identification. In this research, the formation of silhouettes produced binary image after going through several enhancement processes from the previous results.

Image enhancements are needed to improve image quality to make it easier for the next process. The first image enhancement was to eliminate noise in the image using median filter with an 8x8-dimensional matrix. The second process was image morphology using dilation operations as shown in (3) and erosion as shown in (4). However, prior to the morphological process, grayscale images were converted into binary images using the thresholding as shown in (5).

As stated in Eq. 3, the dilation operation closed the gap between two objects by adding pixels around object A to the size of the structure of element B.

$$A \oplus B = \{z \mid (B)_z \cap A \neq 0\} \tag{3}$$

Erosion operation eroded or reduced the area of the object according to the size of the structure of element B.

$$A - B = \{z \mid (B)_z \subseteq A\} \tag{4}$$

Thresholding process is used to convert the image into binary image as presented in Eq. 5. The threshold (T) requirements and desired values were adjusted based on the needs.

$$g(x,y) = \begin{cases} 1 & if \quad f(x,y) \geq T \\ 0 & if \quad f(x,y) \leq T \end{cases} \tag{5}$$

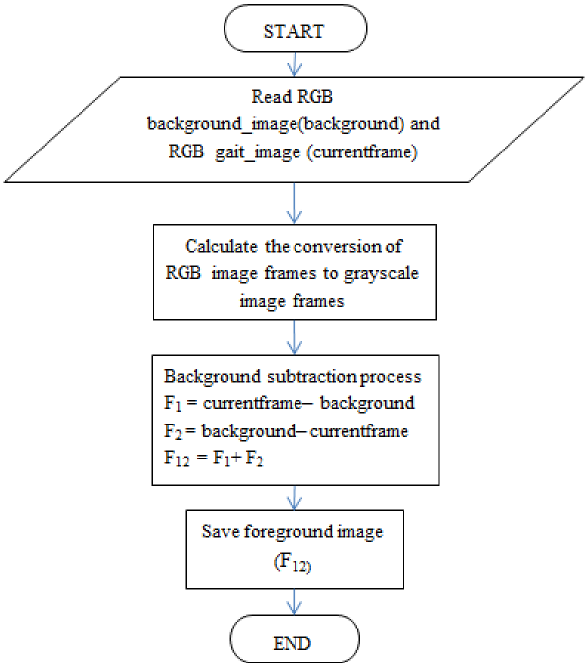
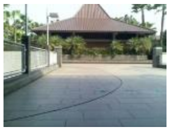
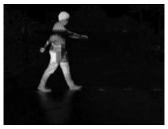

START
Read RGB background_image(background) and RGB gait_image (currentframe)
Calculate the conversion of RGB image frames to grayscale image frames
Background subtraction process
F1 = currentframe– background
F2 = background– currentframe
F12 = F1+ F2
Save foreground image (F12)
END

Fig. 3. Flowchart of Segmentation Process.

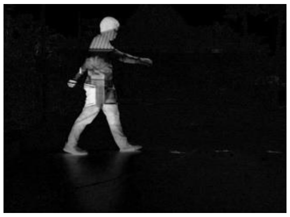
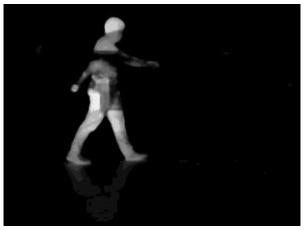
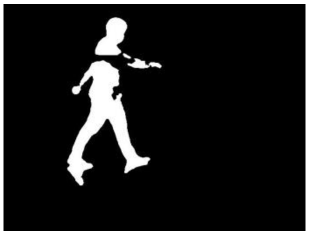
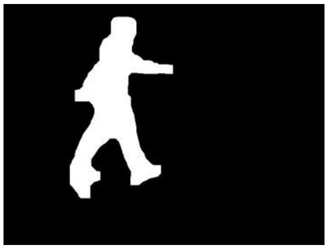
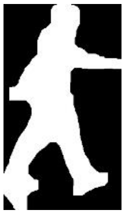
TABLE. III. THE RESULT OF SEGMENTATION PROCESS

Frame	RGB	Grayscale	Background subtraction
Background			--
			

The final stage in silhouette formation is cropping or cutting to produce image frames that focus on gait objects. This cropping process requires the position of x_{min}, y_{min}, width, height. The position of x_{min} was obtained by finding the minimum value of the column, the position of y_{min} obtained by finding the minimum value of the row. The width value was obtained through reducing the maximum column value to the minimum column value, and the height value obtained by reducing the maximum row value to the minimum row value.

Table IV shows the silhouette formation process which consists of four processes. The first column is the sequence of image processing, and the second column is the result of each enhancement process to improve image quality, namely filtering, thresholding, morphology, and cropping.

TABLE. IV. THE RESULT OF SILHOUETTES FORMATION PROCESS

Image Processing	Result
Grayscale imagery, there is noise that can be a nuisance and must be repaired.	
Median image filtering with an 8x8 dimension matrix to disguise the remnants of background images that are considered as noise in the subtraction background image..	
The image thresholding with a threshold value of 32 produces a binary image with the intensity of the color of the background image worth 0 (black) and the color intensity of the gait image value 1 (white).	
Morphological images with dilation and erosion operations close the gap between two objects or holes contained in the image thresholding.	
Cropping on the image frame shows that the frame is more focused on the image of a white object, namely the gait image.	

D. Feature Extraction

Feature extraction is the process of extracting features from each silhouette image to get the CoM of each silhouette image. The CoM (centroid) was generally obtained by using the average coordinate (x, y) value of each pixel composes the object [10]. The center of mass value was stored in the matrix in the form of .mat. As shown in (6), CoM of an object was obtained by calculating the number of pixel in the silhouettes.

$$N = \sum_{x=1}^{n}\sum_{y=1}^{m} B(x,y) \tag{6}$$

Eq. (7) and Eq. (8) shows the calculation the average CoM on the axis of coordinates $(\bar{x}, \bar{y})$ based on the number of pixels [10].

$$\bar{x} = \frac{1}{N}\sum_{x=1}^{n}\sum_{y=1}^{m} B(x,y) \tag{7}$$

$$\bar{y} = \frac{1}{N}\sum_{x=1}^{n}\sum_{y=1}^{m} B(x,y) \tag{8}$$

The results of the silhouette feature extraction and the CoM are shown in Table V.

TABLE. V. EXAMPLE OF THE CENTER OF MASS (X,Y) BASED ON FEATURE EXTRACTION PROCESS

Number	**Frames**	**Center of Mass**	
		x	**y**
1		89	178
2		93	175
3		97	179
4		99	179
5		98	179
6		94	181
7		85	182
8		76	177
9		63	189
10		51	185
11		37	177
12		39	186
13		41	194
14		52	196
15		57	189
16		64	186
17		71	185
18		81	185
19		86	185

20		88	186
21		89	184
22		82	186
23		81	180
24		79	182
25		73	191
26		62	208
27		50	185
28		38	172
29		39	171
30		40	174
31		44	184
32		54	185
33		64	177
34		70	175
35		79	171
36		86	175
37		91	181
38		95	180
39		93	185
40		89	180
41		88	171
42		82	164
43		77	164

44		67	179
45		60	178
46		52	170
47		41	176
48		41	197
49		46	206
50		53	213
51		58	198
52		67	201
53		73	201
54		79	197
55		83	197
56		89	193
57		90	191
58		94	195
59		89	198
60		84	197

As the person start walking, the position of CoM dynamically shifts according to the walking motion. The graph showing the CoM movement on the x-axis and y-axis at 60 frames of silhouette images is shown in Fig. 4 and Fig. 5, which the CoM data were plotted from Table V.

Fig. 4 shows a graph of changes in CoM movement on the x-axis horizontally occurring as a result of changes in position or place when walking. While in Fig. 5 shows a graph of changes in the CoM movement on the y-axis vertically with respect to the coronal plane because when walking, one foot will be on the ground and one foot in the air and ends when the same foot returns to the ground again. This process is known as the gait cycle.

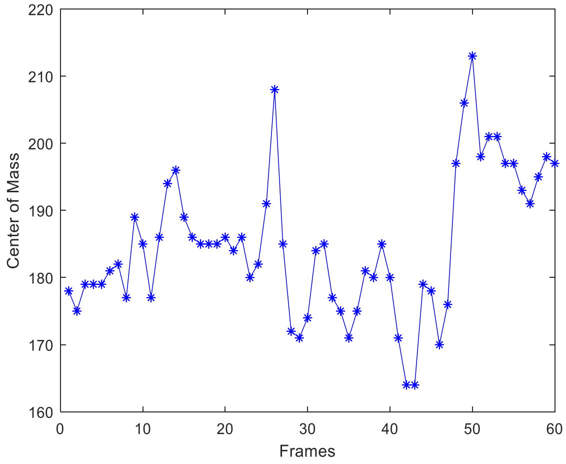


Fig. 4. Movement of the Center of Mass on the X-Axis.

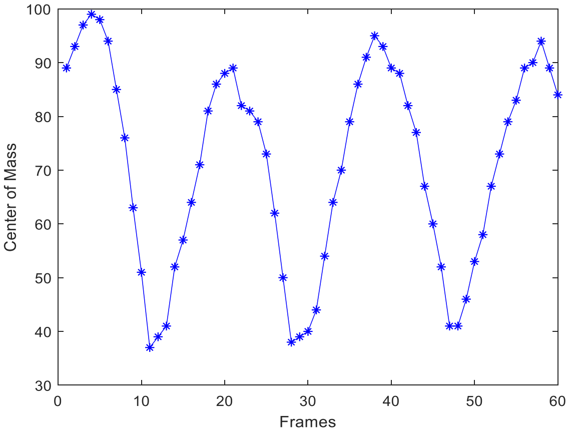


Fig. 5. Movement of the Center of Mass on the Y-Axis.

III. Conclusion

In this research, a frame extraction method based on silhouette and center of mass is presented. The results obtained from this research are 510 frames that were extracted from 17 pedestrian videos. Background subtraction process was successfully separate the gait images from the background. The gait silhouette images were acquired after performing a number of stages by processing grayscale image starting from the noise reduction process to cropping. Based on the silhouette image, feature extraction was performed to obtain the coordinates of the CoM (x, y) for each gait silhouette. The results have shown that the CoM in all image frames were successfully identified. For future work, the CoM can be used as a feature in conducting gait classifications.

References

[1] Carpentier, J., Benallegue, M., and Laumond, J. P., "On The Centre Of Mass Motion In Human Walking," International Journal Of Automation And Computing Vol. 14, Issue 5, October 2017. Pp 542-551. 2017.

[2] Alsaadi, I., M., "Physiological Biometric Authentication Systems, Advantages, Disadvantages And Future Development: A Review," International Journal Of Scientific & Technology Research. Vol. 4, Issue 12, October 2015. Pp 285-289. 2015.

[3] Boulgouris, N., Plataniotis, K., and Hatzinakos, D., "Gait Recogniton Using Linear Time Normalization," Pattern Recognition. Vol. 39, Issue 5, May 2006. Pp 969-979. 2006.

[4] Sudha, L.R. and Bhavani, R.,. "Gait Based Gender Identification Using Statistical Pattern Classifiers," International Journal of Computer Applications. Vol. 40, Issue 8, Pp.30-35, 2012.

[5] Borràs, R., Lapedriza, À., and Igual,L.,. "Depth Information In Human Gait Analysis: An Experimental Study On Gender Recognition". International Conference Image Analysis and Recognition. Vol. Part II, June 2012. Pp 98-105. 2012.

[6] Hustinawaty,. "The Prototype Of Non-Intrusive Skeleton Detection & Feature Extraction Software & Gait Man In Real Time," International Journal Of Sports Science & Engineering. Vol. 07, No. 01, September 2012, Pp. 003-022. 2012.

[7] Sai, S. R., Ravi, R., "Multi-Scale Principal Component Analysis Based Gait Recognition," 1st IEEE/IIAE: International Conference On Intelligent Systems And Image Processing 2013 (ICISIP2013).

[8] Ahmed, M., "Kinect-Based Human Gait Recognition Using Static and Dynamic Features," (IJCSIS) International Journal of Computer Science and Information Security". Vol. 14, No. 12, December 2016.

[9] Gonzalez, R. and Woods, R., Digital image processing. 4th ed. Pearson, 2018.

[10] M.Sayed, "Biometric Gait Recognition Based On Machine Learning Algorithms," Journal of Computer Science. Vol.14, Issue 7, July 2018, Pp 1064-1073, 2018.

Computer Simulation Study: An Impact of Roadside Illegal Parking at Signalised Intersection

Noorazila Asman[1], Munzilah Md Rohani[2], Nursitihazlin Ahmad Termida[3]
Noor Yasmin Zainun[4], Nur Fatin Lyana Rahimi[5]
Department of Infrastructure Engineering and Geomatic[1, 2, 3, 5]
Department of Building and Construction Engineering[4]
Faculty of Civil and Environmental Engineering
Universiti Tun Hussein Onn Malaysia
Parit Raja, Batu Pahat
Johor, Malaysia

***Abstract*—Traffic congestion could be a serious road traffic problem particularly at intersections because of its potential impact on the risk of accidents, vehicle delays and exhaust emissions. In addition, illegal parking by road users at intersections can give additional deterioration to the intersections that may create additional traffic flow interruptions. This paper presented assessment of the illegal parking impact on signalized intersection at Parit Raja, Malaysia using simulation approach using PTV VISSIM simulation software. The results showed that if illegal parkings at Parit Raja intersection were banned, traffic delay and travel time of vehicles will be improved and thus, improving the intersection Level of Service.**

Keywords*—Traffic simulation; traffic flow; signalized intersection; level of service; illegal parking

I. Introduction

Congestion was associated with the necessity and ability to own personal vehicle. This was interrelated to an increment in population and income level that people become affluent to owning vehicles [1]. The increment in traffic volumes on the road contribute to the movement conflict, long queue and stop delays at intersections. In order to reduce congestion at an intersection, traffic signal is introduced. Signalized intersection allows the traffic to cross the road safely as the traffic will be directed to passage in sequence and allow them to cross without obstacles. Signalized intersection, apparently, can reduce right angle accidents, increase road capacity, bring confidence to drivers to cross the roads and provide a good level of service. Although the main function of traffic lights is to reduce the traffic conflict at intersections, installing traffic control devices at intersection areas do not always give an advantage to road users. For example, installing unnecessary traffic control devices can result in disturbing the traffic flows instead of improving it. The installation of traffic lights at intersections can also create other harms to road users such as rear-end crashes and traffic delays [2]. Rear-end accidents at signalized intersections usually occurred because of the leading vehicles' sudden stop due to signal change or traffic situation and drivers' eyes tend to focus on the traffic light rather than the vehicle right in front of them [3]. This kind of situation will slow down the traffic flows at intersections.

In addition, parking on the road sides can have an impact on traffic flows mainly if it is closer to the intersection as it reduces road capacity. Furthermore, side parking can delay or block any movement of vehicles, particularly for left turnings [4]. In India, it was reported that parking was one of the serious problems in urban areas that the country was facing due to the increment of vehicle ownership and the development of mall within the city center [5]. Malaysia is also experiencing similar problems. This is including the illegal parking in unauthorized areas, particularly near the road intersection. Illegal parking on roadsides reduces the effective width of the lane and hence reducing the speed and capacity of the prevailing roadways. As a result, the traffic will be interrupted which cause delays, accidents, congestion, etc.

In previous years, the assessment of the signalized intersections performance had used various approaches that focused on the application of software. Simulation by using software was one among the most popular methods including SUMO, TRANSIMS and PTV VISSIM [6]. These software programs were able to produce simulations of traffic flows and various scenes of traffic operation at signalized intersections without disturbing the traffic at the actual locations to detect any problem occurring at the signalized intersections. The significant of this study was evaluation on the impact of illegal parking by measuring level of service (LOS) on the road for two scenarios, which are road with illegal parking and road without illegal parking.

II. Case Study: Parit Raja Signalized Intersection

This study had been conducted at a signalized intersection located at Parit Raja, Batu Pahat, Malaysia (Fig. 1). The town of Parit Raja was one of the busiest areas in Batu Pahat as it is the main destination for businesses and occupations. There were shop lots, banks, offices, industrial areas and educational institutional buildings within the small town. Parit Raja signalized intersection was an important junction which connects two main cities, Kluang and Batu Pahat.

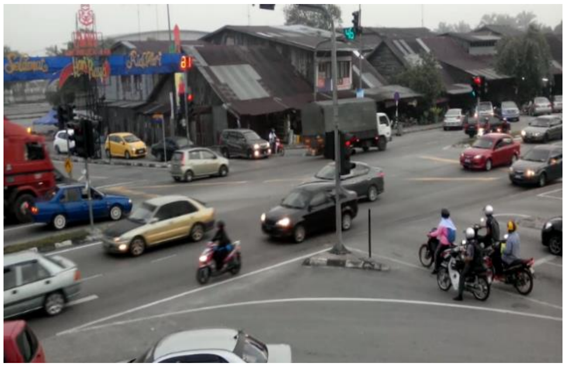


Fig. 1. Study Location.

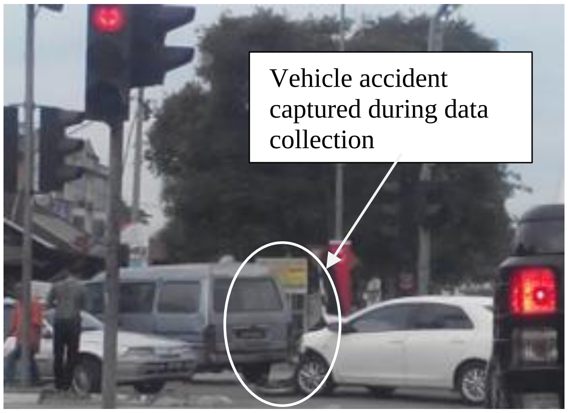


Fig. 2. Vehicle Accident at Parit Raja Signalized Intersection.

Parit Raja intersection was a four-legged junction and controlled by traffic signals. From the traffic study conducted, traffic flow at Parit Raja intersection had the highest traffic volume during morning peak hours at 7.00 to 9.00 am. During this period, traffic flow was mostly dominated by road users who live nearby. Accidents still occur at the Parit Raja signalized intersection even when the intersection was controlled by traffic signals (Fig. 2). Social activities such as businesses on the road curbs encourage people to illegally park their vehicles on the side of the roads, affecting the traffic flow. This study simulates traffic flow at Parit Raja signalized intersection using PTV VISSIM software and investigate the impact of illegal parking at the intersection. This study only limits to 2km radius from the intersection area and pedestrian factor were excluded.

III. Methodology and Research Data

PTV VISSIM is the software which has the function of simulates various scenes of traffic operation [6]. The software is able to simulate the traffic situation either by 2D or 3D. In order to replicate the traffic flow at the study areas, real data were needed. Table I shows specific parameters used as an entry data for this study. To simulate the traffic flow at study site, three groups of data were mandatory to be keyed in the software database were;(1) road geometry; (2) traffic; and (3) signal control data.

In this study, traffic volume study was performed by conducting video recording to record vehicle movements of all traffic directions at the intersection. The data were recorded in two peak hours sessions: two hours in the morning (7:00 am to 9:00 am) and two hours in the evening (5:00 pm to 7:00 pm) for five days during the weekday. Data collected were including number of vehicle movement based on vehicle types and direction of all vehicle movements. As principally required by the traffic flow analysis study, traffic volume count had been conducted according to an hourly basis during data collection period (throughout the 2 session of data collection). However, for data analysis, only 1 hour of any day of uppermost traffic volume data had been selected to be simulated. The selection of data for analysis was done based on critical traffic situation factor. The consideration was taken because during this situation, maximum traffic problem was expected to be occurred. Fig. 3 shows the summary of 1 hour highest volume traffic observed from the site. From the data gathered, it was found that the highest traffic volume was on Wednesday while Sunday has the lowest volume (shown in Fig. 3). Fig. 4 presents specifically peak hour traffic volume summary data used in this study while Table II summarized specific traffic composition used in the simulation.

TABLE. I. DATA PARAMETERS

Parameter	Data for Simulation
Traffic	- Traffic composition
	-Type of Vehicle
	-Traffic movement direction
Signal Control	-Signal Timing for Each Traffic Signal Group (green, red, amber)
	-Cycle Length of Traffic Signal
Geometry	-Type of Road
	-Type of Intersection
	-Number of Lanes in each approach
	-Lane Width
	-Approach Grade
	-Parking
	-Number of right turn lanes
	-Number of left turn lanes

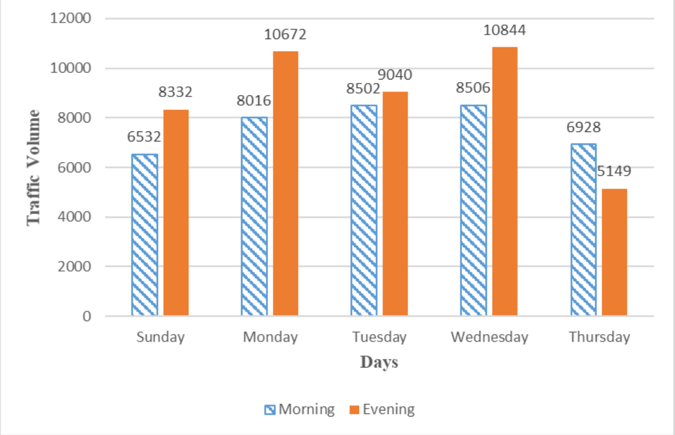


Fig. 3. Showing Graph Traffic Volume at Study Location.

TABLE. II. TRAFFIC COMPOSITION GATHERED FROM STUDY LOCATION

Vehicle movement	**From Batu Pahat**				**From Kluang**				**From Parit Raja Laut**			**From Parit Raja Darat**		
Movement Direction / Type of Vehicle	Turn Right	Through	Turn Left	U-Turn	Turn Right	Through	Turn Left	U-Turn	Turn Right	Through	Turn Left	Turn Right	Through	Turn Left
Car	244	1200	180	64	112	904	96	80	156	112	16	212	60	216
Motorcycle	164	332	80	32	52	192	108	44	60	172	28	60	152	92
Bus	0	8	0	0	0	8	0	0	0	0	0	0	0	0
HGV	8	68	4	0	4	136	8	0	12	0	0	4	4	8
Total	**416**	**1608**	**264**	**96**	**168**	**1240**	**212**	**124**	**228**	**284**	**44**	**276**	**216**	**316**
Volume Percent	17.5	67.5	11.1	4.0	9.6	71.1	12.2	7.1	41.0	51.0	7.9	34.2	26.7	39.1
Total Volume	**2384**				**1744**				**556**			**808**		
% Car	58.7	74.6	68.2	66.7	66.7	73.2	45.3	64.5	68.4	39.4	36.4	76.8	27.8	68.4
% Motorcycle	39.4	20.7	30.3	33.3	31	15.5	50.9	35.5	26.3	60.6	63.6	21.7	70.4	29.1
% Bus	0	0.5	0	0	0	0.65	0	0	0	0	0	0	0	0
% HGV	1.9	4.2	1.5	0	2.4	10.6	3.8	0	5.3	0	0	1.5	1.9	2.5

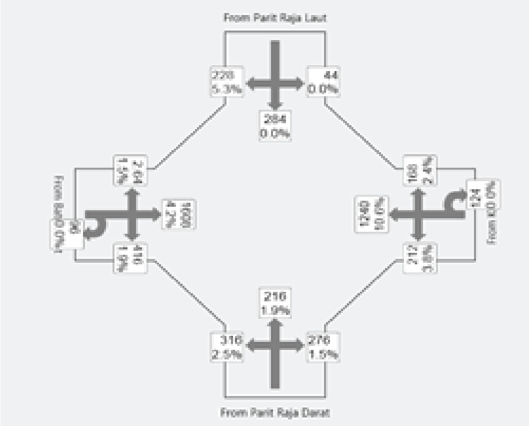
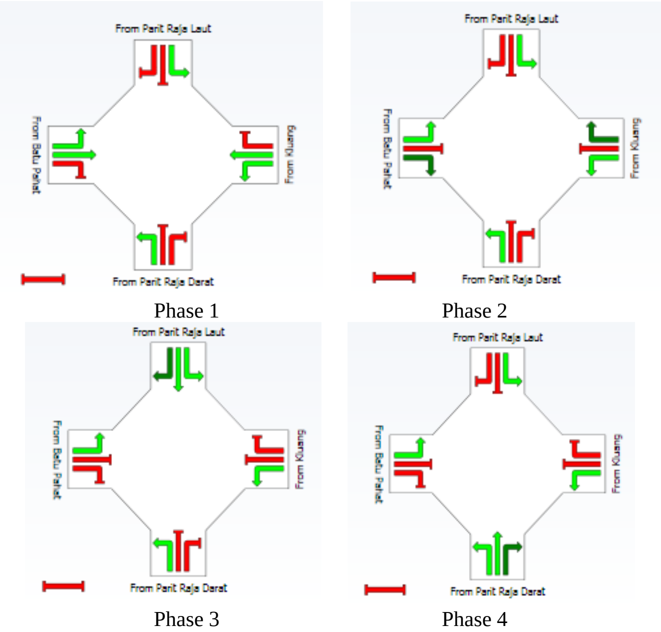


Fig. 4. Summary of Peak Hour Volume at Study Location.



Phase 1

Phase 2

Phase 3

Phase 4

Fig. 5. Traffic Signal Phase for Studied Intersection.

Geometrical data was an important parameter in preparing the simulation in this study. In real condition, geometric data such as number of lanes, lane width, road gradient, parking etc. could influence the capacity and rate of flow. Therefore, to ensure that the simulation represents the actual condition on-site, geometric measurement was performed. Information collected was used as a basis to construct network in the VISSIM. Summary of geometrical data for Parit Raja intersection was shown in Fig. 6, while Fig. 7 shows the layout of the intersection.

Traffic movement at the studied intersection was controlled by pre-timed traffic signal. The signal had been designed with four phases as exhibited in Fig. 5. Specifically, traffic signal cycle time captured on-site was shown in Table III.

TABLE. III. CYCLE TIME ON EACH MOVEMENT PHASE

Phase	All Red	Amber	Red	Green
Phase 1	2	4	89	60
Phase 2	2	4	124	25
Phase 3	2	4	124	25
Phase 4	2	4	124	25

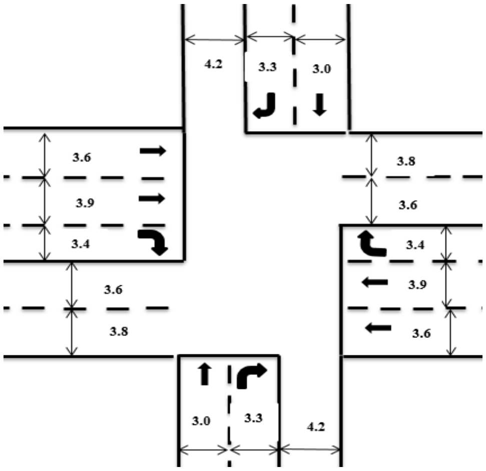


Fig. 6. Lane width of the Studied Intersection.

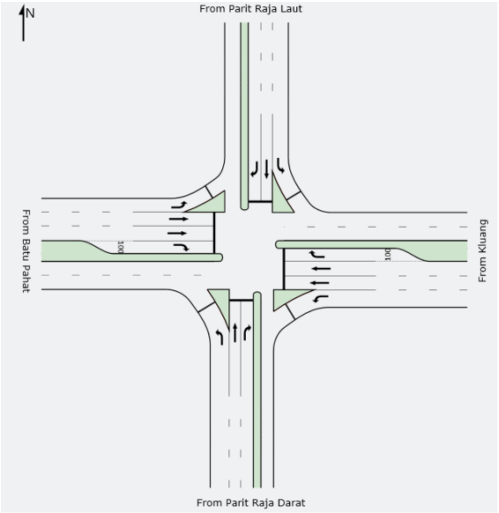


Fig. 7. Intersection Layout at Study Location.

IV. Simulation Model

In this study, traffic flow simulations were simulated using three different random seeds of 32, 42 and 52. The purpose of different random seed was to check the realization of the stochastic quantities in PTV VISSIM [7] such as vehicle capabilities at the study location. Microscopic modeling in this study can produce results such as flow, density, speed, travel and delay time, long queues, stops, pollution, fuel consumption and shock waves [8]. The results from the simulation such as delays and queues for the intersection were used to evaluate the level of service at Parit Raja signalized intersection.

Network and signalized intersection models need to be set before traffic flow simulation could be run or produced. A brief explanation on how to set a few settings in the software before a simulation could be done and how simulation results will be analyzed were explained as follows:

A. Simulation Parameter Setting

Simulation parameter defines the smoothness movement of simulation. Time duration for the simulation was suggested at 4,500 seconds by PTV VISSIM, which was one hour for running the simulations, and 15 minutes for adding input in the software. Simulation resolution was measured as number of steps per second that defines the frequency of vehicle movements in one second of the simulation. 1 to 20 time steps/simulation second was the value that can be used in the simulation resolution. While, 10 to 20 time steps/ simulation second were the suitable values to ensure the movement flow of the simulation and vehicle smooth. In this study, 10 time steps/ simulation second were used. Random Seed could be defined with different values for every simulation running. Simulation needed to be run using different values of random seed to produce better simulation quality. This study used 32, 42 and 52 random seeds and the speed of simulation speed was 10 in the simulation. Simulation speed can be defined as number of simulation seconds compared to the real time second.

B. Network Setting

There are lists in a network setting for simulation such as vehicle behavior, unit of simulation, attribute, display and pedestrian behavior. In this study, only vehicle behavior needed to be set as right-hand traffic. Attribute, display and pedestrian behavior were set as default setting as it was assumed as a standard behavior in this study.

C. Road Link Setting

Field location map for background in the software was imported from Google maps. The background image was set with suitable scale for the model. After setting the scale, road link could be drawn according to the background image. Lane width was entered in the software before the link was drawn. To connect from one link to another, a connector were drawn. All links need to be connected with each other in developing road network.

D. Traffic Based Data

Before the traffic data were entered into a software database, vehicle volume by time interval could be set from 0-900, 900-180 or 900-max. In this study, time interval 0- max was used to collect data from 0-maximum second in the simulation which means the data was collected from the beginning of the simulation until the simulation ended. The data collected were used to determine vehicle queue lengths and vehicle delays at traffic signals. A traffic model needed vehicle volume data per hour as a primary data. Vehicle composition value was entered in the software. Relative flow for each direction could be set according to the percentage data of vehicle for each direction.

E. Signal Control Setting

Signal control needed to be set according to the signal group at the study location. Signal control also needed to be defined according to its type whether it was a fixed time or actuated. However, in a student version of the PTV VISSIM software, signal control type could only be set as fixed time. The dialog box for editing time for signal control group will appear after the edit signal control was clicked. Cycle time, green time, amber and red time have been set based on the actual traffic signal phase time observed from the study area. The advantage of this software was that the sequence of red, amber and red of signal control also could be reformed for further analysis.

V. Result and Discussion

From VISSIM simulations, various results could be obtained to show intersection performances such as vehicle delays and queues. In this analysis, the random seeds used were 32, 42 and 52. Time interval for the result was 100 for every 600 second.

Fig. 8 shows 2D graphical result of the existing traffic condition (with the presence of illegal parking) while Fig. 9 shows the condition if illegal parking was eliminated.

Vehicles in the simulation for this study had been prepared with different colors to characterize four types of vehicles: (1) red for cars, (2) yellow for motorcycles, (3) green for buses, and (4) blue for heavy good vehicles. The white cars in Fig. 8 represented as cars that were parked illegally. Parking lot was added in the simulation to represent illegal parking areas.

Tables IV and V show the analysis outcome of delay and travel time of the existing traffic (with the presence of illegal parking). Average delay stop for this simulation was 44.15 second and average travel time for the vehicle was 139.62 second. This showed that, each driver had to wait at least 2 to 3 minutes to cross the intersection.

Table VII shows the queue length results for every random seed on each traffic direction. The highest queue length was from Batu Pahat to Kluang with 74.74m length and the shortest queue length was from Kluang to Parit Raja Laut with 2.6m distance.

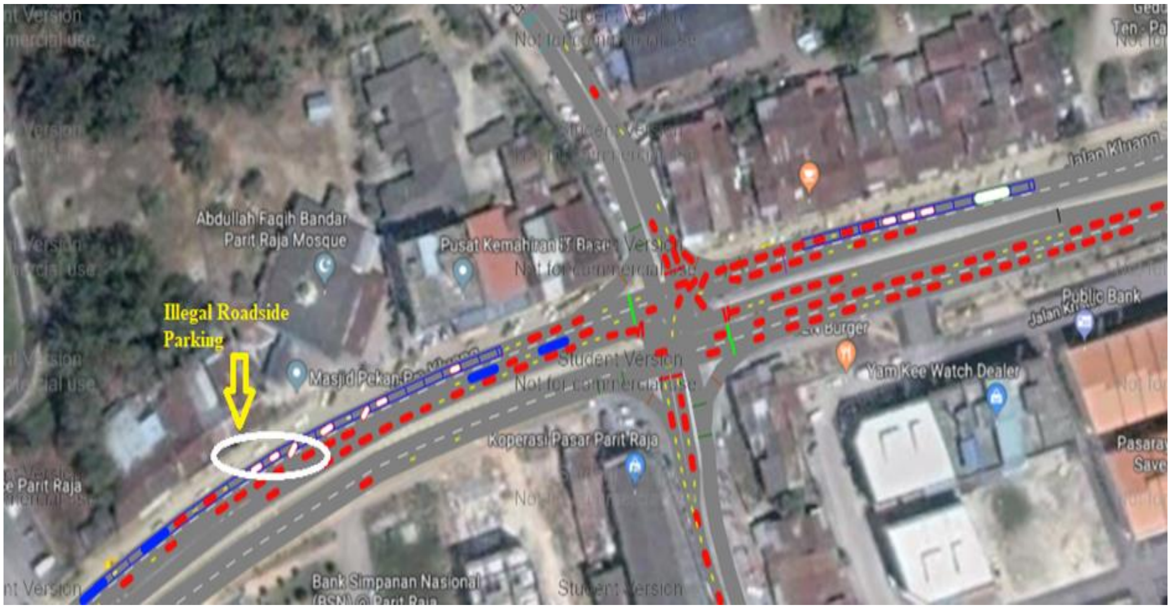


Fig. 8. Simulation of Signalized Intersection Condition at Study Location for Existing Condition (with Illegal Parking).

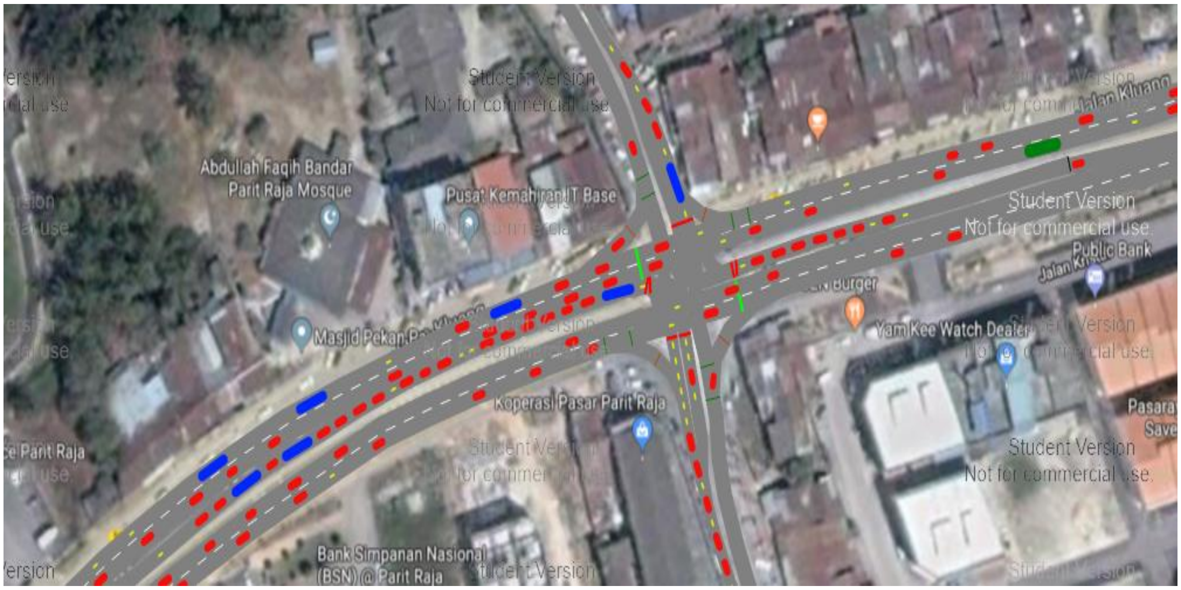


Fig. 9. Simulation of Signalized Intersection Condition at Study Location without Illegal Parking.

TABLE. IV. STOP DELAY OF VEHICLE WITH THE PRESENT OF ILLEGAL PARKING

Random seed (32)	**Random seed (42)**	**Random seed (52)**	**Average**
44.28	42.03	46.15	44.15

TABLE. V. TRAVEL TIME OF VEHICLE WITH THE PRESENT OF ILLEGAL PARKING

Random seed (32)	**Random seed (42)**	**Random seed (52)**	**Average**
134.28	149.41	135.18	139.62

TABLE. VI. SHOWS HIGHWAY CAPACITY MANUAL TABLE OF LOS AT INTERSECTION

LOS	**Control Delay** **Sec/veh (signalized)**	**Delay** **Sec/veh (unsignalised)**
A	≤10	≤10
B	10-20	10-15
C	20-35	15-25
D	35-55	25-35

TABLE. VII. ARE SHOWING QUEUE LENGTH RESULT FOR TRAFFIC FLOW SIMULATION

Vehicle movement		**Queue length(m)**
From Parit Raja Darat	To Kluang To Parit Raja Laut To Batu Pahat	24.13 24.13 47.86
From Batu Pahat	To Kluang To Parit Raja Darat To Parit Raja Laut	74.74 5.07 34.59
From Parit Raja Laut	To Kluang To Parit Raja Darat To Batu Pahat	23.02 14.67 14.67
From Kluang	To Parit Raja Laut To Parit Raja Darat To Batu Pahat	2.60 20.66 70.74

Further analysis, level of service at the Parit Raja intersection could be evaluated by referring to the Highway Capacity Manual (HCM) [9] level of service on Table VI. Delay result from the simulation was used as a base to obtain LOS based on HCM. From the simulation results, delay of 44.15 for the studied intersection falls in the category of LOS D. This showed that the intersection was approaching unstable flow (tolerable delay, occasionally wait through more than one signal cycle before proceeding).

The impact of illegal parking elimination on Parit Raja intersection: In Parit Raja, illegal parking has become norm for local people which, then created problems with the surrounding traffic. Business activities on the road curbs by the road sides encourage illegal parking in Parit Raja particularly at the areas near to the intersection.

The study conducted had found that, on-road illegal parking had influenced the traffic flows particularly for left turn from Batu Pahat approach and Parit Raja Laut approach. This was due to either the reduction of the effective lane width or the lane has been taken by illegal vehicle parking that interrupted the vehicle movements. Furthermore, illegal parking activities near to the study location was also seen to disrupt the traffic through additional traffic delays due to stop delays caused by searching for parking or exiting parking. Fig. 10 shows that traffic to Ayer Hitam direction interrupted due to a cars stopped at the middle of the road. This situation happened because all cars needed to wait for a car at the illegal parking to enter the traffic.

Tables VIII and IX show the comparison of vehicle delay and travel time obtained from the simulation with and without the presence of illegal parking. The result had shown that, if illegal parking could be removed from the site, stop delay on each vehicle could be improved up to 21%. This means traffic flows at the studied intersection could be increased. In addition, the result presented in Table IX, proved that, illegal parking was causing at least 20% of which was 47.17 seconds more than when illegal parking was eliminated in a driver travel time.

Table X shows the comparison of LOS of simulation result with presence of illegal parking, and without illegal parking. The result shows that, without the presence of illegal parking, LOS of intersection can be improved from D to C.

TABLE. VIII. COMPARISON RESULTS OF STOP DELAY BETWEEN SIMULATION TRAFFIC FLOW WITH AND WITHOUT ILLEGAL PARKING

Condition	**Without illegal parking**	**With illegal parking**
Random seed (32)	34.80	44.28
Random seed (42	32.74	42.03
Random seed (52)	36.96	46.15
Average	34.83	44.15

TABLE. IX. COMPARISON RESULTS OF TRAVEL TIME BETWEEN SIMULATION TRAFFIC FLOW WITH AND WITHOUT ILLEGAL PARKING

Condition	**Without illegal parking**	**With illegal parking**
Random seed (32)	92.11	134.28
Random seed (42	94.40	149.41
Random seed (52)	90.83	135.18
Average	92.45	139.62

TABLE. X. COMPARISONS OF LOS WITH AND WITHOUT PRESENT OF ILLEGAL PARKING

Condition	**Delay result (second)**	**Level of service (second)**
Without illegal parking	34.83	C(>20-35)
With illegal parking	44.14	D(>35-55)

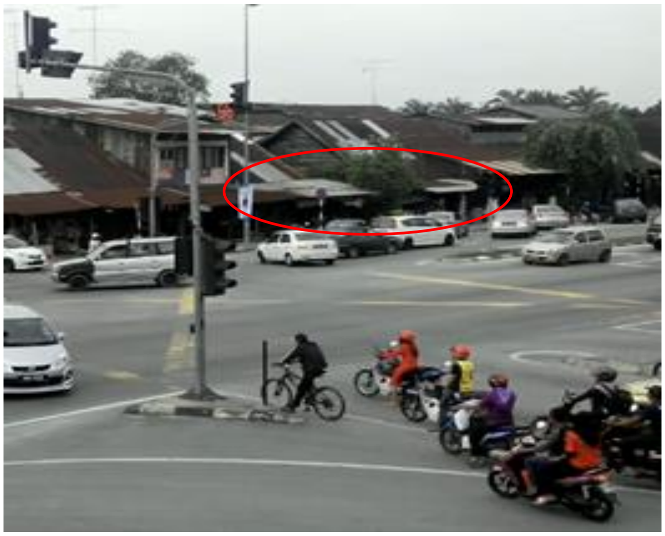


Fig. 10. Showing the Traffic Interruption due to Car Exit from Illegal Parking on Road.

VI. Conclusion

The road network and base data were collected on-site and have been used in developing the simulation model using PTV VISSIM software. This was to ensure that the model could replicated the actual traffic condition. Other than that, PTV VISSIM needed to be calibrated and validate to make the data reliable [10].

Based on the simulation results and video recording of the study location in this study, illegal parking contributed to traffic flow disturbances. Removing illegal parking will help to upgrade the LOS, thus, contributes to smooth traffic flow. This was shown from the improvement of vehicle delay from 44.15 to 34.83 second. Besides that the travel time also has been found to improve by 20% which is 47.17 second less from when the illegal parking existed.

Acknowledgment

The authors gratefully acknowledge the contribution of the Ministry of Education and Universiti Tun Hussien Onn Malaysia. This paper was prepared as part of the research report under Research Grant of GPPS Vot H012.

References

[1] J,Aderamo,A. Traffic Congestion at Road Intersections in Ilorin,Nigeria. Mediterranean Journal of Social Sciences, Volume 3(2),(2012),pp. 201-213.

[2] Washington State Department of Transportation Manual.(2015). Chapter 1300-Intersection Control Type-Design Manual M22-01. In Intersection Control Type Design Manual (pp.1-20).

[3] Yan, X.,Radwan,E. and Abdel-aty,M. Characteristics of rear-end accident at signalized intersection using multiple logistic regression model. Accident Analysis and Prevention, Volume 37,(2005), pp.983-995.

[4] Morillo Carbonell,C. and Campos Cacheda,J.M. Effect of Illegal On – Street Parking On Travel Times In Urban Environment. CIT2016-XII Congreso de Ingeniera del Transporte Valencia,(2016),pp.2491-2503.Available at: http://dx.doi.org/10.4995/CIT2016.2016.3521.

[5] Boro,D.,M,A,Ahmed.,and Goswami,A.Impact of On Street Parking on Traffic Flow Characteristics.Available at: https://www.academia.edu/25582493/Impact_of_OnStreet_Parking_on_Traffic_Flow_Characteristics? Auto.

[6] Tianzi,C.,Shaochen,J.I.N. and Hongxu,Y. Comparative study of VISSIM and SIDRA on signalized intersection. Procedia-Social and Behavioral Sciences,96(Cictp)(2013),pp.2004-2010.Available at: http://dx.doi.org/10.1016/j.sbspro.2013.08.226.

[7] PTV Group,(2004).VISSIM 4.10 User Manual, North America:PTV Planung Transport Verkehr AG.

[8] Nurul Nasuha,N.A. and Munzilah,M.D. Overview of Application Of Traffic Simulation Model.MATEC Web Conferences,(2018), Available at:https:doi.org/10.1051/matecconf/201815003006.

[9] Highway Capacity Manual,(2000).Signalized Intersection 4th ed., United Stated Transportation Research Board of the National Academies of Science.

[10] Rrecaj,A.A. and M,Bombol,K. Calibration and validation of the VISSIM parameter-state of the art. Technology, Education, Management, Informatics, Volume 4(3),(2015),pp.255-269.

Assessment of IPv4 and IPv6 Networks with Different Modified Tunneling Techniques using OPNET

Asif Khan Babar[1]
University of Sindh
Dadu Campus
Dadu, Pakistan

Zulfiqar Ali Zardari[2], Sirajuddin Qureshi[4], Song Han[5]
Faculty of Information Technology
Beijing University of Technology
Beijing, China

Nazish Nawaz Hussaini[3]
Institute of mathematics and
Computer Sciences and IMCS
University of Sindh
Jamshoro

***Abstract*—Currently, all the devices are using Internet protocol version 4 (IPv4) to access the internet. IP addresses of the IPv4 are now depleted from IPv4 pool announced by IANA (Internet Assigned Number Authority) in February 2011. To solve this issue Internet protocol version 6 (IPv6) is launched. But the main problem is current devices can't support directly IPv6 that causes various compatibility issues. Many researchers have proposed various techniques, but still, their efficiency and performance is a big challenge. This study examines several mechanisms of transition IPv6 the backbone of multiprotocol label switching (MPLS) to evaluate & compare their performances. It involves comparing different performance metrics and manual tunneling tunnel efficiency metrics. The main goal of this paper is to examine the dissimilar tunneling techniques and find out which tunneling method is better in all performance, which increases network performance. Experimental results show that ISATAP is better performance in all metrics.**

Keywords*—ISATAP; tunneling techniques; IPv4; IPv6; network performance

I. Introduction

Due to the rapid growth of population demand for IP addresses has increased more and more [1-3]. Eventually, IP addresses of IPv4 pool is completely exhausted. IANA announced on 3 February 2011 that no. of IPv4 addresses are almost exhausted. Many companies and organizations are moving towards IPv6 addresses. IPv4 is 32 bit long and supports an address of only 32-bit, meaning 4.3 billion. IPv6 is 128 bit long and includes an enormous amount of addresses, i.e. trillions of trillion addresses are now accessible. MPLS is a mechanism for packet labeling [4-7]. It is extremely scalable system and commonly utilized in transmission technology by internet service suppliers. It plays a vital role in the IPv4 backbone network for companies. MPLS examines the labels and forward data packets discovered on the label instead of searching for hard routing and examine the packets. Companies, use the backbone of MPLS to link offices and sites together remotely. The integration of IPv6 facilities in the MPLS infrastructure can be seen as ordinary progress by service suppliers and businesses using MPLS networks [8-10].The MPLS backbone provides the option of connecting IPv6 network, using the existing IPv4 network. When using the IPv4 MPLS backbone current, several ways to connect to IPv6 islands. Because the cost of updating the spine in whole or in part is greater and needs network updates, therefore transition mechanisms are deployed. The theory of MPLS is developed and considered as the hybrid technology of ATM and IP. This paper evaluates distinct techniques for clouting current IPv4 network MPLS additional IPv6 facilities lacking the need for backbone adjustments. These techniques are used to isolate IPv6 domains to interact on the present IPv4 MPLS backbone [11]. In the IPV6 tunnels among customer edge and customer edge CE-to-CE routers together with manual tunnels, ISATAP tunnels, 6to4 tunnels. IPV6 tunnels between supplier edge and supplier edge PE-to-PE routers along with manual tunnels, IPV4 automatic tunnels, ISATAP tunnels, and 6to4 tunnels. This paper analyzes performance parameters, i.e. delays in data packets, jitter, and throughput of the network above mentioned techniques and performs statistical analysis [12-15]. The purpose of doing this study is to investigate the various tunneling mechanisms which run both network IPv4 and IPv6. After the deployment of these mechanisms find out the best transition mechanism that provides the highest throughput with very low delay and jitter in the network. For better understanding, the scenarios are shown below in Fig. 1. The figure shows this research paper consist of four phases. Each phase provides the evaluation of the research intention is acclimated. The emulation is done by Graphical Network Simulator (GNS3) tool. Network simulation and data gathering are done by (OPNET) tool. In the last phase data analysis is conducted by MS-office 2013.

A. Contributions/ Findings

The findings of these studies and contributions described as follows:

- Proposed work covers the shortage of IPv4 addresses and provides full IPv6 connectivity.
- Proposed study assessment of a sequence of IPv6 multi-protocol label switching (MPLS) transition mechanisms.
- Analyzing the transition system and identifying which mechanism is best performed in terms of the smallest delay lowest jitter and the greatest performance.

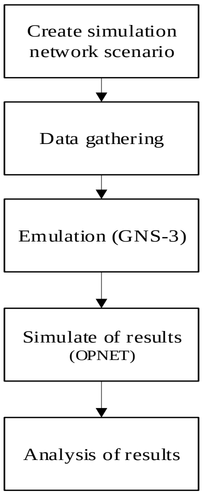

Create simulation network scenario
Data gathering
Emulation (GNS-3)
Simulate of results (OPNET)
Analysis of results

Fig. 1. Proposed Methodology Flowchart.

Our research paper is divided into six sections, the second section of the paper describes some existing techniques and drawbacks, the third section presents the methodology of the proposed analysis, and the fourth section is about the simulation results and their discussion. The fifth section is all about analysis of data through different statistical methods like ANOVA, F-test and T-test, in the last section is the conclusion.

II. Related Research

The various researchers have proposed different tunneling mechanisms, but there are some drawback is still in the network. Following Table I show some tunneling mechanisms and their drawbacks.

TABLE. I. Summary of Previous Work and their Drawbacks

S. No.	Author	Technique	Drawbacks
01	Dr vadym Kaptur	Tunneling, NAT, Dual-stack	Old technique and not ideal
02	Luke smith	Dual-stack and manual tunnel	A circumstance where point-to-multipoint tunnels
03	Zeeshan Ashraf	OSPF V3 in IPV6 tunneling methods	Only focus on OSPF protocol
04	M. S. Ali	Traffic sent from IPV4 network to IPV6 network and only method is used that is 6to4 method in OPNET tool	This research only focuses on one method, but other methods are remaining
05	Sami Salih	6vpe	Onlyfocused delay performance parameters.
06	Yashwin Sookun, Vandana Bassoo	ISATAP, 6RD and Dual Stack Performance Analysis of IPv4/IPv6 Transition Techniques	Network and traffic load is high
07	N. Chuangchunsong at al	DS-Lite, 4over6,	Metrics are not defined clearly
08	Mohammad Aazam et al	Teredo, ISATAP	Tunneling overhead in Teredo

III. Proposed Analysis

For simulation optimized network engineering tools (OPNET) has been used [16]. The Customer edge to customer edge (CE to CE) and the provider edge to provider edge (PE to PE) routers are placed. For configuration of routers well-known emulation is used, called graphical network simulator (GNS3) [17-19]. For simulation, all configurations are imported to the OPNET environment. External Border Gateway Protocol (EBGP) and Multiprotocol Border Gateway Protocol (MP-BGP) are used for PE routers, CE routers for remote access to PE router in MPLS whereas Interior Gateway Protocol (IGP) and Open Shortest Path First (OSPF) are used inside the MPLS. For suitable deployments configured for IPv4 and IPv6 networks, but it also depends on the transition mechanisms. MPLS cloud set to be in IPv4-enabled and IPv6-enabled customers and servers for the transition processes. If customers need to interact with servers on separate islands on each IPv6 island, they must cross the cloud of the IPv4 MPLS. Then a total of eight tunneling scenarios were configured for the various tunneling mechanisms shown in Table II. All customers and servers have IPv6 allowed configuration. These tunneling processes were used to traffic IPv6 throughout the current IPv4 network by encapsulating IPv6 packets in the IPv4 header. Data packet will be decapsulated at the end node of the tunnel, and it will be removed from the IPv4 packet header. An actual data packet of IPv6 transferred to Well-matched Tunnels, and it configured remaining four routers among provider edge routers.

TABLE. II. IPV6 Transition Mechanisms

Manual tunnel	
IPv4 MPLS backbone	Manual tunnel CE to CE
IPv4 MPLS backbone	Manual tunnel PE to PE
Automatic tunnel	
IPv4 MPLS backbone	Manual tunnel CE to CE
IPv4 MPLS backbone	Manual tunnel PE to PE
6to4 tunnel	
IPv4 MPLS backbone	Manual tunnel CE to CE
IPv4 MPLS backbone	Manual tunnel PE to PE
GRE tunnel	
IPv4 MPLS backbone	Manual tunnel CE to CE
IPv4 MPLS backbone	Manual tunnel PE to PE
ISATAP tunnel	
IPv4 MPLS backbone	Manual tunnel CE to CE
IPv4 MPLS backbone	Manual tunnel PE to PE

The CEs were intended to be allowed for IPv4 and IPv6 in the event of CE-to CE tunneling, and only IPv4 was configured for PE routers and all IPv6-enabled client and server configuration. After the encapsulating process of IPv6 data packets in the header of the IPv4 network, these tunneling procedures were used to traffic IPv6 across the existing IPv4 network. The packet will be decapsulated at the tunnel end node and the packet header IPv4 will be deleted. It will then forward the initial IPv6 packet to its final IPv6 place. Next, the

6PE, Native IPv6, and dual-stack transition devices were configured. The MPLS cloud is allowed for 6PE with PE routers supported by IPv4 and IPv6, CE routers supported by IPv6 and PE router supported by IPv4. The MPLS key infrastructure is unaware of IPv6 in the 6PE transition system and to support IPv4/IPv6; only PE routers are updated and 6PE. The 6PE routers use the MP-BGP over IPv4 to exchange accessibility information across the network in a transparent manner.

IV. Experimental Setup and Analysis

In our proposed technique, we have deployed IPv4 and IPv6 networks with MPLS technology. For the simulation of the network, OPNET simulator is used to verify both networks are working smoothly with five routers as shown in Fig. 2. For each of 11 situations, the model was run for five hours with three seeds. Every second, the metrics were set to be gathered, leading in 18000.

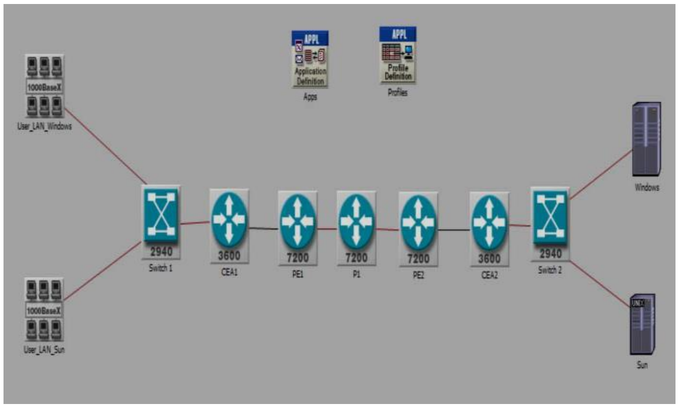


Fig. 2. OPNET Network Simulation.

V. Simulation Setup

The five thousand four hundred values gathered for every metric are highly greater, as shown in the statistics above the numbers are close to one another. As a consequence, the statistical investigation is conducted to define if the mechanisms for each performance metric have any statistically significant distinctions. The statistical analysis conducted to assess the information gathered is described in Section 4.

OPNET Average End-to-End delay for all scenarios as shown in Fig. 3. X-axis is total time in minute’s total running time for simulation is 5 hours and Y-axis is time for delay measured in seconds.

OPNET Average End-to-End jitter for all scenarios as shown in Fig. 4. X-axis is total time in minute’s, total running time for simulation is 5 hours and Y-axis is time for jitter that is delay that is IP delay variation is seconds.

OPNET average End-to-End Throughput for overall simulation as shown in Fig. 5. X-axis is total time in minute’s total running time for simulation is 5 hours and Y-axis is time for throughput of the network that process how many packets can process in a given amount of time.

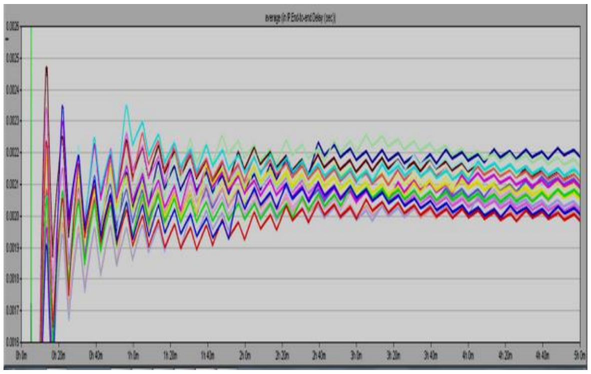


Fig. 3. OPNET Average End-to-End Delay for All Scenarios.

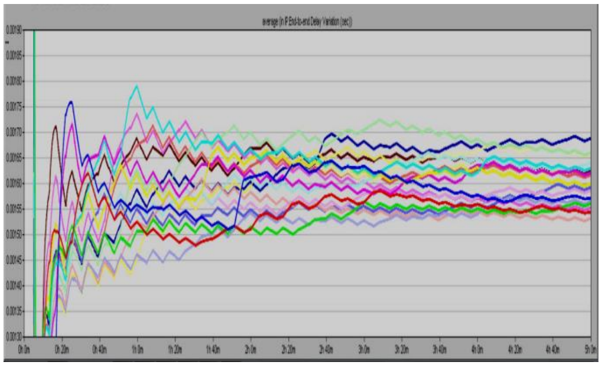


Fig. 4. OPNET Average End-to-End Jitter for All Scenarios.

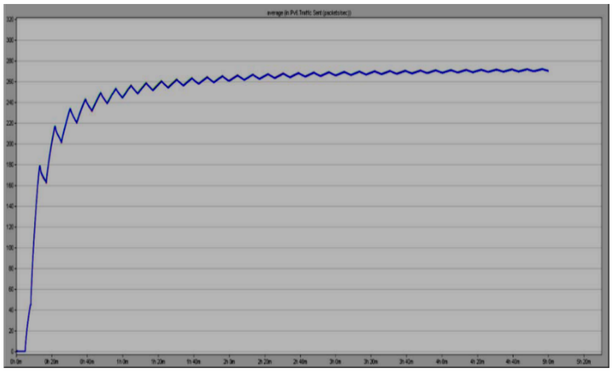


Fig. 5. OPNET Average Throughput for All Scenarios.

VI. Analysis of the Results

Investigation of the collected data and evaluated the resultant data. Methods Below were used to perform the statistical analysis.

- Analysis of variance (ANOVA)
- F-TEST
- T-TEST

ANOVA was used to determine if there is statistically significant difference in means among the scenarios F-Test was used to determine whether modifications were equivalent and whether the mean of one system differed from the mean of the other. Finally, either two sample T-test using degree of freedom was used to determine if the mean of one mechanism is different from the mean of another mechanism.

A. Scenario. 01 Customer Edge to Customer Edge (CE-to-CE)

To calculate significant variations between the metrics of IPv6 CE-to-CE tunneling processes (delay, jitter, throughput) and Calculate which method is the best.

1) Analysis of delay: For ANOVA, the hypothesis below has been recognized. • Null Hypothesis (H0): delay implies equivalent to CE-to-CE tunneling processes Alternative Hypothesis (H1): for CE-to-CE tunneling processes, at least one delay implies different from other means.

TABLE. III. ANOVA RESULTS FOR DELAY CE-TO-CE TUNNEL

ANOVA: Single Factor				
Specified group	***Total***	***Sum (Quantity)***	***Average***	***Difference***
Manual CE-CE	51699	107.4725336	0.0020970	2.44585
Auto CE	51762	106.4065222	0.0020822	2.44083
GRE CE-CE	51758	109.4823187	0.0031417	2.5934
6to4 CE-CE	51792	108.815763	0.0034448	2.53934
ISATAPCE-CE	51750	105.403224	0.0030568	2.33172

(contd.)

Source of variation	***SS***	***DF***	***MS***	***F***	***P-VALUE***	***F-critical***
Between Groups	0.000127451	4	4.24835	16.96009107	5.2	2.37 F test > F critical Reject null Hypothesis
Within Groups	0.519426123	259211	2.50491			
Total	0.519553573	259215				

The result for delay performance metrics from CE-to-CE tunnel is shown in above Table III. where F test>F critical for Seeing that null hypothesis is rejected that reason it is enough evidence that at least one delay mean is different from other delay mean for all scenarios.

2) F-Test for delay CE-to-CE tunnel: F test was used to evaluate if the variances are equal using the hypothesis given below.

Null hypothesis (H0)= delay variance i= delay variance

Alternative hypothesis (H1) = delay variance i=<delay variance j.

The result shown in Table IV since F0>F α n1-1, n2-2 is not true then null hypothesis is not rejected and it is enough evidence that delay variances are equal in that condition we will perform T-test to find which delay mean is less than the others.

3) T-Test for delay CE-to-CE tunnel: The results are shown in Table V, since t0<-tα, n1-1, n2-2 then the null hypothesis was rejected and hence it is enough evidence that delay means of Manual CE-to-CE and Automatic CE-to-CE is less than the 6to4 CE-to-CE and GRE CE-to-CE. Additionally ISATAP CE-to-CE tunnel is less than the GRE CE-to-CE tunnel.

TABLE. IV. F-TEST RESULT FOR DELAY CE-TO-CE TUNNEL

F-TEST	**n1,n2 is the degree of freedom f0=S12/S22**		
Transition Mechanisms	***F α, n1-1, n2-2***	***F0***	***Test if F0>F α n1-1,n2-2***
Auto CE & 6to4 CE	one	0.961206111	Negative
Auto CE & GRE CE	one	0.941160555	Negative
Auto CE & ISATAP CE	one	0.801237	Negative
Manual CE & 6to4 CE	one	0.963112573	Negative
Manual CE & GRE CE	one	0.943143	Negative
Manual CE & ISATAP CE	one	0.85365	Negative
6to4 CE & GRE CE	one	0.979155778	Negative
6to4 CE & ISATAP CE	one	0.89125	Negative

TABLE. V. T-TEST RESULT FOR DELAY CE-TO-CE TUNNEL

Two Sample T-Test			
Transition mechanism	***tα n1+n2-2***	***t0***	***Test if t0<-tα , n1+n2-2***
Auto CE & 6to4 CE	1.63	-4.25017473	YES
Auto CE & GRE CE	1.63	-6.06411939	YES
Auto CE & ISATAP CE	1.63	-3.638063334	YES
Manual CE & 6to4 CE	1.63	-5.453454824	YES
Manual CE & GRE CE	1.63	-1.820066369	YES
Manual CE & ISATAP CE	1.63	-2.639061371	YES
6to4 CE & GRE CE	1.63	-3.999067322	YES
6to4 CE & ISATAP CE	1.63	-2.638067771	YES

4) Analysis of jitter: Jitter is also analyzed using similar techniques. The ANOVA results in Table VI were obtained.

TABLE. VI. ANOVA RSULTS FOR JITTER CE-TO-CE TUNNEL

ANOVA: Single Factor				
Groups	***Count***	***Sum***	***Average***	***Variance***
Manual CE-CE	51753	82.6607	0.001597	4.05
Auto CE-CE	51864	82.52443	0.001591	4.14
GRE CE-CE	51858	85.79141	0.001654	4.28
6to4 CE-CE	51892	84.85763	0.001635	4.37
ISATAPCE-CE	51850	81.33167	0.001568	4.28

(contd.)

Source of variation	***SS***	***Df***	***MS***	***F***	***p-value***	***F-critical***
Between Groups SSB	0.000143	3	4.77	113.4597	2.8	2.604952 F test > F critical Reject Null Hypothesis
Within Groups SSW	0.087257	207363	4.21			
Total SST	0.0874	207366				

5) F-Test for jitter CE-to-CE tunnel: The results shown in Table VII says that there is enough evidence to support that end-to-end jitter means of different tunneling Manual CE-to-CE and Automatic CE-to-CE are less than 6to4 CE-to-CE and ISATAP CE-to-CE tunnel, and 6to4 CE-to-CE has lower mean jitter than ISATAP CE-to-CE.

6) T-Test for jitter CE-to-CE tunnel: The results shown in Table VIII says that there is enough evidence to support that end-to-end jitter means of different tunneling Manual CE-to-CE and Automatic CE-to-CE are less than 6to4 CE-to-CE and ISATAP CE-to-CE tunnel, and 6to4 CE-to-CE has lower mean jitter than ISATAP CE-to-CE.

B. Scenario. 02 Provider Edge to Provider Edge (PE-to-PE)

1) ANOVA results for delay PE-to-PE tunnel: The result for delay PE-to-PE performance parameters are shown in Table IX. Since Ftest >Fcritical therefore the Null Hypothesis is rejected and there is enough evidence to demonstrate that at least one delay mean is different from other delay means among the all scenarios.

TABLE. VII. F-TEST RSULTS FOR JITTER CE-TO-CE TUNNEL

F-TEST	**n1,n2 is the degree of freedom f0=S12/S22**		
Transition Mechanisms	***F α, n1-1, n2-2***	***F0***	***Test if F0>F α n1-1,n2-2***
Auto CE & 6to4 CE	one	0.945501655	Negative
Auto CE & GRE CE	one	0.967344952	Negative
Auto CE & ISATAP CE	one	0.922723464	Negative
Manual CE & 6to4 CE	one	0.945787554	Negative
Manual CE & GRE CE	one	0.976769838	Negative
Manual CE & ISATAP CE	one	0.944407759	Negative
6to4 CE & GRE CE	one	0.956819337	Negative
6to4 CE & ISATAP CE	one	0.945066522	Negative

TABLE. VIII. T-TEST RSULTS FOR JITTER CE-TO-CE TUNNEL

Two Sample T-Test			
Transition mechanism	***tα n1+n2-2***	***t0***	***Test if t0<-tα , n1+n2-2***
Auto CE & 6to4 CE	1.63	-10.88907171	YES
Auto CE & GRE CE	1.63	-15.68834028	YES
Auto CE & ISATAP CE	1.63	-11.22907883	YES
Manual CE & 6to4 CE	1.63	-9.441690775	YES
Manual CE & GRE CE	1.63	-14.25657544	YES
Manual CE & ISATAP CE	1.63	-8.88922133	YES
6to4 CE & GRE CE	1.63	-4.672126496	YES
6to4 CE & ISATAP CE	1.63	-9.889071761	YES

TABLE. IX. ANOVA RSULTS FOR DELAY PE-TO-PE TUNNEL

Summary				
Groups	***Count***	***Sum***	***Average***	***Variance***
Auto PE	52022	107.5243	0.001591	2.45
Manual PE	52022	107.5243	0.001597	2.45
GRE PE	51938	108.4176	0.001654	2.39
6to4 PE	51938	108.4080	0.001635	2.39
ISATAP PE	52010	106.5133	0.001568	2.57

(contd.)

Source of variation	***SS***	***df***	***MS***	***F***	***p-value***	***F-critical***
Between Groups SSB	0.000127451	3	4.24835 E-05	16.96009107	0.029644	2.604952 F test > F critical Reject Null Hypothesis
Within Groups SSW	0.519426123	207363	2.50491 E-06			
Total SST	0.519553573	207366				

2) F-Test results for delay PE-to-PE tunnel: The result is shown in Table X, since F0>F α n1-1, n2-2 is not true then null hypothesis is not rejected and it is enough evidence that delay variances are equal in that condition we will perform T-test to find which delay mean is less than the others.

3) T-Test results for delay PE-to-PE tunnel: The result is shown in Table XI, since t0<-tα, n1-1, n2-2 then the Null Hypothesis was rejected. Therefore there is enough evidence to support that end-to-end delay mean of Manual PE-to-PE and Automatic PE-to-PE are less than the 6to4 PE-to-PE and GRE PE-to-PE. Additionally it demonstrates that ISATAP PE-to-PE tunnel is lower delay mean as compare to GRE PE-to-PE tunnel.

TABLE. X. F-TEST RSULTS FOR DELAY PE-TO-PE TUNNEL

F-TEST			
Transition mechanism	***Fα n1-1 n2-1***	***F0***	***Test if F0 > Fα n1-1 n2-1***
Auto PE & 6to4 PE	one	1.024523462	YES
Auto PE & GRE PE	one	1.024800865	YES
Auto PE & ISATAP PE	one	1.02525635	YES
Manual PE & GRE PE	one	1.024523462	YES
Manual PE & ISATAP	one	1.024800865	YES
Manual PE & 6to4 PE	one	1.024785256	YES

TABLE. XI. T-TEST RSULTS FOR DELAY PE-TO-PE TUNNEL

Two sample T-TEST			
Transition mechanism	***Tα ,a***	***t0***	***Test if t0<-t α ,a***
Auto PE & 6to4 PE	1.63	-2.10904	YES
Auto PE & GRE PE	1.63	-2.1273	YES
Auto PE & ISATAP PE	1.63	-2.13886	YES
Manual PE & 6to4 PE	1.63	-2.10904	YES
Manual PE & GRE PE	1.63	-2.12737	YES
Manual PE & ISATAP	1.63	-2.11678	YES

4) ANOVA Results for jitter PE-to-PE tunnel: The result is shown in Table XII, since Ftest<Fcritical the Null Hypothesis was accepted. Therefore there is enough evidence to show that there is no statistically-significant difference among the jitter means of PE-to-PE tunneling mechanisms.

5) ANOVA results for overall throughput: The result of throughput of overall system is shown in Table XIII that shows there is not a significant difference among all the mechanisms.

TABLE. XII. T-ANOVA RSULTS FOR JITTER PE-TO-PE TUNNEL

Summary				
Groups	***Count***	***Sum***	***Average***	***Variance***
Manual PE	52021	83.55317109	0.001606143	4.47442
Auto PE	52021	83.55317109	0.001606143	4.47442
GRE PE	51937	83.54239534	0.001608533	4.09939
6to4 PE	51937	83.59333751	0.001609514	4.09323
ISATAP PE	52010	83.40317009	0.001603592	4.09221

(contd.)

Source of variation	***SS***	***df***	***MS***	***F***	***p-value***	***F-critical***
Between Groups SSB	4.56307E-07	3	1.52102E-07	0.354921286	0.785586345	2.604951992 F test <F critical Accept Null Hypothesis
Within Groups SSW	0.089101094	207912	4.28552E-07			

TABLE. XIII. T-ANOVA RSULTS FOR JITTER PE-TO-PE TUNNEL

Summary				
Groups	***Count***	***Sum***	***Average***	***Variances***
Auto PE	54000	1.40654	2604703	4.27
GRE CE	54000	1.40721	260533	4.3
GRE PE	54000	1.40671	2605027	4.31
Manual CE	54000	1.40645	2604539	4.34
Manual PE	54000	1.40654	2604703	4.27
6to4 CE	54000	1.40616	2603998	4.31
Auto CE	54000	1.4059	2603527	4.32
6to4 PE	54000	1.40671	2605027	4.3
ISATAP CE	54000	1.41517	2613775	4.32
ISTAP PE	54000	1.41622	2632852	4.36

(contd.)

Source of variation	***SS***	***Df***	***MS***	***F***	***p-value***	***F-critical***
Between Groups	2.49	10	2.49	0.005791	1	1.83072 FTEST<FCRITICAL Accept Null Hypothesis
Within Groups	2.55	593989	4.3			
Total SST	2.55	593999				

C. Summarized Result

The statistical analysis for delay, jitter, and throughput was performed to identify if there is a statistically-significant difference among these scenarios and if so to determine which one(s) are the superior methods, in the order of best to worst. The detailed analysis is described in the above Analysis.

The results for delay including the ordinal ranking values are shown in Table XIV.

1) Lowest to highest delay ipv6 transition mechanism: The results for delay including ordinal ranking values are shown in Table XIV shows that ISATAP PE having lowest delay and 6to4 CE is highest delay.

2) Lowest to highest jitter ipv6 transition mechanism: The results for delay including ordinal ranking values are shown in Table XV shows that ISATAP PE having lowest jitter and 6to4 CE is highest delay.

For throughput, the analysis shows that there is no statistically significant difference among the all mechanisms. Next the main objective of this research is analyzed, which is to rank the aforementioned IPv6 transition mechanisms from best to worst as shown in below Table XVI. The best mechanism offers lowest delay, lowest jitter, and highest throughput.

3) Best to worst overall ipv6 transition mechanism: The result shows that ISATAP PE has the best overall performance metrics with lowest delay lowest jitter and highest throughput.

TABLE. XIV. LOWEST TO HIGHEST DELAY IPV6 TRANSITION MECHANSIM

IPv6 Transition Mechanisms in Order of Lowest to Highest Delay	**Ordinal Ranking Value**
ISATAP PE Delay (0.00204793) jitter (0.001603592) Throughput (2632852)	1
Manual PE-to-PE and Automatic PE-to-PE	2
6to4 PE-to-PE and GRE PE-to-PE	4
Manual CE-to-CE and Automatic CE-to-CE	6
6to4 CE-to-CE and ISATAP CE	8
6to4 CE (0.002133757)	10

TABLE. XV. LOWEST TO HIGHEST JITTER IPV6 TRANSITION MECHANSIM

IPv6 Transition Mechanisms in Order of Lowest to Highest Jitter	**Ordinal Ranking Value**
ISATAP PE (0.001603592)	1
Manual CE-to-CE and Automatic CE-to-CE	2
Manual PE-to-PE, Automatic PE-to-PE, 6to4 PE-to-PE, and GRE PE-to-PE	4
6to4 CE-to-CE and ISATAP CE	8
GRE CE (0.001654)	10

TABLE. XVI. BEST TO WORST OVERALL IPv6 TRANSITION MECHANISM

Overall (including delay, jitter, and throughput) IPv6 Transition Mechanisms in Order of Best to Worst	**Ordinal Ranking Value**	**Overall Ranking**
ISATAP PE Delay (0.00204793) jitter (0.001603592) Throughput (2632852)	1	1
Manual PE-to-PE and Automatic PE-to-PE	13	6
Manual CE-to-CE, Automatic CE-to-CE, 6to4 PE-to-PE, and GRE PE-to-PE	14	7
6to4 CE-to-CE , ISATAP CE	21	10
GRE CE-to-CE	23	11

VII. CONCLUSION

This paper has two phases of contribution, i.e., connectivity of IPv4 and IPv6; secondly, test performance of different tunneling techniques. From the above simulation test result, ISATAP PE is best because of the high throughput and lowest jitter during data packets transmission. Whereas GRE CE is worst due to its high jitter and lowest throughput in the network. The main objective is to provide IPv6 connectively and test which tunneling technique is better to have better performance than others. Future work can extend in payload of the network. Additionally security of these tunneling techniques can be analyzed.

REFERENCES

[1] R. Tadayoni,and A. Henten, ‘From IPv4 to IPv6: Lost in translation?’, In Telematics and Informatics, vol. 33, the year 2016, Issue 2, pp 650-659.

[2] L. Smith, M. Jacobi and S. Al-Khayatt, "Evaluation of IPv6 transition mechanisms using QoS service policies", 2018 11th International Symposium on Communication Systems, Networks & Digital Signal Processing (CSNDSP), Budapest, Hungary, 2018, pp. 1-5.

[3] M. Nikkhah, “Maintaining the progress of IPv6 adoption”, In Computer Networks, vol 102, 2016, pp 50-69.

[4] N. Zhang, M. A. Sirbu and J. M. Peha, "A comparison of migration and multihoming support in IPv6 and XIA", 2017 International Symposium on Networks, Computers, and Communications (ISNCC), Marrakech, 2017, pp. 1-8.

[5] Kamaldeep, M. Malik and M. Dutta, "Implementation of single-packet hybrid IP traceback for IPv4 and IPv6 networks", IET Information Security, vol. 12, no. 1,year 2018, pp. 1-6.

[6] D. R. Al-Ani, A. R. Al-Ani, “The Performance of IPv4 and IPv6 in Terms of Routing Protocols using GNS 3 Simulator”, Procedia Computer Science, Vol 130, 2018, Pages 1051-1056.

[7] J. Beeharry and B. Nowbutsing, "Forecasting IPv4 exhaustion and IPv6 migration", 2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), Balaclava, 2016, pp. 336-340.

[8] Y. Sookun,and V. Bassoo, “Performance analysis of IPv4/IPv6 transition techniques”,2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), Balaclava, year 2016 pp. 188-193.

[9] R. Z. Khan, and A. Shiranzaei, "IPv6 security tools—A systematic review", 2016 International Conference on Computing, Communication and Automation (ICCCA), Noida, 2016, pp. 459-464. The Illustrated.

[10] Network How TCP/IP Works in a Modern Network Book2nd Edition year 2017 by Walter Goralski.

[11] V. Kher, A. Arman and D. S. Saini, "Hybrid evolutionary MPLS Tunneling Algorithm based on high priority bits," *Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015 International Conference on*, Noida, 2015.

[12] The Illustrated Network How TCP/IP Works in a Modern Network Book2nd Edition year 2017 by Walter Goralski.

[13] Hamarsheh, and Ala,“Deploying IPv4-only Connectivity across Local IPv6-only Access Networks”, IETE Technical Review Taylor & Francis year 2018, pp. 1-14.

[14] Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), Balaclava, 2016, pp. 336-340.

[15] Y. Sookun,and V. Bassoo, “Performance analysis of IPv4/IPv6 transition techniques”,2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), Balaclava, year 2016 pp. 188-193.

[16] R. Z. Khan, and A. Shiranzaei, "IPv6 security tools—A systematic review", 2016 International Conference on Computing, Communication and Automation (ICCCA), Noida, 2016, pp. 459-464.

[17] The Illustrated Network How TCP/IP Works in a Modern Network Book2nd Edition year 2017 by Walter Goralski.

[18] Hamarsheh, and Ala,“Deploying IPv4-only Connectivity across Local IPv6-only Access Networks”, IETE Technical Review Taylor & Francis year 2018, pp. 1-14.

[19] F. Siddika, and M. A. Hossen , and S. Saha, “Transition from IPv4 to IPv6 in Bangladesh: The competent and enhanced way to follow”, 2017 International Conference on Networking, Systems and Security(NSysS), Dhaka, year 2017, pp. 174-179.

Security and Privacy Awareness: A Survey for Smartphone User

Md. Nawab Yousuf Ali[1], Md. Lizur Rahman[2], Ifrat Jahan[3]
Department of Computer Science & Engineering
East West University, Dhaka 1212
Bangladesh

***Abstract*—Smartphone becomes one of the most popular devices in last few years due to the integration of powerful technologies in it. Now-a-days a smartphone can provide different services as like as a computer provides. Smartphone holds our important personal information such as photos and videos, SMS, email, contact list, social media accounts etc. Therefore, the number of security and privacy related threats are also increasing relatively. Our research aims at evaluating how much the smartphone users are aware about their security and privacy. In this study, firstly we have taken a survey for smartphone users to access the level of smartphone security awareness displayed by the public. We also determine whether a general level of security complacency exists among smartphone users and measure the awareness of android users regarding their privacy. From survey result we have found that, most of the people are not aware about their smartphone security and privacy. Secondly, based on survey results, we have shown a method to measure the level of awareness (LOA) for the smartphone users. By using this method, a user can easily measure his/her smartphone security and privacy related level of awareness.**

Keywords—Smartphone; Smartphone Problems; Level of Awareness (LoA); Security and Privacy

I. Introduction

The technologies of smartphone have been increasing with a huge rate over last few years. Smartphone provides many services as data sharing, phone calls, internet, different online & offline games etc. Therefore, it increases the chance of security and privacy related threats comparatively. Almost 80% of activities related to the internet, so it is important for us to become aware about security and privacy. Several recent studies shown that, when security comes to smartphone, most of the smartphone users are propitious [1, 2, 3]. In order to authentication of smartphone, people often use different patterns, finger print password, face password, pin passwords etc. All these are not enough to protect us from security related issues [4]. Smartphones are handhold device where different personal information are stored. We have to ensure the security of our personal information. Most of the time, due to lack of our awareness we fail to protect our personal information. If all this information falls into a bad hand, we might be in trouble.

According to a recent study, Google play published more than 3.5 million apps from 2009 to December, 2017 [5]. The number of apps is rapidly increasing over recent few years. Another recent security study showed that, in Google play store, more than 200 malevolent apps were found [6]. These apps collected private information like contact numbers, places etc. from users and sent to the attackers' server. Time to time this information was resending to the attackers' server when users use these apps. In the early 2016, Google banned 13 apps from Google play store because, these apps collected information from users and sell to other server [7].

In this paper, we discuss about results of a security and privacy awareness survey for the smartphone users. The research aims at evaluating how much the smartphone users are aware about their security and privacy. In this survey, we create questionnaire to access the level of smartphone security awareness displayed by the public. We determine whether a general level of security complacency exists amongst smartphone users and based on these result we show a statics model to measure the awareness of android users regarding their privacy.

This paper is organized as follows. We start with a discussion of the various previous related works in Section II. Then we explain about the smartphone problems in Section III and discuss different types of attacks in smartphone. In Section IV, we focus on our research methodology along with pilot study, research instrument and target population, and data analysis. In Section V, we analyze the result of our survey including research questions, evolution of research question and then propose a model that can measure the level of awareness. Finally, we show some concluding remarks and future direction in Section VI.

II. Previous Work

Benenson et al. [8] pointed that IT security plays an important role while someone use smartphone, because of its' broadly acknowledged and well documented feature, which mainly focused on the technical area of a smartphone security system. According to their interview of 24 users on IT security of smartphone, they found the role of user. Based on this result they consecrated five hypotheses and proposed a mental technique after evaluation of these hypotheses.

A recent study in South Africa by Ophoff & Robinson [9] shown that the level of awareness on smartphone security based on public users and determined how much a common security level exists in smartpnone users. According to their survey on smartphone security awareness, they examined 619 South African smartphone users based on the trust of smartphone apps and other third party apps. They found that users showing high level of trust on smartphone apps, rather than when they install other third party apps. In this study, they used an updated version of model developed by Mylonas et al. [10].

Alani [11] noted that android smartphone privacy awareness concern grow with spread in users' perspective. A huge number of apps are downloaded daily by the users, but it is really difficult to differentiate between good terms of service security apps and bad terms of service security apps. In this paper, authors shown a result based on a survey of 4027 android smartphone users for android user security awareness. According to their survey, they tried to show the interactions between users and terms and service security while they install apps.

In a recent study by Mylonas et al. [10] pointed out that when a user installs different third party apps from official apps store-house (e.g., Play store, Google play, Apple's app stores etc.), the risk of smartphone security may increase because sometimes the protected information might be accessed by third party apps. According to their survey, they tried to find out whether users aware about their security of smartphone while they downloaded and installed apps form apps store house. Based on their survey, they developed a model that can identify these users who trust apps storehouse.

Zaidi et al. [12] pointed that due to the advanced technologies, smartphone has become a daily necessary component, and also the chance of security based attacks has increased. In this study, authors' discussed about the different threats in smartphone, security based attacks in smartphone and also the solutions to solve these problems. New attack and old attack are the two types of security-based attacks. According to this study, authors provide a simple view of various smartphone security related attacks, and also provide the possible solutions for these attacks to improve the security of smartphone.

III. Smartphone Problems

The technologies of smartphone have been increasing with a huge rate over last few years. Now-a-days a smartphone can provide different services as like as a computer can provide. Our smartphone holds much information such as mailing information, messaging information, calling information etc., which are very important for us. Therefore, we have to ensure the security and privacy of our smartphone.

The addition of powerful OSs, applications, hardware etc., makes smartphone strong and secure, but all these are not enough to protect our privacy. As the number of privacy and security related threats are raising comparatively. The security and privacy related challenges in smartphone are slightly same as the computer threats environment. Smartphone problems are categorized into four parts [12] including: Data protection and privacy, Attacks, Authorization, and Vulnerabilities (Fig. 1).

A. Data Protection and Privacy

Muslukhov *et al.* [13] found out the problem of data protection and privacy and discussed the types of data a user wants to protect in smartphone. Authors' also showed for the different types of data how the required security protection is change. In another recent study, Muslukhov [14] discussed about data protection and privacy problem and showed that the regular update of smartphone lock screen for users' authentication and accessibility creates the security and protection level more strong.

B. Attacks

Attacks are similar in all smart devices such as smartphone, laptop, tablet etc. Attacks in smartphone categorized into two parts including: old attack and new attacks. Old attacks include physical attacks, different type of smartphone virus, backdoor, threats, Trojan, different types of malware, worms, radio and wireless network attacks, and spam attacks. New attacks include relay attack, counter attack, DOS attack, brute force attack, camera based attacks, SMS based attack, XSS attack, control-flow attack, etc.

C. Authorization

Zaidi *et al.* [12] noted that authentication could be getting by three methods. First one is to get the password or code or PIN which is used by actual user for authentication on smartphone. For example, if someone gets your smartphone cleverly and if he/she knows the password or code or PIN which you use, easily can get your personal information from smartphone. Second one is to find which users have used certain code to authentication of his/her smartphone. The third one is to get the fingerprint which is used by users also known as biometric.

D. Vulnerabilities

Vulnerabilities are the weak points of a smartphone, and it causes several different problems such as insecurity of personal information, privacy broken by malicious attackers etc. Users are not much aware about their personal information because, most of the time users' e-mail account, social media account etc. are logged in their smartphone. The vulnerabilities of smartphone contain many parts such as lack of awareness on personal information in smartphone, system fault, insecure apps in smartphone, insecure wireless network etc.

Among all these categories of smartphone problems 'Attacks' is the most common one. Two types of attacks are old attack and new attack. Both attacks have some individual impact to the smartphone. Table I and Table II show the impact of smartphone due to old attack and new attack.

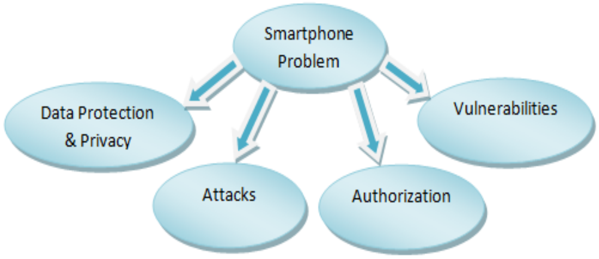


Fig. 1. Categorization of Smartphone Problems.

TABLE. I. OLD ATTACKS AND THEIR IMPACT TO THE SMARTPHONES

Attack Name	Impact to the Smartphone
Physical Attack [15]	• Makes the security of smartphone weak • Causes abnormal behavior in smartphone • Unauthorized code can be effect to the users privacy
Smartphone Virus [16,17]	• Causes abnormal behavior in application and smartphone • Private information can be leaked via applications
Backdoor [18]	• Makes the security of smartphone weak • Create a backdoor for smartphone viruses
Threat [19]	• Makes the security of smartphone weak • Data may be hacked • Creates backdoor into private information
Malware [20,21]	• Interfere in smartphone operations • Collects private information
Wireless Attack [22]	• Data may be hacked • Makes the security of smartphone weak • Private information may be leaked.
Spam [23]	• Fill the e-mail inbox with unnecessary information • Decrease the smartphone internet speed • Collect different important information like contact list, message etc.

TABLE. II. NEW ATTACKS AND THEIR IMPACT TO THE SMARTPHONES

Attack Name	Impact to the Smartphone
Counter Attack [24]	• Target information can be accessed
Relay Attack [25]	• Private information may be hacked.
DOS attack [26]	• Slow the network • Busy the smartphone services
Camera based attack [27]	• Makes the security of smartphone weak • Collects users private information
SMS based attack [28]	• Slow the smartphone operations • Collects sensitive information
Control flow attack [29]	• Collect different important information like contact list, message etc. • Memory information can be accessed
Brute force attack [30]	• Slow the CPU speed • Users password may be hacked

IV. METHODOLOGY

The aims of this research at evaluating how much the smartphone users are aware about their security and privacy. Data collection based on industrial survey is the most common process for research project, but this process requires large time to complete, and data analysis is costly [31]. However, a recent study by Couper [32] discussed about the different technologies of data collection, which can be used to analyze the data automatically (e.g. Google form). Another study by Granello *et al.,* [33] pointed that online data collection has become very popular strategy in many research methodologies.

A. Pilot Study

In our study, we have used survey strategy to find the quantitative results. The survey was planned to find out the level of security and privacy awareness among the smartphone users. To understand the topic better on "security and privacy awareness survey for smartphone users" we consulted with many smartphone users and discussed about their smartphone security related problems. We found three types of users. Some users treated their phones as normal phone, although their phones contain smartphone functionalities, they just use their phone for call or SMS related work only. Some user installed different third party apps without knowing the terms and service related conditions. Some users utilize the full smartphone functionalities.

B. Research Instruments and Target Population

An online tool was used here based on the questions to analyze the collected data. This research contains 20 questions and the answers might be one or multiple. All these questions are based on security and awareness of smartphones. Among these questions we have used just 7 in our study, which can fulfill our goal and objectives. Our aim is to evaluating whether smartphone users aware about their security and privacy related issue, and to evaluate how much aware they are. The target population of this study was smartphone users, especially university students of different countries on the age group between 20 to 26 ages. The purpose of this study is to understand the security and privacy awareness from the smartphone users.

C. Data Analysis and Discussion

At first, we set our questionnaires in a Google form. By using this Google form, we have taken survey from university students' age group in between 20 to 26 year. Then we have stored these results in Microsoft Excel format for further use. After completion the survey, we have found how many responses are there, whether everything is okay or not. We also check every necessary question is answered clearly or not, whether the result fulfills our objectives. Then we combined our survey result together and found out our objectives. Since, our problem statement is related to the security and privacy awareness of smartphone and we combine the survey results and try to find the level of smartphone security awareness displayed by public, whether the general level of security exists amongst smartphone users etc. To present our survey results, we use bar chart. In this study, we have used Google form, computer, Microsoft Excel to find out the security and privacy awareness of smartphone.

V. SURVEY RESULTS

In total 3,424 responses recorded in this survey, among them 175 (5.11%) responses were rejected during initial exploration of data analysis because, all required questions were not answered. Of the remaining 3,249 responses are used in this study. We have analyzed the survey results based on seven research questions which have discussed in this section. All these questions are important to find out the awareness of smartphone security and privacy because all these questions are addressed to smartphone problems.

A. Research Questions

The aim of this research is to measure the level of smartphone security awareness displayed by the public. Also to determine whether, a general level of security complacency exists amongst the smartphone users and to measure the awareness of android users regarding their privacy. The research questions are planned in very simple language, which is easy to understand. All these objectives lead to the following questions:

- Q1: For what purpose do you use Smartphone?
- Q2: From where you mostly install applications?
- Q3: Do you ever install third party applications or applications from Unknown sources in your Smartphone?
- Q4: Before installing application do you read application provider's privacy and policy for using application's?
- Q5: Before installing application do you ever read through application's phone access permissions?
- Q6: What authentication system do you use to lock screen for security?

B. *Evaluation of Research Questions*

Q1: Now-a-days smartphone can perform different services as like computer such as email, SMS, location tracking, contact list, stores photos and videos, social media account etc. Q1 is about the purpose of using smartphone to find how many people use all these services in their smartphone. Fig. 2 shows the result of this question, we can see that only 7.3% of the people use smartphone just for communication and they are less insecure than 88.20% of the people who use smartphone for all these activities.

Q2: Since a smartphone provides different facilities such as email, Google drive, SMS, and different social media, etc. It contains a lot of personal information that is very important for us and we should keep these secure. But most of the time we keep our personal accounts (e.g. Email, Facebook, Google drive etc.) logged in to our smartphone. Suppose, someone lost his/her smartphone and if personal accounts logged in to the smartphone, he/she might be lost his/her personal information. Since, our study is about security and privacy awareness we have used this question to find out how much people aware about their security and privacy. Fig. 3 shows the result of this question, and we can see 65.5% people are not aware in this concern.

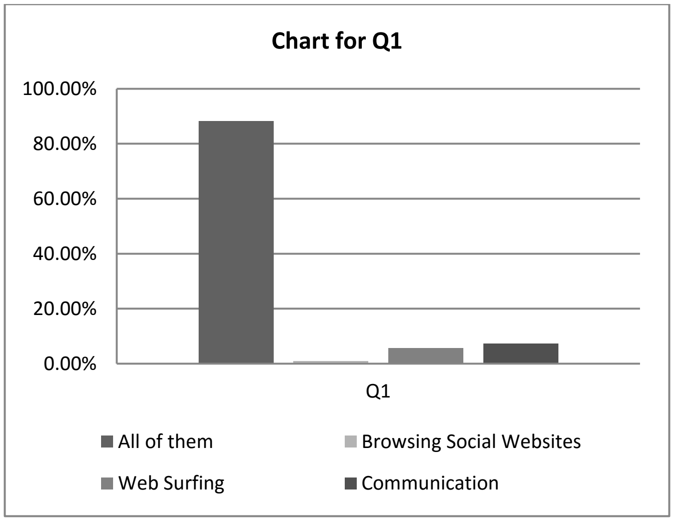


Fig. 2. For What Purpose do you use Smartphone?

Q3: Third party applications are not same as the operating system or manufacture of smartphone, as they are created by vendor. Third party apps contain most of the malware rather than system apps, that's why third party apps are more insecure than system apps. In another scenario, third party apps from unknown sources are more insecure than third party apps from built-in source for system (e.g. play store). In our survey result for question Q3 in Fig. 4, we can see 60% people installed third party apps from unknown sources.

Q4: Before installing apps, the application provider provides the privacy and policy of their apps. This privacy and policy contains about the policy of information about users' access. For example, your application extracts the contact list information from user, so you must have to notify the user about it. From the privacy and policy user can know where, for what and how long his/her information will be used. This is very important and users should read these privacy and policy before installing application. In Fig. 5, we have shown the survey result for Q4, we can easily observe that 25.50% people never read privacy and policy and 52.70% of the people read privacy and policy sometimes.

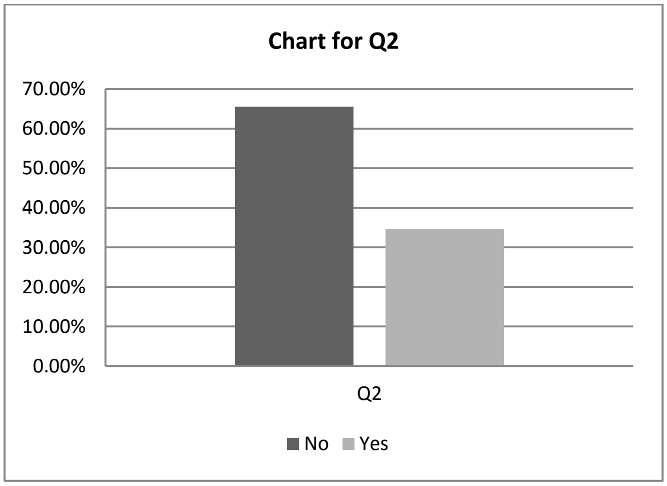


Fig. 3. Do you Sign out from your Personal Accounts (e.g. Email, Facebook, Google Drive Etc.) after using it with Smartphone?

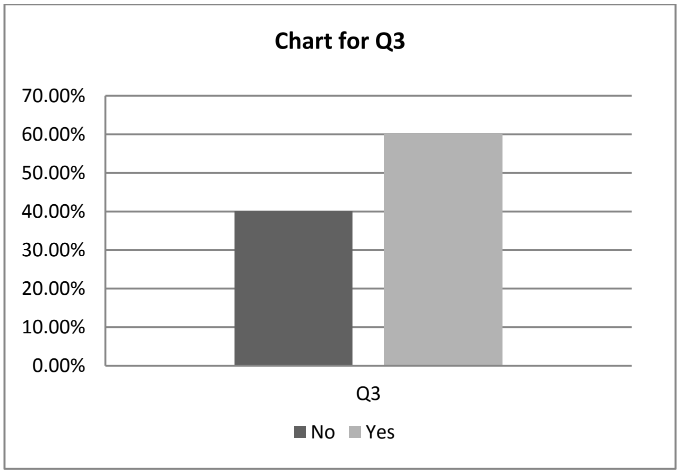


Fig. 4. Do you ever Install Third Party Applications or Applications from unknown Sources in your Smartphone?

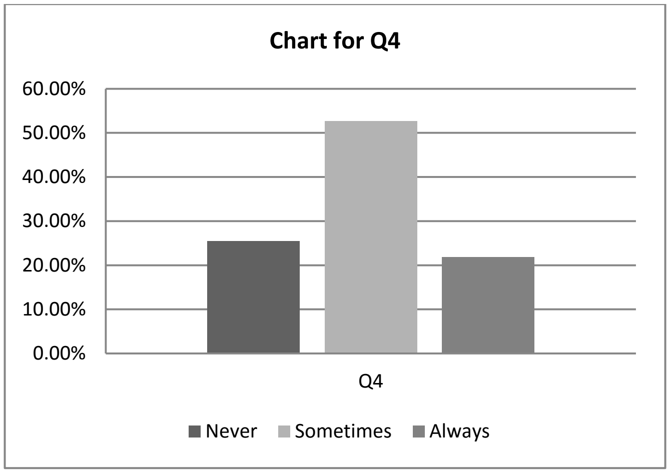


Fig. 5. Before Installing Application, do you Read Application Provider's Privacy and Policy for using Application's?

Q5: The technologies of smartphone have been increasing day by day. A smartphone can hold our different personal information such as photos and videos, mail, SMS etc. and other important information. Before installing application, the application provider shows the applications' phone access permission. Applications' phone access permission means what information of your smartphone access by the application. A user should always read through the applications' phone access permission carefully. Fig. 6 shows the bar chart of the Q5 questions' result. The chart shows that 11.6% people never read the applications' phone access permission and 46.5% people sometimes read the applications' phone access permission before install application.

Q6: Authentication is one of the major problems of smartphone. Suppose, someone gets your phone cleverly for a short time, if he/she does not know your smartphone authentication system, he/she cannot access any information from your smartphone. There are many authentication systems for smartphone including: pin code, password, pattern, fingerprint etc. Among them fingerprint is more secure than others. Fig. 7 shows our survey result for question Q6. We can see almost 98% of the people use authentication system to unlock the smartphone lock screen.

C. Proposed Model

Smartphone security is not limited to those six questions but, when we think about smartphone security and privacy awareness survey those questions gets the top priority. In recent days, those reasons are more responsible for losing the privacy of smartphone. Based on survey result we have developed (1) which can measure the level of awareness (LoA) for a smartphone user.

$$LoA = 1 - \left(\frac{2}{1+e^{-Q}} - 1\right) \tag{1}$$

Where, Q = Q1+Q2+Q3+Q4+Q5+Q6

We have considered some safe option and unsafe options for each question which denote to secure and insecure zone respectively. Safe options for every question carry the value of 0 (zero) and unsafe options carry value of 1 (one). Table III shows the safe and unsafe options for each question.

This proposed equation shows the level of awareness (LoA) in between the range of 0 to 1. Table IV shows the percentage level of awareness (*LoA*) for every possible value of '*Q*'. In Table IV, we can see if a user answered safe option of all the questions his/her LoA is 100%, if user contain 1 unsafe option his/her LoA is 53.78% and so on. Since we have mentioned earlier that only those questions do not responsible for losing the privacy of smartphone, many others reasons also harmful for privacy of smartphone. If we add more questions to this system, it will produce more accuracy.

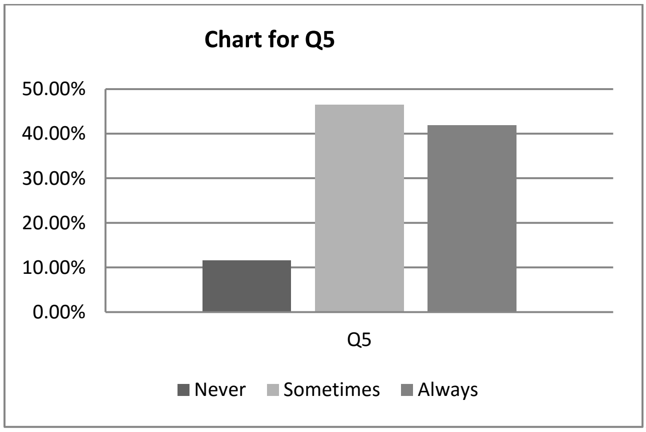


Fig. 6. Before Installing Application, do you ever read through Application's Phone Access Permissions?

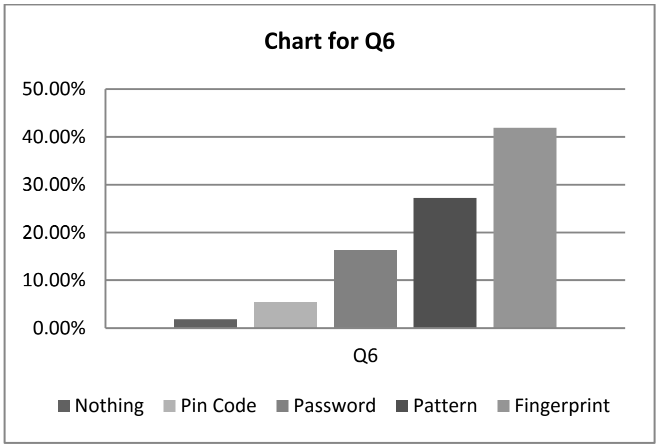


Fig. 7. What Authentication System do you use to Lock Screen for Security?

TABLE. III. CONSIDERED OPTIONS FOR QUESTIONS

Question	**Safe option**	**Unsafe option**
Q1	• Communication	• All • Browsing social websites • Web surfing
Q2	• Yes	• No
Q3	• No	• Yes
Q4	• Always	• Never • Sometimes
Q5	• Always	• Never • Sometimes
Q6	• Pin Code • Password • Pattern • Fingerprint	• Nothing

TABLE. IV. PERCENTAGE LOA FOR THE VALUE OF 'Q'

Value of *'Q'*	**Percentage *LoA***
0	100%
1	53.78%
2	23.84%
3	9.49%
4	3.60%
5	1.34%
6	0.49%

VI. CONCLUSION

This research aims at evaluating how much the smartphone users are aware about their security and privacy. In this study, firstly we have taken a survey from smartphone users to access the level of smartphone security awareness. We have found that on average 60% people do not aware about their smartphone security and privacy. Secondly, we have proposed a model to measure the level of awareness for smartphone users. We have found that almost 50% of the smartphone user contains 9.49% level of awareness. Although, the addition of new technologies makes a smartphone smarter, the security and privacy related threats also increases relatively. In future work, we will extend this study by adding others security and privacy related behavior and make our model more efficient and accurate.

REFERENCES

[1] Roesner, F., Kohno, T., & Molnar, D. (2014). Security and privacy for augmented reality systems. Communications of the ACM, 57(4), 88-96.

[2] Chin, E., Felt, A. P., Sekar, V., & Wagner, D. (2012, July). Measuring user confidence in smartphone security and privacy. In Proceedings of the Eighth Symposium on Usable Privacy and Security (p. 1). ACM.

[3] Jones, B. H., & Heinrichs, L. R. (2012). Do business students practice smartphone security?. Journal of Computer Information Systems, 53(2), 22-30.

[4] Yildirim, N., Daş, R., & Varol, A. (2014, May). A Research on Software Security Vulnerabilities of New Generation Smart Mobile Phones. In 2nd International Symposium on Digital Forensics and Security (pp. 6-16).

[5] PhoneArena, "Android's Google Play beats App Store with over 1 million apps, now officially largest," [Online]. Available: http://www.phonearena.com/news/ [Accessed: 07 July,2019].

[6] Dr.Web, "Android.Spy.277.origin," [Online]. Available: http://vms.drweb. [Accessed: 07 July, 2019].

[7] Dan, G., "Malicious apps in Google Play made unauthorized downloads, sought root,"[Online]. Available: http://arstechnica.com/information-technology/2016/01/malicious-apps-in-google-play-made-unauthorized-downloads-sought-root/. [Accessed: 07 July,2019].

[8] Benenson, Z., Kroll-Peters, O., & Krupp, M. (2012, September). Attitudes to IT security when using a smartphone. In Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on (pp. 1179-1183). IEEE.

[9] Ophoff, J., & Robinson, M. (2014, August). Exploring end-user smartphone security awareness within a South African context. In Information Security for South Africa (ISSA), 2014 (pp. 1-7). IEEE.

[10] Mylonas, A., Kastania, A., & Gritzalis, D. (2013). Delegate the smartphone user? Security awareness in smartphone platforms. Computers & Security, 34, 47-66.

[11] Alani, M. M. (2017). Android Users Privacy Awareness Survey. International Journal of Interactive Mobile Technologies (iJIM), 11(3), 130-144.

[12] Zaidi, S. F. A., Shah, M. A., Kamran, M., Javaid, Q., & Zhang, S. (2016). A Survey on security for smartphone device. IJACSA) International Journal of Advanced Computer Science and Applications, 7, 206-219.

[13] Muslukhov, I., Boshmaf, Y., Kuo, C., Lester, J., & Beznosov, K. (2012, April). Understanding users' requirements for data protection in smartphones. In Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on (pp. 228-235). IEEE.

[14] Muslukhov, I. (2012). Survey: Data protection in smartphones against physical threats. Term Project Papers on Mobile Security. University of British Columbia.

[15] Kataria, A., Anjali, T., & Venkat, R. (2014, February). Quantifying smartphone vulnerabilities. In Signal Processing and Integrated Networks (SPIN), 2014 International Conference on (pp. 645-649). IEEE.

[16] La Polla, M., Martinelli, F., & Sgandurra, D. (2013). A survey on security for mobile devices. IEEE communications surveys & tutorials, 15(1), 446-471.

[17] Cheng, J., Wong, S. H., Yang, H., & Lu, S. (2007, June). Smartsiren: virus detection and alert for smartphones. In Proceedings of the 5th international conference on Mobile systems, applications and services (pp. 258-271). ACM.

[18] Durairaj, M., & Manimaran, A. (2015). A study on security issues in cloud based e-learning. Indian Journal of Science and Technology, 8(8), 757-765.

[19] Pfleeger, C. P., & Pfleeger, S. L. (2002). Security in computing. Prentice Hall Professional Technical Reference.

[20] Khouzani, M. H. R., Sarkar, S., & Altman, E. (2012). Maximum damage malware attack in mobile wireless networks. IEEE/ACM Transactions on Networking, 20(5), 1347-1360.

[21] Peng, S. C. (2013). A survey on malware containment models in smartphones. In Applied Mechanics and Materials (Vol. 263, pp. 3005-3011). Trans Tech Publications.

[22] Mandke, K., Nam, H., Yerramneni, L., Zuniga, C., & Rappaport, T. (2003). The evolution of ultra wide band radio for wireless personal area networks. Spectrum, 3, 10-6.

[23] Xu, Z., & Zhu, S. (2012, August). Abusing Notification Services on Smartphones for Phishing and Spamming. In WOOT (pp. 1-11).

[24] Lee, H. T., Kim, D., Park, M., & Cho, S. J. (2016). Protecting data on android platform against privilege escalation attack. International Journal of Computer Mathematics, 93(2), 401-414.

[25] Yalcin, S. B. O. (2010). Radio Frequency Identification. Security and Privacy Issues. In 6th international workshop, RFIDSec (pp. 8-9).

[26] Dondyk, E., & Zou, C. C. (2013, January). Denial of convenience attack to smartphones using a fake Wi-Fi access point. In Consumer Communications and Networking Conference (CCNC), 2013 IEEE (pp. 164-170). IEEE.

[27] Amravati, M. E. S. (2015). A Review on Camera Based Attacks on Andriod Smart Phones. International Journal of Computer Science & Technology, 6(1), 88-92.

[28] Stites, D., & Tadimla, A. A Survey Of Mobile Device Security: Threats, Vulnerabilities and Defenses./urlhttp.afewguyscoding.com/2011/12 /survey-mobile-devicesecurity-threatsvulnerabilities-defenses.

[29] Davi, L., Dmitrienko, A., Egele, M., Fischer, T., Holz, T., Hund, R., & Sadeghi, A. R. (2012, February). MoCFI: A Framework to Mitigate Control-Flow Attacks on Smartphones. In NDSS (Vol. 26, pp. 27-40).

[30] Kim, I. (2012). Keypad against brute force attacks on smartphones. IET Information Security, 6(2), 71-76.

[31] Kumar, S., & Phrommathed, P. (2005). Research methodology (pp. 43-50). Springer US.

[32] Couper, M. P. (2005). Technology trends in survey data collection. Social Science Computer Review, 23(4), 486-501.

[33] Granello, D. H., & Wheaton, J. E. (2004). Online data collection: Strategies for research. Journal of Counseling & Development, 82(4), 387-393.

Support Vector Machine for Classification of Autism Spectrum Disorder based on Abnormal Structure of Corpus Callosum

Jebapriya S[1], Shibin David[2], Jaspher W Kathrine[3], Naveen Sundar[4]
Department of Computer Science and Engineering
Karunya Institute of Technology and Sciences, Coimbatore, India

***Abstract*—Autism Spectrum Disorders (ASD) is quite difficult to diagnose using traditional methods. Early prediction of Autism Spectrum Disorders enhances the in general psychological well- being of the child. These days, the research on Autism Spectrum Disorder is performed at a very high pace than earlier days due to increased rate of ASD affected people. One possible way of diagnosing ASD is through behavioral changes of children at the early ages. Structural imaging ponders point to disturbances in various mind regions, yet the exact neuro-anatomical nature of these interruptions stays misty. Portrayal of cerebrum structural contrasts in children with ASD is basic for advancement of biomarkers that may in the long run be utilized to enhance analysis and screen reaction to treatment. In this examination we use machine figuring out how to decide a lot of conditions that together end up being prescient of Autism Spectrum Disorder. This will be of an extraordinary use to doctors, making a difference in identifying Autism Spectrum Disorder at a lot prior organize.**

***Keywords*—Autism Spectrum Disorder (ASD); ASD screening data; ABIDE; machine learning**

I. Introduction

Autism Spectrum Disorder (ASD) is a neuro-developmental issue that can be described by distinct issues that may arise during social collaboration, correspondence and conduct. There are numerous other mental disorders like Autistic disorder and Asperger's disorder which have similar classification of symptoms. Depending upon the severity of symptoms, it prevails in numerous forms from extremely gentle ASD to exceptionally serious ASD. ASD must be analysed immediately so as to follow the advancement of the youngster and give legitimate treatment. An underlying screening is done in order to check the development of the child. ASD cannot be analysed by utilizing a solitary screening. A second screening test must be performed when the child is 2-3 years of age. It very well may be precisely analysed after the second screening test process. In the on-going years, researchers are working on it to anticipate it precisely within 18 months of the child. There are a lot of formal screening tools accessible for doctors to expand the exactness for assessing the formative status of the children. Anyway, just a significant number of doctors utilize those accessible tools. And furthermore, this needs a periodical screening way to track the development status of the child.

To address this issue, computer-aided learning for individuals with mental imbalance was created [1], [2]. Furthermore it was concluded that it would be progressively valuable if this would assist in identifying three noteworthy territories like social and relational abilities, unbending nature of reasoning and relational abilities. Despite the fact that it is hard to get a correct number of Autism cases and it is generally recognized that the predominance has been expanding in the course of recent years. There are signs that some ASD range issue might be ascribed to a mix of certain hereditary susceptibilities; for example, introduction to mercury at basic formative stages and diminished capacity to discharge mercury [3]. There are a few investigations that have appeared even environmental toxicity may play a major role in Autism. Analysts are yet dealing with these clutters to comprehend the shrouded manifestations and subtleties of Autism.

Autism spectrum disorders (ASD) is a disability in the socio - development which was first identified by Kanner. Kanner characterized ASD by the features of difficulties of social behaviours, limited repetitive interests, difficulties in interaction and behaviours [4]. In [5], it is stated that ASD is a disability related to development which progresses from childhood to adult and will exist as long as they live. Some of the common distinctive features of ASD are lack in language and communication skills, inadequacy in social interaction, show of inappropriate behaviours, etc.

In DSM 5, the grouping of pervasive developmental disorders as specified by DSM-IV-TR is modified by placing autism spectrum along with social, communication and restricted or repetitive behaviours [6]. In DSM-IV-TR criteria, pervasive developmental disorders are classified as five separate disorders such as, autistic disorder, childhood disintegrative disorder, Rett’s disorder, Asperger’s disorder and pervasive developmental disorder. Children diagnosed with ASD has been increasing as the years are progressing [7].

A. Problem Statement

Machine learning has been utilized to anticipate instances of youngster misuse utilizing organized information and literary data. This has nonetheless, not been done frequently and scarcely ever been finished for ASD and building up a choice emotionally supportive network that helps doctors with the recognition of ASD, has hardly been done, which is demonstrated by the absence of writing on the utilization of it.

II. Research Methods

A. Literature Survey

The literature review mainly focused on how the customary techniques were utilized to anticipate Autism Spectrum Disorder and what was the outcome obtained from the analysis. The study was finished utilizing two sorts of datasets M-Chat which is utilized for little children and rs-fMRI which is utilized for all age gatherings [8, 9].

a) Identifying the risk caused by ASD: The seriousness of ASD was resolved utilizing Modified Checklist for Autism in toddlers, otherwise called the M-CHAT. This is a screening apparatus which comprises of inquiries that must be replied by the guardians. The objective is to find the preeminent classifier for an autism dataset through feature relevance analysis. The classification algorithm is also used for predicting the threat level of autism [10, 11]. Among various classification algorithm applied, for example, BVM, CVM and MLR created high exactness of 95.21% utilizing Runs Filtering and it likewise precisely grouped the test dataset. This is valuable for age group of 16-30 years yet the forecast was not exact in all cases.

b) Auxiliary Imaging using Voxel-based Morphometry: The examination demonstrated that utilizing a Multivariate Pattern Analysis(MPA) and voxel-based morphometry (VBM) to order structural magnetic resonance imaging data obtained from 24 children and young people with mental imbalance IQ coordinated neurotypical participants was pertinent just to small data sets [12]. Multivariate Pattern Analysis (MPA), which is a pattern recognition technique that is solely based on machine-learning, can be used to group information by isolating between at any rate two classes. In MPA, the groups are distinguished with about 90% of accuracy based on gray matter in the medial prefrontal cortex, posterior cingulate cortex (PCC), and bilateral medial temporal lobes which are all regions within the default mode network (DMN) [13].

c) Machine Learning Techniques: There are many machine learning algorithms that can be used for classification. The three popular classification algorithms used are Random Forest, Naive Bayes and Support Vector machines .Apart from these algorithms we can use the java implementation of the C4.5 algorithm known as the J48 algorithm [14]. Using these assorted algorithms makes sure that the outcomes are highly reliable and it additionally helps us in finding whether the algorithm is usable for not for the classification task. When these algorithms were applied to the dataset that was isolated into two classes-ASD or No ASD, the results are as in Table I.

Be that as it may, utilizing the extra data we can discover how seriously the individual is influenced with ASD. This is finished by detaching into four groups. They are No ASD, Mild ASD, Moderate ASD and Severe ASD [14]. On utilizing similar characteristics and the previously mentioned machine learning algorithms the outcome got are as in Table II.

So as to build this we apply the 1-away strategy to the J48 calculation as it is the best one. Thus, the exactness is expanded from 54.1% to 90.2%.

TABLE. I. Results Obtained using Two Classes

Algorithm Applied	Accuracy	Precision	Recall
Naïve Bayes	0.865	0.866	0.865
Support Vector Machine	0.833	0.835	0.833
J48 (Decision tree)	0.871	0.871	0.871
Random Forest	0.851	0.854	0.851

TABLE. II. Results Obtained using four Classes

Algorithm Applied	Accuracy	Precision	Recall
Naïve Bayes	0.512	0.479	0.512
Support Vector Machine	0.493	0.475	0.493
J48 (Decision tree)	0.541	0.524	0.541
Random Forest	0.507	0.489	0.507

d) Using deep learning algorithms and Resting state functional magnetic resonance imaging (rs-fMRI): The classification of brain imaging data is stricter when using deep learning algorithms than using supervised methods. On utilizing significant neural framework a mean game plan exactness of 70% and a precision somewhere in the range of 66% and 71% in individual folds was acquired [15,16]. An expansion of 5% in classification accuracy was acquired while using deep learning classification method instead of Support Vector Machine [17]. In spite of the fact that the ABIDE dataset contains sensitive varieties, the deep learning methods envelop such assortments and yield better outcomes over machine learning algorithms. The neural patterns obtained from the classification show an anti-correlation of brain function between posterior and anterior areas of the brain or cerebrum work among foremost and back regions of the brain.

B. Data Exploration

The Modified Autism Checklist in Toddlers (M-CHAT) is a validated developmental screening tool for children aged 16 to 30 months. It is intended to identify children who may benefit from a more thorough evaluation of development and autism. This helps to find the best autism dataset classifier by analysing feature relevance and classification algorithm. Among the various classification algorithms used, algorithms such asBVM, CVMand MLR which produced an accuracy of 95.21 % using Runs Filtering method [18]. This method accurately classified the test dataset.

The Autism Brain Imaging Data Exchange (ABIDE) initiative has totalled practical and auxiliary brain imaging information gathered from research facilities around the globe to quicken our comprehension of the neural bases of autism. Every gathering was made through the accumulation of datasets freely gathered crosswise over in excess of 24 global brain imaging research centers and are being made accessible to examiners all through the world, predictable with open science standards. Since the data set contains more than 1112 records, before the data can be used for our machine learning process, the data must be cleaned. The data will be explored after cleaning the ABIDE data to determine if co-occurring conditions are present in the data. If there are any clusters that could help us predict ASD, it will also be investigated [19]. Support Vector Machine algorithm will be used on the data set, to determine if there are attributes that seem strongly correlated to ASD. In this algorithm, each data item is plotted as a point in n-dimensional space with each feature being a coordinate value. 'n' denotes the number of features selected for identification [20].

III. Proposed Work

A survey was done on various methods of predicting Autism Spectrum Disorder (ASD) using different machine learning algorithms. Based on the survey a conclusion is drawn that the algorithm Support Vector machine (SVM) can be used for identifying patterns from autistic brain images [21]. Here we have used the region called "Corpus Callosum" to identify the differences between the autistic and non-autistic brain images. The Corpus Callosum is a fibre bundle which connects the left and right hemisphere of the brain. Using SVM algorithm, there were striking differences noted in that particular region where the thickness of Corpus Callosum was either too thick or thin compared to those brain images without ASD [21], [22]. Furthermore it is found that the autistic brain had a decreased white matter volume and larger ventricles. Using these observations will ensure our result outcomes are more reliable with higher accuracy rates. The proposed architecture is shown in Fig. 1.

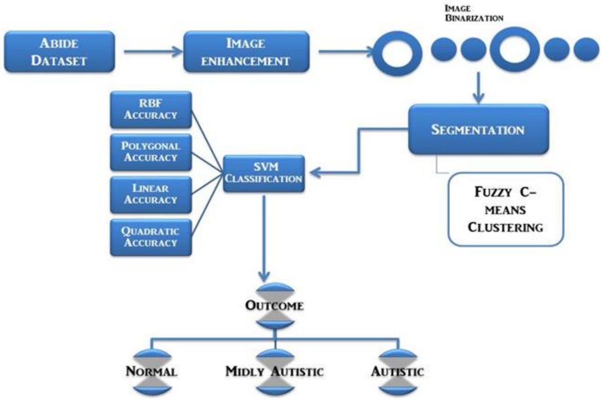


Fig. 1. Architecture of the Proposed Work.

A. Data Sets

The ABIDE dataset contains 1112 records. This includes data from 539 individuals with ASD and data from 573 typical controls (ages 7-64 years). The Data Processing Assistant for Resting-State fMRI (DPARSF), the Configurable Pipeline for the Analysis of Connectomes (CPAC), the Neuro imaging and the Connectome Computation System (CCS) are used for Functional pre-processing. This large pre-processed data is used as an input for the classification algorithm.

IV. Results

Below, the results from machine learning are described. We divide our data set into three different classes, namely Normal, Mildly Autistic and Autistic. A total of 100 iterations are performed to classify the individuals falling under the three classes.

A. Machine Learning Results using Three Classes

We applied the Support Vector Machine classification algorithm to the data set when it was divided into three classes, either the individual was Normal, or the individual was Mildly Autistic or Autistic. This was performed by using the attributes resolved in the data set step mentioned above. We have used the MATLAB software for accuracy calculation.

B. Enhancement of Image

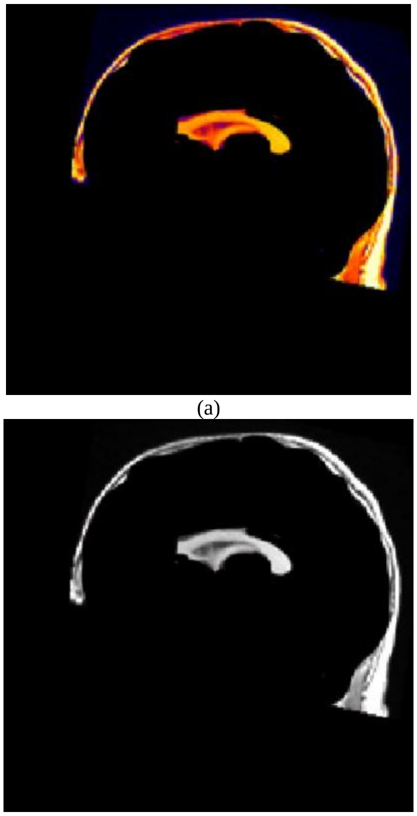
Image enhancement is a process that is widely utilized in numerous image processing applications, to amplify the quality of images. In MATLAB software, the image is enhanced by converting the image into gray scale. The difference is seen in the below given images (Fig. 2(a), Fig. 2(b)).

C. Image Binarization

Image binarization is the process of converting a pixel image into a binary image. Here, we use two main functions, one to normalize the gray scale image by defining a threshold value and the other to convert the indexed image to black and white intensity. The binary image after thresholding is shown in Fig. 3.

D. Segmentation

Image segmentation is the process of splitting images into multiple fragments. This division into fragments is mostly based on the characteristics of the pixels of the image. We have applied fuzzy c-means (FCM) clustering to produce one or more clusters of the given binary image. After FCM segmentation, the following clusters (Fig. 5) are generated out of which one is chosen.



(a)

(b)

Fig. 2. (a) MRI Scan of An individual before Enhancement. (b) MRI Scan of an Individual after Enhancement.

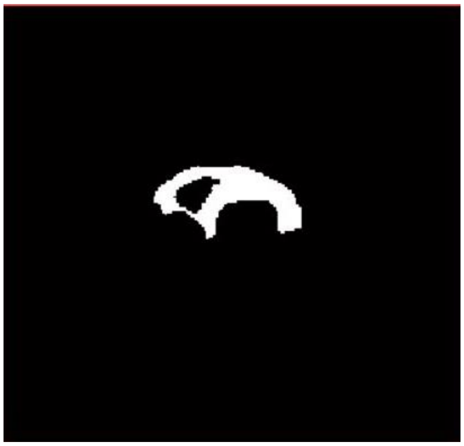


Fig. 3. Binary Image of an MRI Scan.

E. *Accuracy Evaluation*

A support vector machine (SVM) is a supervised learning algorithm that creates a hyper plane in between data sets to classify the belongingness of the data to its appropriate class. The maximum accuracy of different types of SVM with 100 iterations to classify the images into the aforementioned three classes has been evaluated. The confusion matrix generated is given in Fig. 4.

The accuracy in percentage for different types of SVM is mentioned in Table III and Fig. 6.

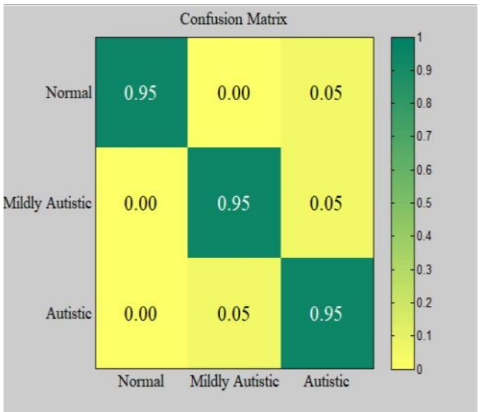


Fig. 4. Confusion Matrix for Classifier using three Classes.

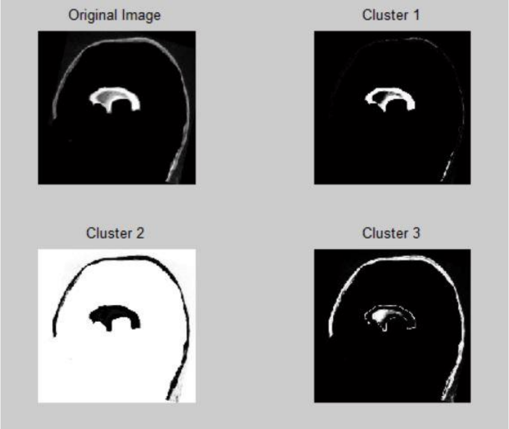
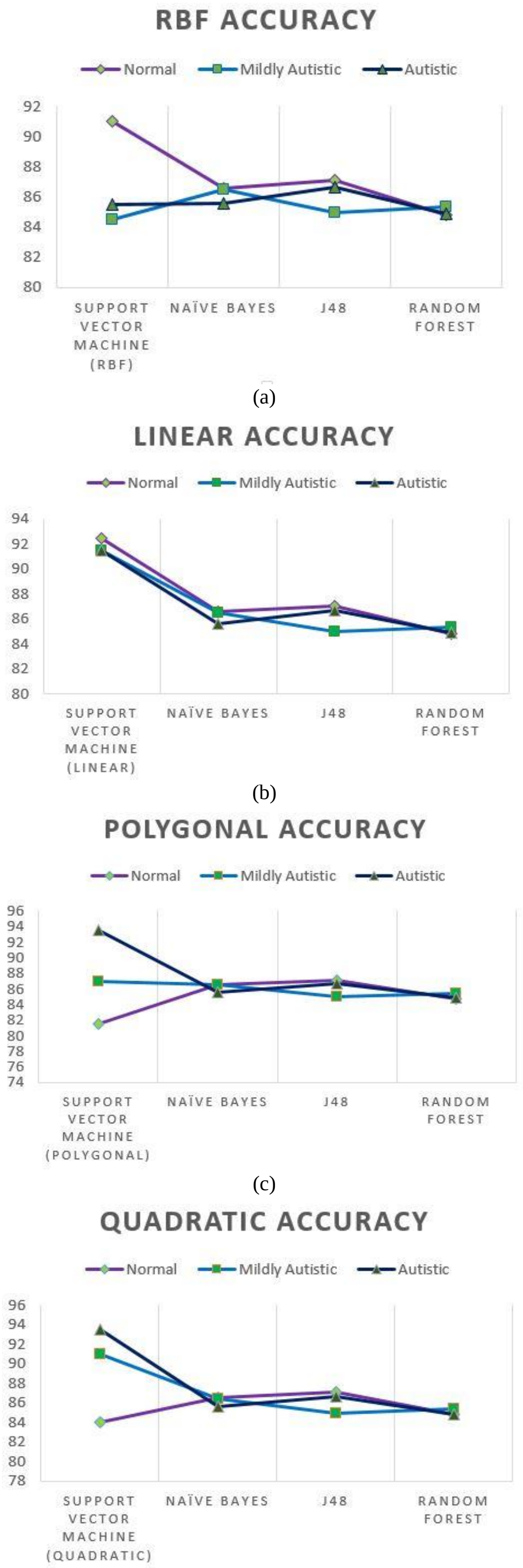


Fig. 5. Clusters Generated after FCM Segmentation.



(a)

(b)

(c)

(d)

Fig. 6. (a) Rapid Basis Function (RBF) Accuracy versus Regular Machine Learning Techniques, (b) Linear Accuracy versus Regular Machine Learning Techniques. (c) Polygonal Accuracy versus Regular Machine Learning Techniques. (d) Quadratic Accuracy versus Regular Machine Learning Techniques.

TABLE. III. MAXIMUM ACCURACY ACQUIRED FROM VARIOUS TYPES OF SVM

Types of SVM	Normal	Mildly Autistic	Autistic
RBF Accuracy	91%	84.5%	85.5%
Linear Accuracy	92.5%	91.5%	91.5%
Polygonal Accuracy	81.5%	87%	93.5%
Quadratic Accuracy	84%	91%	93.5%

V. DISCUSSION AND CONCLUSION

The objective was to identify the conditions that demonstrate to be prescient of ASD. This data can be utilized by physicians to enable them to confirm a complete formal screening for ASD. Complex arrange parameters were utilized to plan and analyze discriminate examination along with bolster vector group of classifiers with a most extreme reachable exactness of 94.7% utilizing four highlights and a second request polynomial bit in SVM. The investigation has endeavored to characterize the chemical imbalance range scatter and creating subjects utilizing administered learning systems as depicted in Figure 6. For future work, the focus is towards the investigation of likelihood by utilizing profound learning approaches for the programmed acknowledgment of SMM practices inside and crosswise over subjects.

REFERENCES

[1] Kathleen M. Carroll, Bruce J. Rounsaville, “Computer-assisted Therapy in Psychiatry: Be Brave—It’s a New World”, Current Psychiatry Reports, 2010 Oct; 12(5): 426–432.

[2] Jorn Moock, “Support from the Internet for Individuals with Mental Disorders: Advantages and Disadvantages of e-Mental Health Service Delivery”, Frontiers in Public Health. 2014; 2: 65.

[3] Amy E. Kalkbrenner, Rebecca J. Schmidt, Annie C. Penlesky, “Environmental Chemical Exposures and Autism Spectrum Disorders: A Review of the Epidemiological Evidence”, Current Problems in Pediatric and Adolescent Health Care. 2014 Nov; 44(10): 277–318.

[4] Rachel Cooper, “Diagnostic and Statistical Manual of Mental Disorders (DSM), Encyclopedia of Knowledge Organization, ISKO, available at: https://www.isko.org/cyclo/dsm.htm

[5] Vihang N. Vahia, “Diagnostic and statistical manual of mental disorders 5: A quick glance”, Indian Journal of Psychiatry. 2013 Jul-Sep; 55(3): 220–223.

[6] Murat Gök, “A novel machine learning model to predict autism spectrum disorders risk gene”, Neural Computing and Applications, 2018, pp. 1-7.

[7] Centers for disease control and prevention (CDC), available at: https://www.cdc.gov/ncbddd/autism/data.htm

[8] Anibal Slon Heinsfeld , Alexandre Rosa Francop, R. Cameron Craddockt,, Augusto Buchweitzp,, Felipe Meneguzzio “Identification of Autism Spectrum Disorder using Deep Learning and the ABIDE Dataset“,2017.

[9] Felix D.C.C Beacher ,Eugenia Radulescu ,Ludovico Minati, Simon Baron-Cohen, Michael V.Lombardo, Meng-Chuan Lal, Anne Walker, Dawn Howard, Marcus A.Gray ,Neil A. Harrison, Hugo D. Critchley “Sex Differences and Autism: Brain Function during Verbal Fluency and Mental Rotation”,2012.

[10] Lucina Q. Uddin, Vino Menon, Christina B. Young, Srikanth Ryali, Tianwen Chen, Amirah Khouzam, Nancy J. Minshew, and Antonio Y. Hardan; “Multivariate searchlight classification of structural MRI in children and adolescents with autism”,2017.

[11] LamyaaSadouk ,TaoufiqGadi,andEl Hassan Essoufi; “A Novel Deep Learning Approach for Recognizing Stereotypical Motor Movements within and across Subjects on the Autism Spectrum Disorder”,2018

[12] M. S. Mythili and A. R. Mohamed Shanavas; “An improved autism predictive mechanism among children using fuzzy cognitive map and feature extraction methods (feast)”,2016

[13] Christine Ecker, Susan Y Bookheimer, Declan G M Murphy; “Neuroimaging in autism spectrum disorder: brain structure and function across the lifespan”,2015.

[14] Bram van den Bekerom “Using Machine Learning for Detection of Autism Spectrum Disorder”,2017.

[15] R.Geetha Ramani, K.Sivaselvi “Autism Spectrum Disorder Identification Using Data Mining Techniques”,2017.

[16] M. S. Mythili, A. R. Mohamed Shanavas; “A Study on Autism Spectrum Disorders using Classification Techniques “, 2014.

[17] Tabtah, F. “Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment.” Proceedings of the 1st International Conference on Medical and Health Informatics ,pp.1-6. Taichung City, Taiwan, ACM.,2017.

[18] Thabtah, F. “Machine Learning in Autistic Spectrum Disorder Behavioural Research: A Review.” To Appear in Informatics for Health and Social Care Journal. December, 2017 (in press),2017.

[19] Thabtah F, Kamalov F., Rajab K “A new computational intelligence approach to detect autistic features for autism screening.” International Journal of Medical Infromatics, Volume 117,2018.

[20] Christina Schweikert, Yanjun Li, David Dayya, David Yens, Martin Torrents, D. Frank Hsu “Analysis of Autism Prevalence and Neurotoxins Using Combinatorial Fusion and Association Rule Mining”, 2009.

[21] Koyamada S, Shikauchi Y, Nakae K, Koyama M, Ishii S (2015) “Deep learning of fMRI big data: a novel approach to subject-transfer decoding”, 2015.

[22] Plis SM et al. (2014) Deep learning for neuroimaging: a validation study frontiers in neuroscience 8,2014.

IoT based Temperature and Humidity Controlling using Arduino and Raspberry Pi

Lalbihari Barik
Department of Information Systems, Faculty of Computing and Information Technology in Rabigh
King Abdulaziz University, Kingdom of Saudi Arabia

Abstract—**Internet of Things (IoT) plays a pivotal part in our mundane daily life by controlling electronic devices using networks. The controlling is done by minutely observing the important parameters which generate vital pieces of information concerning the functioning of these electronic devices. Simultaneously, this information will transmit these vital statistics from the transmitting device as well as save the same on the cloud to access by the applications and supplementary procedures to use them. This scrutiny associates the outcomes of the environmental observances like the humidity and temperature measurements using sensors. The gathered information could be profitably used to produce actions like distantly dominant cooling, heating devices, or long term statistics, which will be useful to control the same. The detected data are uploaded to the cloud storage through network and associate using android application. The system employs Arduino UNO with Raspberry Pi, HTU 211D sensor device, and an ESP8266 Wi-Fi module. The experimental results show the live temperature and humidity of the surroundings and the soil moisture of any plant using Arduino UNO with Raspberry Pi. Raspberry Pi is mainly used here for checking the temperature and humidity through the HTU 211D sensor element. The sensors are used for measuring the temperatures from the surroundings, storing displayed information with different devices. Here, the ESP8266 Wi-Fi module has been used for data storing purpose.**

Keywords—IoT; Raspberry Pi; Arduino UNO; data transmission; sensors

I. Introduction

IoT is used for connecting the electronic devices with the internet. The devices may vary from the temperature measuring equipment and vehicles SOS system to other electronic devices such as sensors, software's, and network connectivity facilities, which sanction collecting and exchanging data. The twenty-first century has witnessed a massive paradigm shift to and focusing on global attention onto IoT as a burgeoning discipline with multiple possibilities and diverse opportunities for growth and development [1]. Internet connection facilitates the smooth functioning of the devices that have become indispensable parts of our day-to-day lives and existence. The Internet offers the provision to link and network different kinds of devices like sensors and fitness devices. In the changed scenario post the September 11, 2001 attack on the United States where surveillance has gained paramount importance in proposed model security and survival, the internet facilitates wholesome and perfect monitoring systems using closed-circuit cameras [2].

All these devices that enable them to upload input as well as output to the Internet using cloud provisioning. The information thus garnered is accessible for monitoring and analysis anywhere in the globe via the internet [3]. In order to cut down on human effort and involvement, of late people increasingly depend on embedded systems to control and monitor the factors affecting the ecosystem. Temperature and humidity are vital in observing and understanding nature. IoT comes into the picture here by significantly enhancing the efficiency of the mechanism and systematically cutting down on human involvement, and thereby overall expenditure [4].

Practically, every part of exercise contains controlled schedules of temperature as well as humidity. However, the exact value of temperature with its significant feature in any field is essential in monitoring [5]. Constant perception in temperature is utilized in various industries like the pharmaceutical industry as the driving force behind these monitoring systems, computerized and straightforward temperature sensors can use [6]. Resistors, semiconductors, thermistors estimate temperatures values. These components are present inside the sensor to retrieve the temperature in consonance with the circumstances. The primary goal of our system is to supervise the live temperature and humidity within a low cost [7].

Raspberry Pi is the observational system or controller which is used for the cloud saving. Python is the programming language which is utilized in Raspberry Pi. HTU 211D sensors is a temperature sensor which is used here for the sensing purpose [8]. This comprises of temperature ascertaining capacity and favorable fundamental position of utilizing HTU 211D sensors, which boasts of less weight and ease of use. The sensor is associated with Raspberry Pi utilizing connecting wires. Temperature sensor HTU 211D sensors is utilizing is perused put away, and shown in the Raspberry Pi unit [9].

IoT based devices in homes and industries are used for controlling all the electrical or electronic devices which are present. Additionally, the saved information of the IoT devices can be controlled from anywhere [10]. The sensor analyzes the graphical representation of the observed data in every user-defined format wherever in the world. In this work, IoT based Arduino with Raspberry Pi microcontroller is used. Humidity and temperature monitoring using Arduino is an exciting and secure process. Furthermore, this flexible system obtains more values in calculating the actuator from the data saved on the internet [11]. For connecting the Arduino board with Raspberry Pi, USB line serial interface is essential to connect with any application [12].

II. Literature Review

Arduino controller system is used for measuring the temperature and wetness of the devices, pressure, and height measurement. The setup contains the height measuring device and a measuring or controlling instrument. In this work, they proposed an Arduino UNO with Raspberry Pi data processing unit [13]. Along with this setup, a Cube satellite is included here to supply the data of weather condition when no network coverage is available. This method has advantages like ease of construction, portable device; price is economical, low power, and a reliable system.

However, there are some disadvantages like not used in long-distance while powerful transceiver sections are not present, and the gas balloon is also hauling, and parts could be broken in the rain or during practice. Alternative energy panel method did a significant role in measuring the aforementioned parameters [14]. The apparent statistical data are first collected and then sent by the use of a GSM module through the receiver. A server is used here for connection and collection of information.

Weather surveillance development system can be used in gathering real-time information as well as for transmission [15]. They have achieved by introducing a VAISALA WXT520 weather transmitter device to transmit the information from one place to another place [16]. This device senses all the ecological parameters and their ratio. Then the collected real-time information's are transmitted wirelessly over a long distance through GSM [17]. This method provides flexibility because, in this method, one can add or remove the measuring parameters. Santhosh et al. [18] had projected a new model for ecological observation applications.

In this method, they proposed a system for patient monitoring and transmitting their measured data to the doctors. The medical base station and variety of distributed wireless device nodes are employed in this work for transmitting and receiving purpose [19]. Here the Base station is developed by Raspberry Pi device node along with the Zigbee [20]. The device is accessed by the use of nodes with internet Wi-Fi for gathering the information from the transmitting place. The web application which is developed is Apache protocol internet server [21]. This method has the following advantages like affordable, compacted, easy to modify, and simple to keep up and have the drawback of group action sensing modalities to receiver nodes [22].

Additionally, internet interface based systems are developed that utilizes a wireless device network (WDN) for agricultural area monitoring [23]. WDN consists of a frequency transmitter and a receiver. AVR-ZigBee, Bluetooth-module, temperature, humidity, soil wet sensors, and LCDs are used in this research [24]. Smart PZT sensors were used as an actuator and receiver, coupled with two XBee's and two Ardiuno as signal generator and signal receiver in damage identification [25]. This method is reliable and economical for agricultural unit's observation system.

III. Architecture of the System

This part concludes that the combination of Arduino UNO with Raspberry Pi is a perfect method for electronic device monitoring, data collection, and data saving. This paper helps to calculate the humidity and temperature values of the particular surroundings using the DHT-11 Temperature Sensor, and the results are graphically analyzed by one ThingSpeak platform software and an ESP8266 Wi-Fi module [7]. Fig. 1 shows the overall block diagram setup of the proposed system.

In this system, the Arduino mega 2560 Raspberry Pi microcontroller is used for controlling the temperature, water level, and the humidity levels. The Arduino mega 2560 Raspberry Pi microcontroller is the heart of this system also the power supply used here is a solar panel instead of the regular electric supply. Along with this controlling process, the system is capable of switching on and off the DC motor based on the soil water content level as well as the rainy season. Moreover, the system is also portable on the weather forecast.

In this work, the calculation of the humidity and temperature value of the particular surface area is being done using the HTU-211D sensor. The sensor is preferably used for sensing the humidity and temperature of any area. Then the sensors sensed data are controlled and collected by the Arduino UNO with Raspberry Pi microcontroller: Power Supply from solar Panel is served for the microcontroller.

The system is beneficial for farmers to work in the non-power sector to save electricity. Moreover, the results are graphically analyzed by using the one ThingSpeak platform through the ESP8266 Wi-Fi module. Finally, collected data from the HTU 211D sensor, saved the same data, and analyzed the data by using the cloud data module [8].

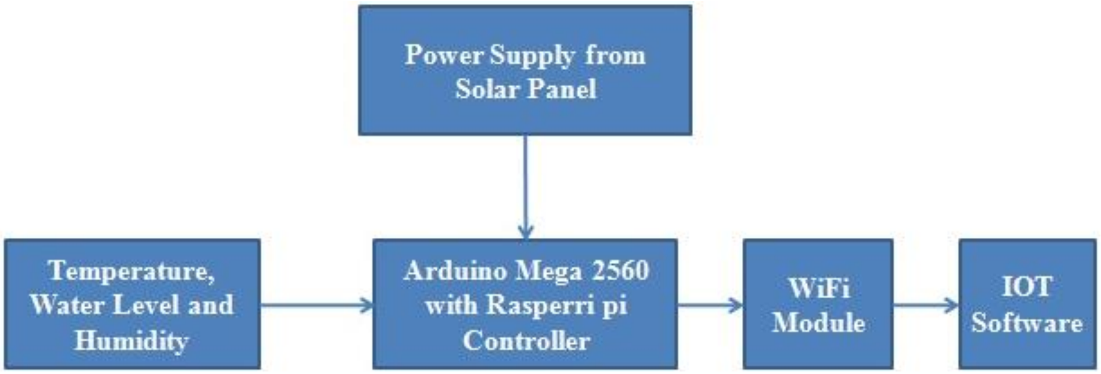

Power Supply from Solar Panel
Temperature, Water Level and Humidity
Arduino Mega 2560 with Rasperri pi Controller
WiFi Module
IOT Software

Fig. 1. Overall Block Diagram.

Fig. 2 shows the solar panel power supply supplied to the whole system. During periods of rain, the battery saved data is supplied. So, this will be beneficial for the farmers in monitoring the agriculture area. Fig. 3 shows the circuit diagram for measuring the humidity and temperature of an area. It shows the solar power supply of the system with an Arduino MCU with Raspberry Pi microcontroller, HTU 211D sensor, and an ESP8266 Wi-Fi module.

Temperature detector which is used here is a 4-pin low price extremely reliable detector named HTU 211D SENSORS. The first pin is connected with a Vcc node point. Here, the utilized power supply is a solar panel. The second pin is an information pin which will collect all the information from outside and provides information to the microcontroller. The sensor pin configuration of HTU 211D sensor detectors is represented in Fig. 3.

The temperature detector is very much useful for getting digital signal output. The HTU 211D SENSOR detector includes a resistive wetness component and is connected to an extraordinarily high-performing-8-bit-microcontroller. This sensor provides the best worth output, fast response, low cost, and is interference-proof. Their temperatures vary from 00C-550C, and the wetness value is among 20-90%.

To transfer the readings of the device from HTU 211D sensor to open the supply cloud ThingSpeak software, Arduino UNO with Raspberry Pi interfaces at the output with a LAN module named ESP-8266. In this module, ESP-8266 LAN semiconductor device is connected with a full TCP/IP protocol stack. A voltage of 3.3V is ideal, which is then connected by Arduino UNO with Raspberry Pi on PC. The calculation is performed using the AT command and wants the desired sequence, to be used as a saver. The module will work on each saver and server. It gets associated when connected to LAN through the module so that it can transmit over the web.

During the testing of the ESP-8266 module, the module is connected with the Arduino UNO Raspberry Pi. Then the programmed Arduino UNO Raspberry Pi set up is connected with the ThingSpeak platform through the ESP8266 LAN module. ESP8266 LAN module acts as a protocol shopper, and it will send the knowledge to ThingSpeak server. ThingSpeak is the best IoT platform used for data collection and storage purposes. Another unique feature of ThingSpeak is the data analysis and comparison module. Comparison between two different days can be accomplished using ThingSpeak platform.

The HTU 211D sensors are used to senses the humidity and temperature, and transfer the collected data through the 5th pin of Arduino MCU connected Raspberry Pi, as shown in Fig. 4. This set up can also control the DC fan, motor, and water levels for supporting farmers. Then the measured values of humidity and temperature values from the Arduino MCU are uploaded to the Cloud.

Fig. 4 is the overall hardware setup of the proposed system. This arrangement the Arduino mega 2560 Raspberry Pi microcontroller is worn to design for calculating the temperature, water level, and the humidity levels of the primary agricultural areas also help the farmers. The Arduino mega 2560 Raspberry Pi microcontroller is the sole of this arrangement, as well as the used power supply in this work, is the solar panel as a replacement for the regular home power supply. Because the solar panel installation only required some amount; after the installation, there is no money spending required. Next to the controlling process of such arrangement, in addition to that, it can switch on and off the DC motor support on the soil water content level as well as the weather season. Furthermore, the system is convenient for the weather forecast.

Then the collected data are transferred to the farmers live through the GSM to their cell phones. Based on the water level measuring system, the collected data are sending to the farmer’s cell phone continuously. They can switch on or off their motor based on the collected data from the water level measuring system. This is beneficial for the farmers to control the motors as well as can watch their plants from their house. Moreover, this helps the plants from the overwatering. The system is beneficial for water scarcity problems. The values are uploaded within the stipulated time period through the ESP-8266 Wi-Fi system. Then, from the Cloud, the humidity and temperature standards are measured using one ThingSpeak platform from anywhere.

Here, the used open data platform source is ThingSpeak software, which is free. Two-parameter tabs as humidity and temperature are selected in Fig. 5. After the new channel log in two API keys are generated. The original String API Key is “NTIM1RXET6YVUVWF”. Then replace the above line with the given program API key. Next, substitute the Host Name and Password with Wi-Fi name and Wi-Fi password. The original String-Hostname is “Jonah” and the password is “2569696”.

Fig. 2. Solar Panel.

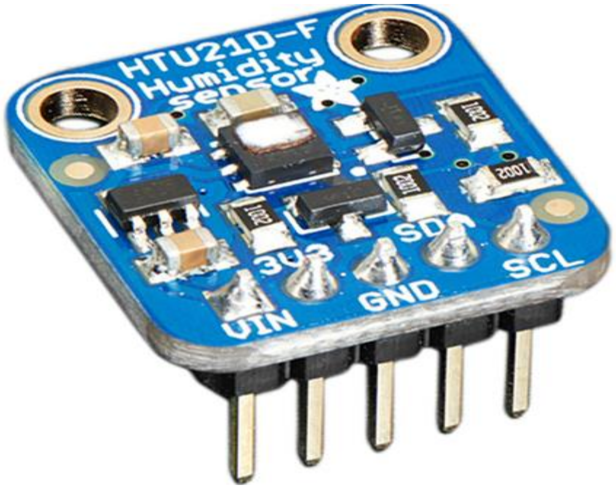


Fig. 3. Adafruit HTU 211D Sensors.

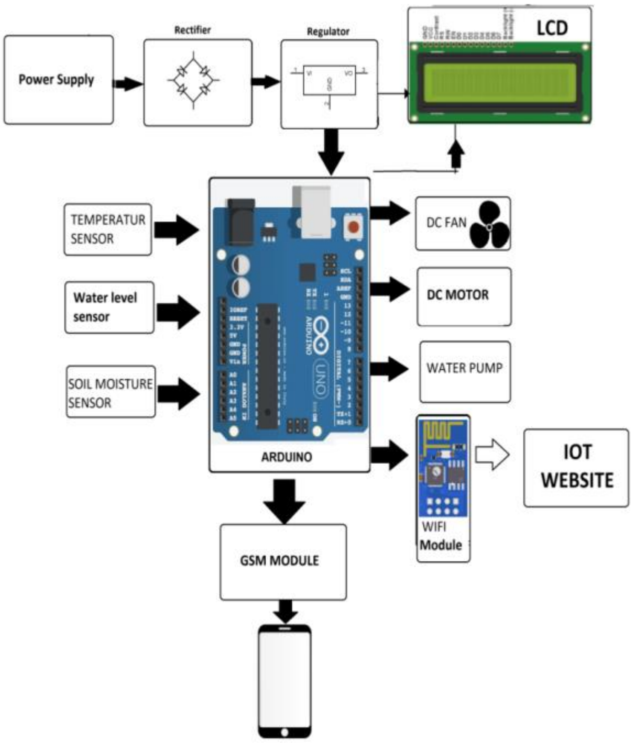
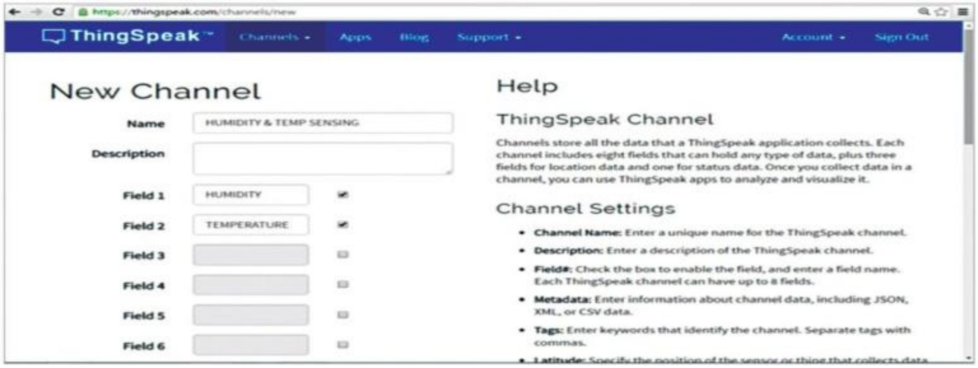


Fig. 4. Circuits and Boards.


https://thingspeak.com/channels/new

ThingSpeak™ Channels ▾ Apps Blog Support ▾ Account ▾ Sign Out

New Channel

Name: HUMIDITY & TEMP SENSING

Description:

Field 1: HUMIDITY ☑

Field 2: TEMPERATURE ☑

Field 3: ☐

Field 4: ☐

Field 5: ☐

Field 6: ☐

Help

ThingSpeak Channel

Channels store all the data that a ThingSpeak application collects. Each channel includes eight fields that can hold any type of data, plus three fields for location data and one for status data. Once you collect data in a channel, you can use ThingSpeak apps to analyze and visualize it.

Channel Settings

- **Channel Name:** Enter a unique name for the ThingSpeak channel.
- **Description:** Enter a description of the ThingSpeak channel.
- **Field#:** Check the box to enable the field, and enter a field name. Each ThingSpeak channel can have up to 8 fields.
- **Metadata:** Enter information about channel data, including JSON, XML, or CSV data.
- **Tags:** Enter keywords that identify the channel. Separate tags with commas.
- **Latitude:** Specify the position of the sensor or thing that collects data

Fig. 5. Cloud, Humidity and Temperature Measurement.

Fig. 6 shows the overall setup with all sensors and cloud. The figure clearly shows the cloud, humidity, and temperature measurement systems. The program should be verified with the Wi-Fi setup. To import the DHT library in Arduino Integrated Development Environment (IDE), select the input sketch from the selected input folder. Then click 'import' to retrieve the data from the library. To save cloud in the library, click 'add library'; then select the library that has downloaded. Compile the sketch/program and upload to Arduino MCU through Arduino IDE. For these steps, better internet connectivity is indispensable, and hence, it should ensure beforehand.

The central unit may be a microcontroller (Arduino UNO) and acts as the central processor unit for the complete system. This unit interfaces with the device chip as the input for receiving temperature and humidness readings. For output, it interfaces with the Wi-Fi module to send the received information to the cloud over the web. The microcontroller polls the device to retrieve information and sends over the web to ThingSpeak.

To begin with, Raspberry Pi should be ready, and for that, need NOOBS. It is a software-based system manager that simplifies transfer, install, and then acquired wind of Raspberry Pi. Boot the NOOBS system once in the beginning; then get a variety of operating systems (OS) to decide on from the system. NOOBS makes obtaining started with Pi simple and includes a bunch of in OS to decide. The Raspberry Pi itself does not go together with the software system. Raspbian is the "official" software system of the Raspberry Pi. Raspbian has been the quality Raspberry Pi in OS like UNIX.

Since the system includes temperature and humidity watching, one device interface is required and no native storage of information. Designate Arduino UNO with Raspberry Pi microcontroller that serves the purpose well because of its simplicity, lustiness, and low value. Fig. 6 shows an image of

Arduino UNO with Raspberry Pi microcontroller utilized in the proposed system. This microcontroller board is predicated on the ATmega328P. The controller has a USB port, 14 digital input/output pins, 6 analog input pins, 16 megacycle quartz with a power jack, and a button. It is battery-powered with a battery. It is programmable with the Arduino IDE via a sort B USB cable.

The all humidity and temperature values will be uploaded on the ThingSpeak platform. After that, one can see its graphical representation of both humidity and temperature values in a separate view window, as shown in Fig. 7. If one wishes to change the channel or field name, one can change it from the channel settings. Finally, the collected data are transferred to the farmers time to time with the GSM to their cell phones. This is beneficial for the farmers to control the motors as well as can watch their plants from their house. The proposed hardware prototype is shown in Fig. 17.

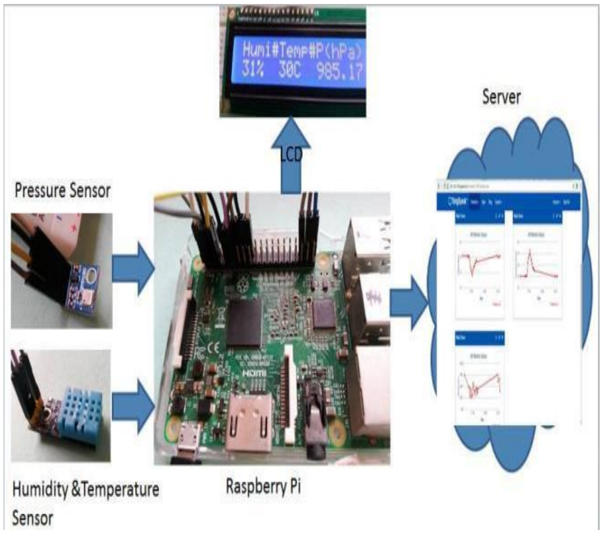
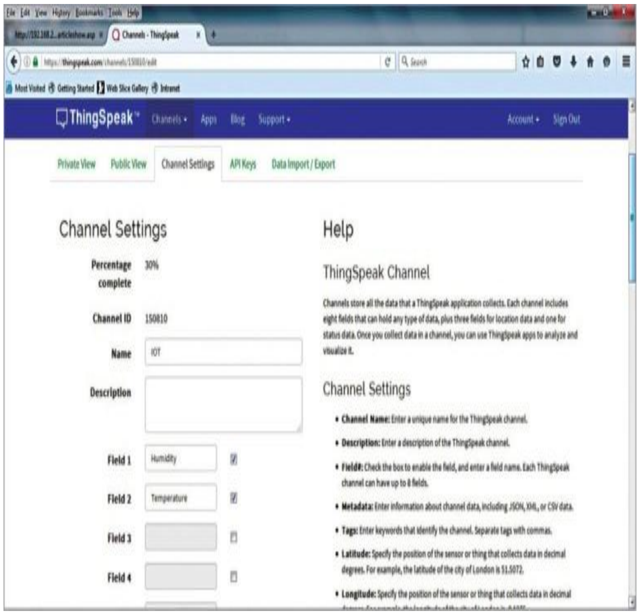


Fig. 6. Overall Set up with All Sensors and Cloud.



Channel Settings

Percentage complete 30%

Channel ID 150810

Name IOT

Description

Field 1 Humidity ☑

Field 2 Temperature ☑

Field 3 ☐

Field 4 ☐

Help

ThingSpeak Channel

Channels store all the data that a ThingSpeak application collects. Each channel includes eight fields that can hold any type of data, plus three fields for location data and one for status data. Once you collect data in a channel, you can use ThingSpeak apps to analyze and visualize it.

Channel Settings

- **Channel Name:** Enter a unique name for the ThingSpeak channel.
- **Description:** Enter a description of the ThingSpeak channel.
- **Field#:** Check the box to enable the field, and enter a field name. Each ThingSpeak channel can have up to 8 fields.
- **Metadata:** Enter information about channel data, including JSON, XML, or CSV data.
- **Tags:** Enter keywords that identify the channel. Separate tags with commas.
- **Latitude:** Specify the position of the sensor or thing that collects data in decimal degrees. For example, the latitude of the city of London is 51.5072.
- **Longitude:** Specify the position of the sensor or thing that collects data in decimal

Fig. 7. Channel Settings.

IV. Results and Discussion

IoT based temperature and humidity measurement system provides an economical and safe system. This is very useful for the detection of agricultural-related parameters. The results of the temperature and humidness will see on the Raspbian OS terminal. The central hardware element of the proposed system is the microcontroller that interfaces with alternative elements of the system. Since the system includes temperature and humidity controlling that one device interface is required and no primary storage of information. This designated an Arduino UNO with Raspberry Pi microcontroller.

In this regards, the Arduino mega 2560 Raspberry Pi microcontroller's controlling the temperature, water level, and the humidity levels measurement plots are plotted. The Arduino mega 2560 Raspberry Pi microcontroller used the solar panel instead of the proposed electric supply. The comparisons between the supply usage and their advantages are studied. Moreover, the graphs of the controlling process system, switch on and off the DC motor based on the soil water content level and the weathers forecasting are plotted. Water supply content levels, as well as the rainy season are drawn in the ThingSpeak software.

Fig. 8 shows the water level checking setup. In this setup, the sensor first senses the water level. Then data is then transferred to the farmers live through the GSM to their cell phones. This is beneficial for the farmers to control the motors as well as can watch their plants from their houses.

Fig. 9 shows the comparison between the solar panel powers to standard power. In this work, instead of the standard power supply, a solar panel supply is used. The ThingSpeak software attains the generated power difference to standard power.

Fig.10 graph shows the humidity results of the proposed system. Humidity is the quantity of water vapor present in the air. The water vapor amount is essential to attain the saturation state in proportion with the increase in temperature value. As the temperature of a parcel of air is low, it will eventually reach the saturation point without adding or losing the mass of water. The quantity of water vapor enclosed in the air can vary significantly.

Fig. 11 shows the comparison graph between the temperature and humidity levels every two hours. This is essential for the system. These comparisons are used for the weather forecasting unit of the proposed model. The temperature and humidity levels are proportional to each other.

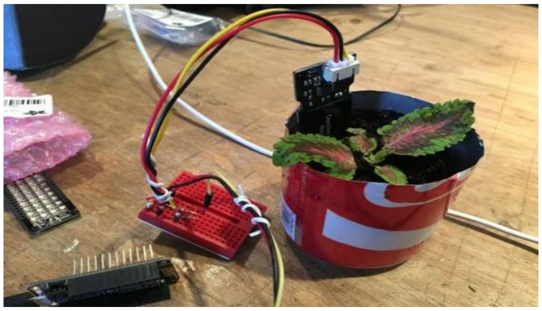
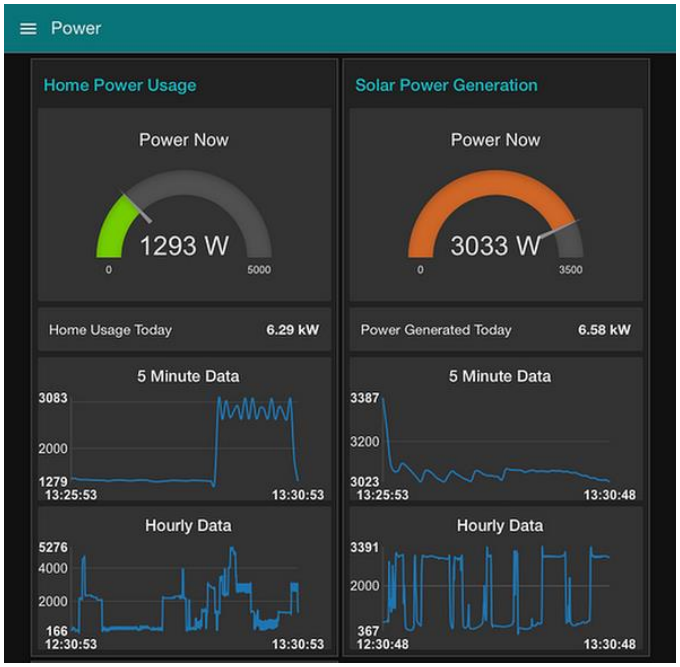


Fig. 8. Water Level Checking Setup.



Home Usage Today 6.29 kW

Power Generated Today 6.58 kW

Fig. 9. Comparison between the Solar Panel Powers to Standard Power.

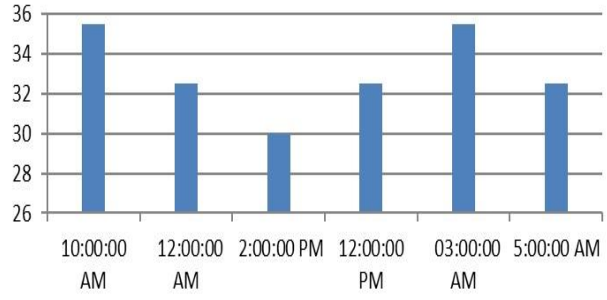


Fig. 10. Graphical view of the Humidity.

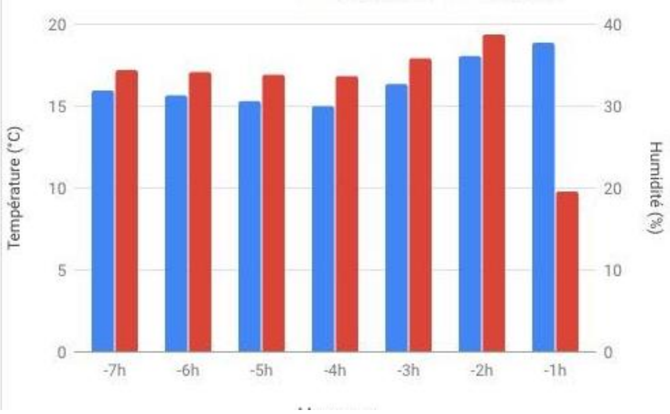


Fig. 11. Comparison between Temperature and Humidity.

Thus, from these results, if the temperature of the surrounding increases, the humidity value of air decreases. Similarly, from the predicted humidity results, if the season is rainy, much water need not be supplied to the plants. This set up can also control the DC fan, motor, and water levels for supporting farmers. Then the measured values of humidity and temperature values from the Arduino MCU are uploaded to the cloud. Then the collected data is communicated uninterruptedly to the farmers live through the GSM to their cell phones. Based on the water level measuring system, the collected data will be sent to the farmer’s cell phones continuously. They can switch on or off their motor based on the collected data from the water level measuring system. This is beneficial for the farmers to control the motors.

Fig. 12 shows the Temperature and Humidity in Real-time. It is essential for plants growth level-based systems. The plant's growths are depended upon both the parameters. If the farmers are not conscious in that the plants will die. Besides, it enables them to monitor their plants from within the comforts of their homes. It will also help to ensure that the plants never suffer from over-watering. Again, the system is beneficial when it comes to water scarcity problems. The values are uploaded within the stipulated time period through the ESP-8266 Wi-Fi system. Then, from the cloud, the humidity and temperature standards are measured using one ThingSpeak platform from anywhere.

Fig. 13 shows the Real-time temperature and humidity of the proposed system. This figure shows the flow of the system functionality wherever HTU 211D SENSORS offers live readings of temperature and wetness at the same time to the microcontroller that sends these reading through the Wi-Fi module over the web to the ThingSpeak cloud.

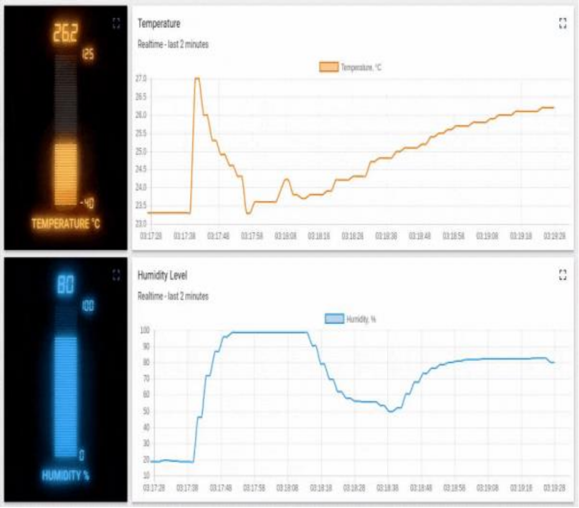


Fig. 12. Temperature and Humidity in Real-Time.

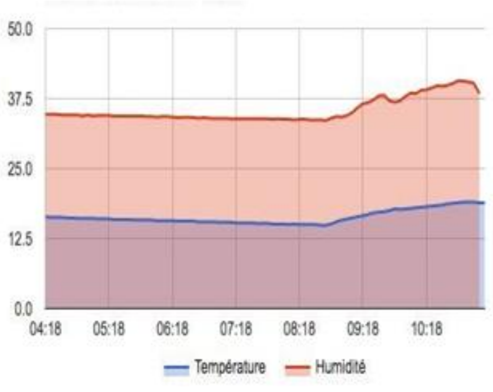


Fig. 13. Real-Time Temperature and Humidity.

Fig. 14 shows the graphical view of the temperature measurement every two hours. Temperature measurement is a widespread technique in IoT based agricultural monitoring system. Its importance cannot be overstated, whether it is summer or winter, spring or autumn. So, HTU 211D temperature sensor is used in the proposed work environment to sense and monitor the temperature. The temperature of that place is monitored with the help of the internet using IoT.

Fig. 15 shows the measured atmospheric pressure of the proposed model. Temperature monitoring is employed in various applications like temperature, pressure, flow rate, and capacity. In the agricultural area, temperature monitoring is essential because based on the temperature, the plants' growth and photosynthesis happen. So, to monitor the agriculture field, as well as to save the plants from overheat death and to reduce the human effort in this work, the effective utilization of IoT is proposed.

The weather forecasters are essential in the agricultural field. The plant's growths are depended on environmental changes only. Fig. 16 shows the Weather forecast output of the proposed work. It is necessary for the time of agriculture. The farmers can watch all the things from their home. In this work, the humidity, light, wind, and the percentage of rainfall is in the same area. It can monitor and arrange the necessary actions from their own place.

Fig. 17 shows the proposed hardware setup of the humidity and temperature using the Arduino Mega 2560. IoT web based temperature monitoring is a type of temperature recorder that monitors the temperature of the field. Then, it stores the data in a database and displays the live temperature on the website through a web server. The system will continuously monitor the temperature condition of the particular area at anytime and anywhere using the web. The primary reason for promoting this proposed model is user-friendly and makes agriculture smart.

Fig. 18 shows the constant temperature maintenance of the system enhances the process of agriculture at the time of the harvesting process. In this work, the constant humidity and temperature control are necessary.

Fig. 19 shows the single-day temperature graph of the proposed system. It is beneficial for the weather forecasting and the day by day weather monitoring process.

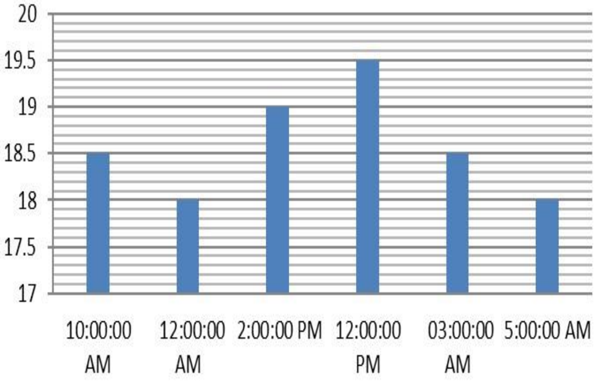


Fig. 14. Graphical view of the Temperature Measurements Every Two Hours.

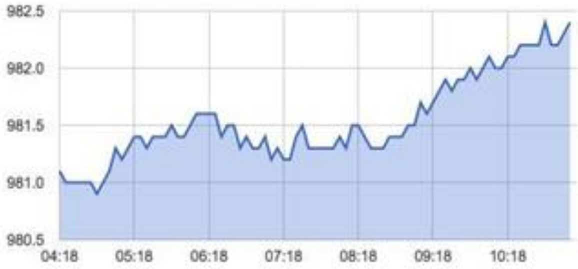


Fig. 15. Measured Atmosphere Pressure.

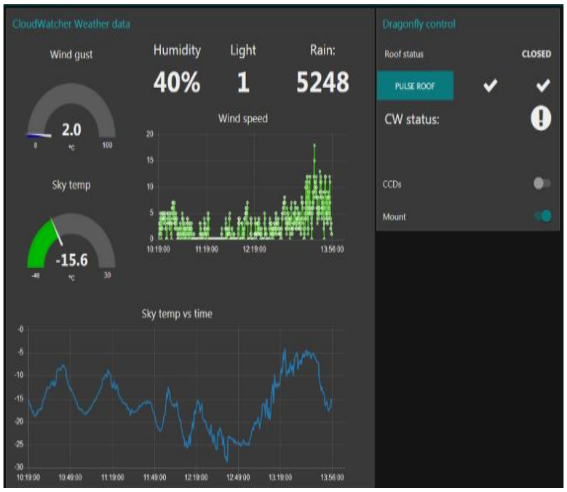


Fig. 16. Weather Forecast Output.

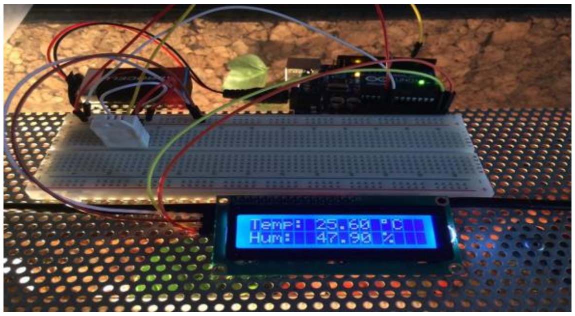


Fig. 17. Proposed Setup of Humidity and Temperature using Arduino Mega 2560.

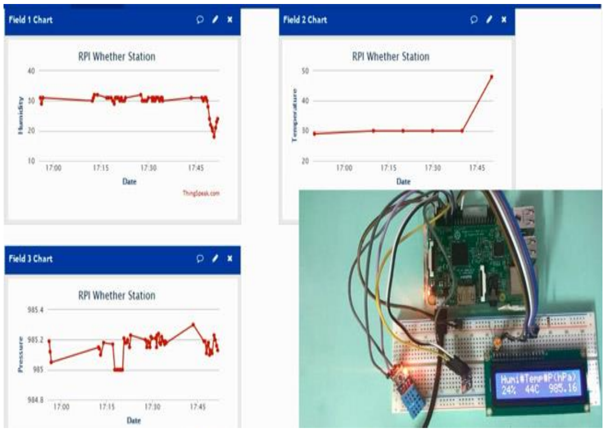


Fig. 18. Constant Temperature Maintenance.

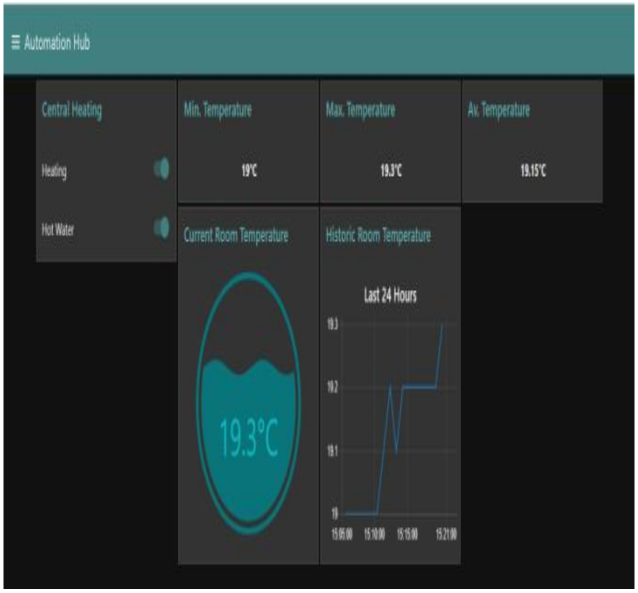


Fig. 19. One Day Temperature Graph.

Table I portrays the comparison between the state of the art methods. In temperature monitoring in the proposed model utilizes sensor monitoring and controlling with Arduino microcontroller. Power usage the Solar panel is used for power saving purpose. In this work, the sensor monitoring and controlling with Arundino controller is used for monitoring the pressure. Finally, the weather conditions and all the above said monitoring plots are done with the ThingSpeak platform.

Fig. 20 shows the DC motor control based on the water level of the soil. The water levels are measured with the water level checking set up shown in Fig. 8. Moreover, from the measured values of the Arduino control, the DC motors are controlled. It is an excellent method for farmers. The proposed work is beneficial for physically challenged peoples who are doing agriculture. It will make a great revolution in agriculture.

Table II shows the comparison between the blockchain models with the IoT system. IoT is centralized with a low latency system. IoT main advantage is the large system adaptation model. That is the reason behind the proposed model.

TABLE. I. COMPARISON BETWEEN STATE OF THE ART METHODS

Methods	Existing	Proposed
Temperature Monitoring	Sensor monitoring	Sensor monitoring and controlling with Arundino
Solar Power usage	Nil	The solar panel is used
Controlling systems	Nil	DC motor monitoring and controlling by water level
Humidity Monitoring	Sensor monitoring	Sensor monitoring and controlling with Arundino
Pressure Monitoring	Sensor monitoring	Sensor monitoring and controlling with Arundino
Weather forecasting	Nil	With ThingSpeak platform

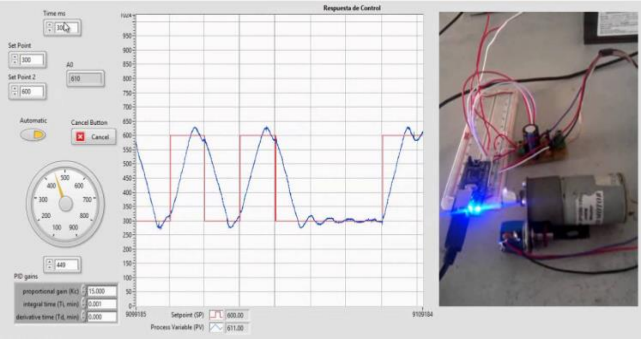


Fig. 20. DC Motor Control by Water Level.

TABLE. II. COMPARISON BETWEEN BLOCKCHAIN MODELS WITH IOT SYSTEM

Blockchain	IoT
Decentralized	Centralized
Resource consuming	Resource restricted
Block mining is time-consuming	Demand low latency
Scale poorly with a large network	IoT considered to contains a large number of devices
High bandwidth consumption	IoT devices have limited bandwidth and resources
Has better security	Security is one of the big challenges of IoT

V. CONCLUSION AND FUTURE SCOPE

IoT based temperature and humidity detecting device provides an efficient and definitive system for monitoring agricultural parameters. The system also provides a corrective movement or decision-making system. IoT based monitoring of area is a handiest, but it also allows the consumers to research the correct modifications within the surroundings and for taking possible action. It is inexpensive and consumes much less electricity. The Gross Domestic Product (GDP) per capitals in agriculture can be multiplied and helps to add our need parameters.

This set up can also control the DC fan, motor, and water levels for supporting farmers. Then the measured values of humidity and temperature values from the Arduino MCU are uploaded to the cloud. Then the collected data are transferred to the farmers live through the GSM to their cell phones. Based on the water level measuring system, the collected data are sending to the farmer’s cell phone continuously. They can switch on or off their motor based on the collected data from the water level measuring system. It is beneficial for the farmers to control the motors as well as can watch their plants from their house. Moreover, also it will help the plants from the overwatering. This system is beneficial for water scarcity problems.

IoT based system can be extended for controlling extraordinary electronic and electric devices from remote locations. Moreover, the system also can be extended for finding the moisture of soil and the farm monitoring for animals growth.

In the future, the extensive Arduino system can put into practice as agriculture automation system and weather-based fertilizer flower and monitor the value of the plants' growth via the mobile application. IoT based systems are a vital step in sympathetic, relevance growth, accomplishment, and serve as a construction block for a numeral of practical modernization technique controller.

ACKNOWLEDGMENT

The author thanks King Abdulaziz University for this work.

REFERENCES

[1] Bhargav Goradiya, and H. N. Pandya, "Real time Monitoring & Data logging Systemusing ARM architecture of Raspberry pi & Ardiuno UNO" International Journal of VLSI and Embedded Systems-IJVES. ISSN: 2249 – 6556. Vol 04, PP: 513-517, July 2013.

[2] M. Rahaman Laskar, R. Bhattacharjee, M. Sau Giri, and P. Bhattacharya, "Weather Forecasting using Arduino Based Cube-Sat", Twelfth International Multi-Conference on Information Processing (IMCIP) – 2016.

[3] Vinayak Appasaheb Pujari, M. M. Raste, and A. A. Pujari, "Cost Effective Automatic Weather Station-a Review", International Journal of Electrical and Electronics Engineers (IJEEE)-Vol. No. 8 Issue 01, January-June 2016.

[4] C. H. Chavan, and V. Karande, "Wireless Monitoring of Soil Moisture, Temperature and Humidity using Zigbee in Agriculture", International Journal of Engineering Trends and Technology (IJETT)-Volume 11 Number 10 – May 2014.

[5] Mayur Randhir, R. R. Karhe, "Monitoring Of Environmental Parameters by Using Cloud Computing" International Journal of Computer Science Trends and Technology (IJCST) – Volume 3 Issue 3, PP: 151-155. May-June 2015.

[6] Nelson Gonzalez, Charles Miers, Fernando Red´ıgolo, Marcos Simpl´ıcio, Tereza Carvalho, Mats N¨aslund and Makan Pourzandi, "A quantitative analysis of current security concerns and solutions for cloud computing" Journal of Cloud Computing: Advances, Systems and Applications 2012, 1:11.

[7] Mahesh D. S, Savitha S, and Dinesh K. Anvekar, "A Cloud Computing Architecture with Wireless Sensor Networks for Agricultural Applications"International Journal of Computer Networks and Communications Security Vol.2, No.1, January 2014, 34–38 Available online at: www.ijcncs.org ISSN 2308-9830.

[8] C. H. Chavan, and P. V.Karande, "Wireless Monitoring of Soil Moisture, Temperature & Humidity Using Zigbee in Agriculture" International Journal of Engineering Trends and Technology (IJETT) – Volume 11 Number 10 - May 2014.

[9] Basil Ahammed, Design & Implementation of Smart House Control Using LabVIEW, International Journal of Soft Computing and Engineering (IJSCE), 1 (6), 2012.

[10] Dingrong Yuan, Shenglong Fang, Yaqiong Liu, The design of smart home monitoring system based on WiFi electronic trash, Journal of Software, 9 (2), 2014, 425-428.

[11] Jiansheng PENG W.L, Qiwen HE, Design of smart home system based on the wireless MCU CC2510, Journal of Hechi University, 10, 2008.

[12] Patricio G, Gomes L, Smart house monitoring and actuating system development using automatic code generation, Industrial Informatics, 7th IEEE International Conference, 256-261, 2009, 23-26. Vinay Sagar K.N, Kusuma S.M, Home Automation using Internet of Things, IRJET, 02,2015.

[13] Girish Birajdar "Implementation of Embedded Web Server Based on ARM11 and Linux using Raspberry PI" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-3 Issue-3, July 2014.

[14] Roselle B. Anire et al., "Environmental Wireless Sensor Network using Raspberry Pi 3 for Greenhouse Monitoring System", IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), 2017.

[15] KonstantinosTzortzakis, et al., "Wireless Self Powered Environmental Monitoring System for Smart Cities based on LoRa", Panhellic Conference on Electronics and Telecommunications (PACET), 2017.

[16] Munsyi et al., "An Implementation of Data Exchange Using Authenticated Attribute-Based Encryption for Environmental Monitoring", 2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC).

[17] Somansh Kumar, "Air Quality Monitoring System Based on IoT using Raspberry Pi", International Conference on Computing, Communication and Automation (ICCCA 2017).

[18] Cho ZinMyint, Lenin Gopal et al., "WSN-based Reconfigurable Water Quality Monitoring System in IoT Environment", 2017 14th International Conference on Electrical Engineering or Electronics, Computer, Telecommunication and Information Technology.

[19] SanketSalvi, Pramod Jain et al., "Cloud Based Data Analysis and Monitoring of Smart Multi-level Irrigation System Using IoT" International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), 2017.

[20] Alif Akbar Pranata, Jae Min Lee et al., "Towards an IoT-based Water Quality Monitoring System with Brokerless Pub/Sub Architecture" IEEE Transaction on Instrumentation and Measurement 2017.

[21] HakanUcgun, et al., "Arduino Based Weather Forecasting Station", 2nd International Conference on Computer Science and Engineering 2017.

[22] J.Cabra, D.Castro et al., "An IoT approach for Wireless Sensor Networks applied to e-health environmental monitoring", 2017 IEEE International Conference on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social computing and IEEE Smart Data.

[23] Pablo Velasquez, et al., "A low-cost IoT based Environmental Monitoring System. A citizen approach to pollution awareness", 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON).

[24] Nikolas Vidakis et al., "Environmental Monitoring through Embedded System and Sensors", 52nd International conference on Power Engineering Universities, 2017.

[25] Shirazi et. al.. "Damage identification using wireless structural health monitoring system through smart sensor application", International Journal of Advanced and Applied Sciences 4.2. 2017: 38-43.

The Use of Geospatial Technology for Epidemiological Chagas Analysis in Bolivia

Natalia I. Vargas-Cuentas[1], Alicia Alva Mantari[2], Avid Roman-Gonzalez[3]
Image Processing Research Laboratory (INTI-Lab), Universidad de Ciencias y Humanidades, Lima, Peru

***Abstract*—Chagas disease is caused by the parasite Trypanosoma Cruzi and transmitted by the Vinchuca. Bolivia is the country with the highest prevalence in the South American region; for example, in 2015, there was a prevalence of 33.4%. This disease causes severe intestinal and cardiac problems in the long term, 30% of the cases register cardiac symptoms, and 10% have alterations in the esophagus or colon. This research aims to analyze the relationship between environmental factors and Chagas outbreaks in an area of Bolivia to identify the environmental conditions in which the disease is transmitted, using epidemiological, meteorological data and also environmental indexes extracted from Landsat 8 satellite images. Through a Principal Components Analysis (PCA) of the environmental indexes extracted from the satellite images and the meteorological information, has been found that the environmental conditions that have a correlation with the occurrence of cases are: temperature, relative humidity, visibility, Normalized Difference Soil Index (NDSI) and Modified Normalized Difference Water Index (MNDWI).**

Keywords—Trypanosoma Cruzi; Vinchuca; Landsat 8; PCA; Normalized Difference Soil Index (NDSI); Modified Normalized Difference Water Index (MNDWI)

I. Introduction

In the world, the environmental characteristics and climatic changes of the different ecosystems have a decisive influence on some diseases that affect man, especially those associated with vectors such as Chagas.

Chagas disease is one of the most worrisome vector diseases in Latin America; the World Health Organization (WHO) declares it as one of the most critical public health problems in America.

According to [1] since 1990, significant success has been achieved in the control of the parasite and the vector in Latin America. [2] In the last 20 years, different Initiatives achieved a reduction in the transmission by domestic vectors as well as blood transfusions. The burden of Chagas disease has been reduced significantly (from around 30 million to 7 -8 million). [3] Chagas disease is caused by the flagellated protozoan, Trypanosoma Cruzi, which is transmitted to humans and other mammals mainly through the droppings of blood-sucking hemipteran insects on their host immediately after feeding.

The transmission of the disease to humans is due to the ability of the insect to explode and adapt to environments in different areas. The distribution of the disease is not uniform. Previous studies like [4] show environmental differences in the situation of the problem of this disease. For the planning of protection measures, it is essential to know how this disease is distributed at the provincial level. Also, it is necessary to identify the risk that each sector has, in order to assess the needs for control, care and planning according to local needs.

The conception of the study of space as a new perspective for epidemiological analysis in order to identify possible epidemiological outbreaks and the development of the disease as a global process in a population space is called panoramic epidemiology. In [5] for example, the reality of Argentina is described regarding the occurrence of dengue outbreaks, in order to carry out epidemiological surveillance for the control of the disease-causing vector. In work an exploration of the problem is carried out from the discipline of panoramic epidemiology to analyze the distribution, ecology, and behavior of the vector. Besides, an exploration of techniques in classification and image processing is developed, in order to generate a temporal space dispersion model of the vectors. Predictive maps were developed, of the focal density of Aedes aegypti, based on environmental information derived from SPOT 5 HRG1 high-resolution spatial images and images of average spatial resolution of surface temperature derived from Landsat 5 satellite information. A predictive model of biweekly aedic indices was generated, which was based on macro-environmental information from Landsat TM and ETM sensors, and vector monitoring and control information. Then, space-time epidemiological patterns and population parameters of the vector were estimated; the effectiveness of vector control measures during the outbreak was also estimated.

Also in [6] as part of the investigation of CONAE (National Commission for Space Activities) and the Ministry of Health of Argentina, a project was developed to implement a Dengue Early Warning computer system, which calculates Dengue's environmental risk in Argentina. The risk is assessed based on a static component related to historical environmental conditions and a dynamic component related to current environmental conditions. The stratification of the localities of Argentina is shown as a result according to their environmental risk of Dengue.

The study published by Neteler [7] analyzes the environmental conditions of the spread of the Asian tiger mosquito Aedes Albopictus in Europe, as part of a constant concern for public health due to outbreaks transmitted by vectors. Data from the reconstructed daily satellite time series (MODIS Land Surface Temperature maps, LST) of northeastern Italy were used, as well as the annual average temperature by areas, in order to reconstruct areas and compare them with the current known distribution of Ae. Albopictus in northeastern Italy. LST maps show peculiar microclimatic characteristics. From this data, surface maps are reconstructed to predict the infection areas of the vector with an accuracy of 200m pixels. This is an important study that could be applied to other arthropod species in which the temperature is a relevant factor.

According to [8], the dispersal capacity of a disease that is transmitted by vectors depends on multiple environmental, climatological, biophysical and social variables [1]. The information extracted from the satellite images will allow us to observe the climatic changes related to epidemiological outbreaks indirectly [8]. With satellite images of different spectral bands, different environmental parameters can be identified and measured [2].

Images from the Landsat 8 satellite and the OLI-TRIS sensors will be used. With this analysis, we try to describe if there are tools to generate models that describe the propagation habitat of vectors that are the transmitters of endemic diseases.

It is from these images that for certain period of time we analyze the indices that are extracted from them to generate a model that gives us a simplified representation of reality. This model allows finding variables to understand the values of the incidence and prevalence, as well as the process of transmission of the disease in a certain period.

The importance of this work is based on the fact that space is a privileged place. For this reason, satellite images can give us an overview of the area that one wishes to analyze. The idea is to have global information about areas with a possibility of epidemiological outbreaks. All this information can contribute to better decision making when planning prevention tasks and epidemiological management.

The limitations of the present study are that the temporal resolution of the Landast 8 satellite is 16 days, which implies having only one image to analyze per month, in addition in the study area there are not many weather stations to complement the data obtained from the satellite images.

The content of this work is broadly divided into three main parts: in the first part the Chagas disease is described. Afterwards a spatial temporal epidemiology analysis is developed. Finally an analysis of the relationship between all the data extracted is implemented.

II. Description of Chagas Disease

This section seeks to expose the basic aspects of Chagas disease, besides shows the growing presence of Chagas in South America and mainly in Bolivia in recent years.

A. Definition and basic Aspects of the Disease

The disease of American trypanosomiasis or Chagas, it is a type of zoonosis, a disease typical of the encounter with animals of contact with people, typical of this continent, because has existed in Latin America since before the conquest.

In 2018 WHO reported 6 or 7 million people infected in 21 countries by the parasite Trypanosoma Cruzi [3], which causes this disease and 90 million people are at risk of contracting the infection.

Chagas can be transmitted through the consumption of contaminated food, stinging of the infected vector, blood transfusion, transmission of the pregnant mother to the child, transplantation of infected organs.

The process of the disease has an incubation period of 4 to 10 days, mostly asymptomatic. Then the three phases are distinguished, acute, indeterminate and chronic.

The acute phase lasts between 2 to 4 months, is asymptomatic, so the diagnosis is difficult at this stage, it can also present very mild symptoms. It is characterized by the high concentration of parasites in blood. Some of the symptoms presented on occasion are: fever, headache, swollen lymph nodes, slight pallor, muscle pain, breathing with difficulty, abdominal pain.

The indeterminate phase begins after the acute phase and lasts for 8-10 weeks, regardless of the symptomatology. During this phase, the disease is usually asymptomatic. Although it still has quantities of the parasite in the blood, it is only possible to diagnose it in 20% to 60% of the cases. It is during this phase that the contagion becomes a great problem of public health, due to the ignorance of the presence of the disease.

During the chronic phase the parasite usually hides in the cardiac and digestive organs, where they multiply and begin to generate irreversible tissue damage. Only 30% of cases have any cardiac symptoms, and about 10% have alterations in the esophagus or colon. Moreover during this stage the amount of parasitaemia is low. The symptoms will depend on the damage of the parasite in the host organ. Chronic myocarditis is the most common heart affection due to Chagas disease.

Among the symptoms that derive from the presence of the parasite in the heart, are myocardial damage, arrhythmia and heart failure.

Ventricular fibrillation is probably the most frequent mechanism of sudden death in chronic Chagasic patients.

The disease registers a percentage of mortality that oscillates between 55% and 65%, mainly due to cardiac complications.

The treatment is especially useful in the acute phase of the disease, where it can cure up to 100% of cases. The effectiveness of the treatment decreases in an inverse manner over time, because the parasite can cause irreversible damage to some organs.

This treatment is based mainly on benznidazole and nifurtimox capable of killing the parasite depending on the phase of infection.

In some chronic cases, the treatment although it may not kill the parasite completely, can prevent or delay the progression of the disease, although it is important to consider the adverse consequences of dosing them over time.

Unfortunately there is no vaccine against Chagas disease, but the prevention method is the control of the vector in the areas of greatest incidence.

The T. Cruzi can infect several species, depending on the geography of the area, for this reason it is recommended for its prevention the fumigation of areas of greater risk, cleaning and improvement in the conditioning of homes, hygiene, adequate conservation of the food and develop regular serological tests in risk areas.

B. *Presence of Chagas in Latin America*

Chagas is a parasitic disease that has become a public health problem in Latin America [9], because it has a presence in at least 21 countries in the region.

In America the infection is located from the south of the United States to Argentina and Chile. According to [10] due to this disease in this continent about 50,000 people die each year and more than 100 million people are at risk of becoming infected.

In 2013 according to the World Health Organization [11], it has been estimated that of the 8 million people carrying T. Cruzi, the largest number of cases are concentrated in Latin America, although important number of cases are also calculated in the United States, Canada, Spain, Japan and Australia.

Chagas disease in Latin America can be associated to multiple factors, including dwelling houses built with materials such as adobe, mud and straw, this type of structures can be seen in rural and suburban areas of the region.

There are around of 140 species of vinchucas in the world, of which the majority are distributed throughout the American continent, only a few species are present in Asia, Africa and Australia. [11].

In Latin America, together with the Pan American Health Organization (PAHO) and the World Health Organization (WHO), a horizontal technical cooperation strategy was developed among countries to prevent and control Chagas disease in the region. [3] For example, these initiatives have been developed in the Southern Cone (1992), the Andean countries (1998) and the Amazonian countries (2003), among other cooperation strategies in the region.

This cooperation has contributed to the elimination of allochthonous species of vectors, the detection of congenital cases, the reduction of prevalence in children and the improvement of the quality of treatment of infected and sick people, among others.

In the specific case of South America, the vector Triatoma Infestans predominates in Argentina, Bolivia, Brazil, Chile, Paraguay, Peru and Uruguay. [12] Besides in Colombia and Venezuela the predominant vector is R. Prolixus.

According to figures calculated by WHO in 2010, Bolivia is the country with the highest incidence of Chagas in South America.

C. *Presence of Chagas in Bolivia*

According to the Institute of Development Health Research (IINSAD) of the Universidad Mayor de San Andrés (UMSA) in [13] there are 140 species of vinchucas in the world, in Bolivia there are 21 types of vinchucas identified, the most common vinchuca is the Triatoma Infestans, which is responsible for the largest number of Chagas (CH) cases recorded in the country.

The country with the largest dispersion area of the CH vector (Triatoma Infestans) is Bolivia. [14] In the endemic vector map of the Ministry of Health Chagas disease is dispersed in approximately 60% of the Bolivian territory.

In the country, three endemic zones were identified: the valley area comprised by the departments of Cochabamba, Chuquisaca, Tarija and Potosí, the Chaco area comprised by the departments of Santa Cruz, Chuquisaca and Tarija, and finally the Amazon area departments of Beni, Pando, part of Santa Cruz, north of La Paz and north of Cochabamba.

According Médecins Sans Frontières (MSF) in [15] the South American region, Bolivia register more than 600,000 people infected with the disease. Besides, an average of 8,000 new cases of Chagas is registered each year.

The Chagas Prevention Program of the Ministry of Health of Bolivia indicates in [13] that the country register the highest prevalence in the South American region, for example in 2015 there was a prevalence of 33.4%.

In 2016 the Ministry of Health of Bolivia recorded 17,892 new cases, in the department of Santa Cruz, 57.72% of the total cases of Chagas were concentrated, followed only by the department of Cochabamba, which concentrated a total of 23.13% of infected persons [14].

III. Spatial Temporal Epidemiology Analysis

For this research it is necessary to select a study area of interest in Bolivia, acquire the different types of data and performs the treatment of the satellite images, this section exposes these main procedures.

A. *Selection of the Study Area*

Chagas has become an emerging disease in Bolivia, since it is dispersed in approximately 60% of the Bolivian territory. The CH vector (Triatoma Infestans) it has found in the warm departments of the country, such as Santa Cruz, Tarija, Chuquisaca and Cochabamba, the adequate conditions to prevail and transmit the disease.

We can also observe that Chagas disease has presence in the nine departments of Bolivia. For this reason we must select among the most affected departments of the country by this viral disease and where there is a high number of cases.

In order to select a suitable research study area the National Program of Preventive and Control of Chagas in Bolivia was revised, where the epidemiological situation of the country in 2017 can be observed:

As we can observe in Table I, there is a high number of Chagas cases in three departments: Santa Cruz, Cochabamba and Tarija. Also we can observe a moderate number of cases in three departments: Potosi, Chuquisaca and Beni. Finally there is a low number of cases in three departments: La Paz, Pando and Oruro.

TABLE. I. TOTAL CASES OF CHAGAS DISEASE IN BOLIVIA IN 2016

Departments	**MALE**	**FEMALE**	**TOTAL**
BENI	82	85	167
CHUQUISACA	350	483	833
COCHABAMBA	1,705	2,433	4,138
LA PAZ	79	131	210
ORURO	4	17	21
PANDO	6	5	11
POTOSI	408	448	856
SANTA CRUZ	4,334	5,994	10,328
TARIJA	545	783	1,328
TOTAL:	7,513	10,379	17,892

In the nine departments of Bolivia there is a total of 15911 cases of Chagas that have been recorded throughout the year of 2017, as can be seen in Fig. 1.

According to the collected data, the department of Santa Cruz, is the department that has registered the most high number of cases in 2017, with a total of 7777 cases, in second place is the department of Cochabamba with a total of 4794 cases and in third place is the department of Tarija with a total of 1457 cases of Chagas.

It can be observed that in eight of the Bolivian departments the number of female cases is greater. The departments of Santa Cruz and Cochabamba are those that report the highest number of female cases.

In order to correctly identify the study area, the incidence rate and point prevalence of Chagas disease in the nine departments of Bolivia were calculated in Table II.

As can be seen in Table III, the departments with the highest incidence rate and prevalence percentage are: Tarija, Santa Cruz and Cochabamba respectively. However, the department of Santa Cruz has the highest number of cases, because it has a larger population (more than three millions of inhabitants), for this reason Santa Cruz is selected as the area of analysis.

Since the department of Santa Cruz is the largest department in Bolivia, is necessary to select only one specific area to carry out the research.

The department of Santa Cruz has 15 provinces, as can be seen in Fig. 2, of which according to the Ministry of Health in Bolivia 14 provinces have the presence of Chagas in 2017, the information collected can be observed in Table III.

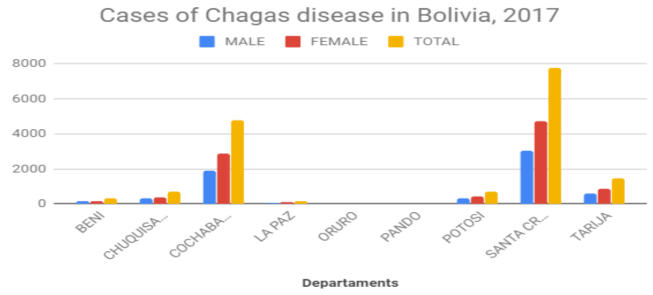


Fig. 1. Chagas Disease in the Bolivian Departments in 2017.

TABLE. II. TOTAL CASES OF CHAGAS DISEASE IN BOLIVIA IN 2017

	TOTAL		
Departments	**MALE**	**FEMALE**	**TOTAL**
BENI	153	147	300
CHUQUISACA	293	379	672
COCHABAMBA	1,915	2,879	4,794
LA PAZ	60	99	159
ORURO	5	6	11
PANDO	4	15	19
POTOSI	317	405	722
SANTA CRUZ	3,050	4,727	7,777
TARIJA	603	854	1,457

TABLE. III. INCIDENCE RATE AND POINT PREVALENCE OF CHAGAS

Departments	**TOTAL**	**Population**	**Incidence rate (100,000 inhabitants)**	**Point prevalence (%)**
BENI	300	462,081	64.924	0.065
CHUQUISACA	672	621,148	108.187	0.108
COCHABAMBA	4,794	1,943,429	246.677	0.247
LA PAZ	159	2,862,504	5.555	0.006
ORURO	11	531,890	2.068	0.002
PANDO	19	139,018	13.667	0.014
POTOSI	722	880,651	81.985	0.082
SANTA CRUZ	7,777	3,151,676	246.758	0.247
TARIJA	1,457	553,373	263.294	0.263

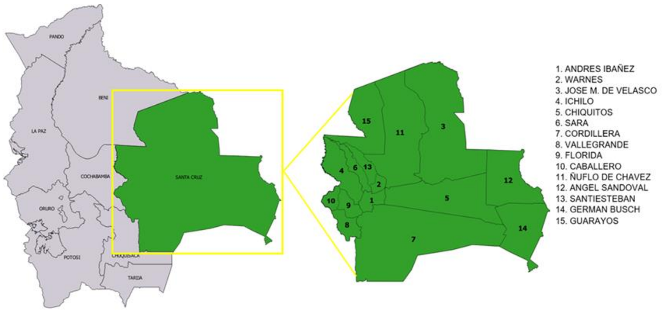


Fig. 2. Provinces of the Department of Santa Cruz.

As can be seen in Table IV, the provinces with the highest number of cases in the department of Santa Cruz are: Andres Ibañez (with 6,315 cases), Obispo Santistevan (with 811 cases) and Warnes (with 330 cases).

As can be observed in Fig. 3, in 9 of the 14 provinces affected by Chagas in Santa Cruz it can be seen a higher number of cases in the female population, except in the province of German Busch where it can be seen that the number of cases is the same in the female and male population.

Andrés Ibañez province represents 81.9% of the total number of cases registered in the department of Santa Cruz throughout 2017, followed in the second place by the province of Obispo Santistevan, which represents 11.2% of the cases and in the third place is the province Warnes which represents 2.2% of the total cases, as can be seen in Fig. 4.

TABLE. IV. CASES OF CHAGAS IN THE PROVINCES OF SANTA CRUZ IN 2017

	TOTAL		
Province	**MALE**	**FEMALE**	**TOTAL**
ANDRES IBAÑEZ	2,498	3,817	6,315
CHIQUITOS	11	46	57
CORDILLERA	14	18	32
FLORIDA	1	4	5
GERMAN BUSCH	4	4	8
GUARAYOS	13	16	29
ICHILO	14	15	29
MANUEL MARIA CABALLERO	5	2	7
ÑUFLO DE CHAVEZ	43	40	83
OBISPO SANTISTEVAN	341	470	811
SARA	31	19	50
VALLEGRANDE	5	12	17
VELASCO	3	1	4
WARNES	67	263	330

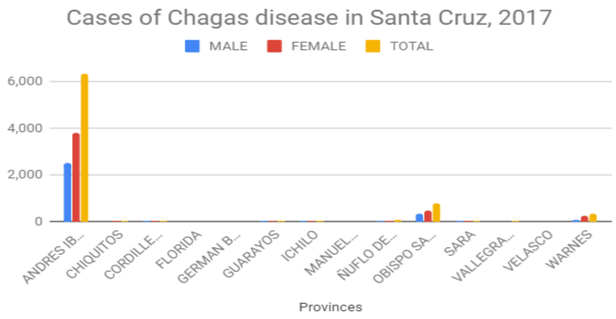


Fig. 3. Cases of Chagas Disease in the Provinces of Santa Cruz.

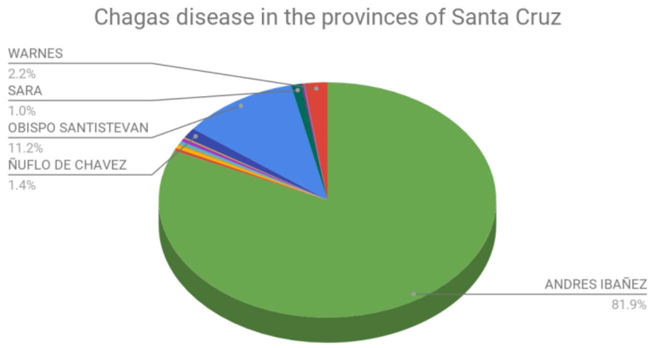


Fig. 4. Chagas Disease in the Provinces of Santa Cruz.

Thus these three provinces of Santa Cruz, representing the 95.3% of Chagas cases, are selected as the study area analyzed in the present project, as can be observed in Fig. 5.

The province Andrés Ibañez it is the most important province of the department of Santa Cruz, has five municipalities, which are: Cotoca, El Torno, La Guardia, Porongo and Santa Cruz de la Sierra. It is located at the coordinates: 17°50'00"S 63°18'00"W. Besides, the province Obispo Santistevan has five municipalities, which are: Montero, Saavedra, Mineros, Fernández Alonso and San Pedro. It is located at the coordinates: 16°30'00"S 63°30'00"W. Finally the province Warnes has two municipalities, which are: Warnes and Okinawa. It is located at the coordinates: 17°20'00"S 63°00'00"W.

B. Data Acquisition

For the present work, one needs three types of data on which the analysis, correlations, and conclusions will be obtained. These data are: satellite images, epidemiological data, and meteorological data. The description of each of these data is detailed below.

a) Satellite images: There are different satellites that provide us multispectral images, from which it is possible to collect environmental information, which in this case will be used to identify the environmental parameters that influence in Chagas outbreaks in three provinces of Santa Cruz in Bolivia.

For this research we collected information from the Landsat 8 satellite, launched on February 11, 2013, these satellite images are downloaded for free from the web: https://earthexplorer.usgs.gov/.

The Landsat 8 satellite has two sensors, OLI and TIRS. Among the most important features observed in [16], can be mentioned that it is heliosynchronous, at an orbital height of 705 km, WRS-2 (Worldwide Reference System), with an inclination of 98.2°, a temporal resolution of 16 days and having a radiometric resolution of 12 bits, spatial resolution of 30m, and spectral resolution of nine bands 5 of them in the visible field and the others in the non-visible field.

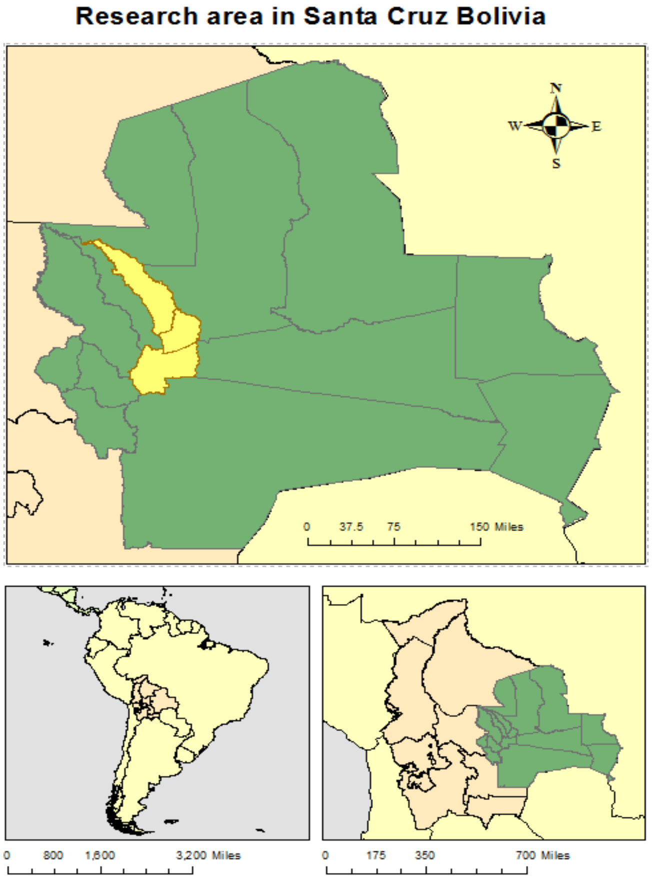


Fig. 5. Research Area in Santa Cruz Bolivia.

The images selected for this research were compiled from the USGS website. It was identified that the scene that covers 83% of the study area, is in the location Path 231 and Row 72, according to the WRS system.

The scenes of all the months of the year 2017 were analyzed, choosing only those that have a minimum cloud percentage, the details can be observed in the following Table V.

Those images registering a minimum cloudiness of less than 40% were selected. The selected scenes have an average cloudiness of 21.80%, the selected months are: February, July, August, October and November.

Of the five datasets selected, we have been downloaded: 11 raster bands (GeoTIFF), the BQA file (16 bits quality control file), and the metadata file (MTL), in total 14 files.

The spatial subset of the 11 bands of the image is: 7622x7732 pixels, with a vertical and horizontal resolution of 96pp, and a depth of 32 bits. The image corresponding to band 8, is the only one that has a resolution of 15242x15242 pixels, with a vertical, horizontal and depth values similar to those already mentioned.

b) Epidemiological data: The Ministry of Health of Bolivia, through its Departmental Health Service (SEDES), by its acronym in Spanish “Servicio Departamental de Salud”, in a joint effort of the nine departments of Bolivia, has developed and implemented the National Health Information System.

This information system contains fourteen-year data from 2005 to 2018, also it contains different groups of variables, such as immuno-preventable diseases, sexually transmitted diseases and vector-borne diseases, among many others.

The national information system has been developed to maintain an epidemiological surveillance program in all the health establishments of the nine departments of Bolivia. The tool is available on the web portal of the Ministry of Health Bolivia in the informatics tools section available at: https://snis.minsalud.gob.bo/.

This informatic tool allows to perform an epidemiological surveillance of vector-borne diseases, including: Dengue, Zika, Chikungunya and Chagas disease among others.

The tool allows a search of data by departments, provinces, municipalities, networks and health establishments throughout Bolivia and in the different months of the year of interest.

For this study, the epidemiological database corresponding to the 2017 period is used, as can be seen in Table VI.

In the database collected, as can be seen in Table IV, the number of cases of acute Chagas reported in the study area is obtained, but in addition these cases are classified by gender and by age, from less than 6 months to over sixty years of age.

It can be observed in Fig. 6 that the age group with the highest number of cases of Chagas is the group of 60 years and over, with 2,560 cases. Moreover, the age group with the least presence of Chagas is the group of six months to less than one year, with 40 cases.

The database compiled also contains data on the types of health facilities where cases of Chagas were reported in 2017. It is important to mention that in Bolivia there are first, second and third level hospitals, health centers and medical posts among other smaller health facilities.

TABLE. V. SATELLITE IMAGE DATASETS OF THE STUDY AREA

Data set	Path-Row	Date acquired	% Cloud coverage
LC08_L1TP_231072_20170119 _20170311_01_T1	231 - 72	19/01/2017	44.76
LC08_L1TP_231072_20170220 _20170301_01_T1	231 - 72	20/02/2017	33.55
LC08_L1TP_231072_20170308 _20170317_01_T1	231 - 72	08/03/2017	64.03
LC08_L1TP_231072_20170425 _20170502_01_T1	231 - 72	25/04/2017	69.88
LC08_L1TP_231072_20170527 _20170615_01_T1	231 - 72	27/05/2017	71.7
LC08_L1TP_231072_20170612 _20170628_01_T1	231 - 72	12/06/2017	82.26
LC08_L1TP_231072_20170714 _20170726_01_T1	231 - 72	14/07/2017	10.9
LC08_L1TP_231072_20170831 _20170915_01_T1	231 - 72	31/08/2017	5.61
LC08_L1GT_231072_20170916 _20170929_01_T2	231 - 72	16/09/2017	100
LC08_L1TP_231072_20171018 _20171025_01_T1	231 - 72	18/10/2017	20.27
LC08_L1TP_231072_20171103 _20171109_01_T1	231 - 72	03/11/2017	38.69
LC08_L1GT_231072_20171205 _20171222_01_T2	231 - 72	05/12/2017	100

TABLE. VI. DESCRIPTION OF THE COLLECTED EPIDEMIOLOGICAL DATA

Database	Form	Disease	Year	Months	Weeks
Notification for epidemiological surveillance	302 a	Acute Chagas	2017	January to December	52
Monthly notification for epidemiological surveillance	302 b	Acute Chagas	2017	January to December	52

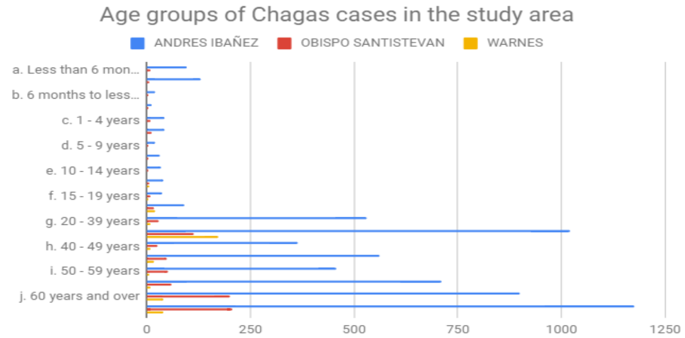


Fig. 6. Age Groups of Chagas Cases in the Study Area.

As it can be observed in Fig. 7 the Chagas disease has been found in 11 types of health facilities, both public and private, belonging to the health network of the Bolivian state. It can be seen that there is a greater presence of Chagas in the Ambulatory Health Centers, on the contrary it can be seen that there is a lower presence of the disease in the poly-medical centers.

Table VII is a summary of the epidemiological information extracted and shows the total cases of Chagas in three provinces selected as the study area. We can only observe the cases of Chagas of certain months, this is due to the cloud coverage, as was previously explained we only select the months of the year with a low percentage of clouds.

c) Meteorological data: The meteorological data are essential since they are the data that we seek to equate with the environmental parameters that will be obtained from the satellite images to correlate them with the epidemiological data.

The meteorological data come from the meteorological stations of the National Service of Meteorology and Hydrology (SENAMHI) of Bolivia (http://senamhi.gob.bo/index.php /inicio).

The extracted data are daily precipitation, daily relative humidity, and the daily visibility.

In Table VIII, one can see a table with the collected meteorological data for February, July, August, October, and November of 2017.

The data collected corresponds to the year 2017 (which corresponds to the year that is being analyzed in the present work with the satellite images) in 852450 SLET meteorological station, since this station is in the area under analysis. One chose February, July, August, October, and November because these months correspond to the satellite images.

C. *Treatment of Satellite Images*

It is necessary to extract the environmental parameters of the satellite images, for this the images must be cropped with the shape of the area of interest, corrections must be made as part of the preprocessing stage and finally an algorithm must be programmed in MATLAB to extract the necessary information from the satellite images.

a) Preprocessing: To begin with the preprocessing of the images, the maps of departmental, provincial and municipal limits were used in shape file format, selecting those that correspond to the area of interest identified in the epidemiological analysis.

The shapefiles were downloaded from the website of GeoBolivia, which were based on the maps created by the Bolivian Geographical Institute, which has the information provided by the Ministry of Autonomies with data updated to April 2015, this divides Bolivia into 339 municipalities.

These geographic information files were created on April 15, and its publication as an available web resource occurred on May 10, 2015, under the OGC protocol: WMS-1.1.1-http-get-map.

After obtaining the shapefile of the area of interest, which as indicated above, comprises three provinces: Andrés Ibañez, Obispo Santistevan and Warnes. The next process was to crop the 11 bands and the BQA raster of each multispectral image using the obtained shapefile, in each subset the spatial resolution of the input satellite image is preserved.

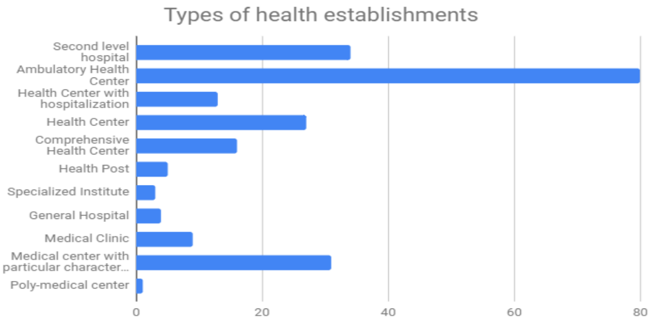


Fig. 7. Type of Health Establishments in the Study Area.

TABLE. VII. CHAGAS CASES IN THE STUDY PROVINCES

Province	Municipality	February	July	August	October	November
Andres Ibañez	COTOCA	3	13	23	23	31
	EL TORNO	5	1	2	9	7
	LA GUARDIA	7	22	15	6	11
	PORONGO	0	0	0	0	0
	SANTA CRUZ DE LA SIERRA	472	531	640	469	466
Obispo Santistevan	FERNANDEZ ALONSO	17	1	1	2	4
	GENERAL SAAVEDRA	0	0	2	2	0
	MINEROS	11	0	0	0	0
	MONTERO	44	50	18	55	26
	SAN PEDRO	3	4	2	18	13
Warnes	OKINAWA	2	0	0	1	1
	WARNES	41	49	6	0	0

TABLE. VIII. Meteorological data for 2017 (852450 SLET meteorological station)

	February	July	August	October	November
Temperature	27.5	25.4	22.4	31.4	28.4
Relative humidity	73.0	60.0	34.0	52.0	67.0
Visibility	11.3	11.4	12.9	6.9	12.9

For achieve this objective QGIS 3.4 Madeira was used, this procedure is important in order to obtain the data only from the area of interest, in this case the area contained only in the path 231 and row 72.

After cutting the raster according to the area of interest, we proceed to perform the correction procedures using the ENVI 5.3 software.

Due to interferences in the satellite instruments, different effects can occur that can affect the images acquired, which is why, in order to acquire valid information from the satellite images, two corrections are made, the first is a radiometric correction and the second an atmospheric correction.

Radiometric correction is a technique used to reduce anomalies caused by the sensor system or the conditions of the image capture, this correction resets the digital values of the image.

In order to change the original multispectral images obtained by the sensors at a radiometric scale, the spectral radiance is calculated [17]. Radiometric correction is an indispensable step in the creation of high-level images for subsequent processes.

The following parameters, observed in Table IX, were obtained after performing the radiometric correction.

It must be remembered that the electromagnetic radiation of the earth is captured through the satellite sensors. This energy is known as radiance [18].

The atmospheric correction is the technique of evaluate and eliminate the atmospheric and terrain distortions that are introduced in the values of radiance that arrive at the sensor from the Earth's surface [19]. The main objective of this correction is recovering the physical parameters of the terrestrial surface including the reflectance of the surface, the visibility of the ground and the temperature.

The model used for atmospheric correction is Fast Line of Sight Atmospheric Analysis of Spectral Hypercubes (FLAASH). For the use of this module in ENVI 5.3, some input parameters that detail the characteristics of the image are required. [20] Besides, the input data type must be a radiance file of floating-point and in the format of Band Interleaved by Line (BIL).

Moreover, for the appropriate atmospheric correction of the image, meteorological data are needed, also information such as height, date and time, the type of sensor that models the relative spectral response. For which it was identified that the average altitude of the study area is 384 m.a.s.l and the meteorological data were extracted from the Trompillo weather station, located at the GPS point (Latitude -17.8, Longitude -63.16).

TABLE. IX. Results of the radiometric correction

Variables	20/02	14/07	31/08	18/10	3/11
Sun Azimuth	82.68	39.1	50.84	76.58	88.72
Sun elevation	57.45	40.64	51.09	63.95	65.63
Cloud cover	33.55	10.9	5.61	20.27	38.69
Earth sun distance	0.99	1.01	1.01	1	0.99

Within the specific configuration for this correction, the multispectral settings must be configured, an important step is to select specific bands for aerosol retrieval, for which the Kaufman-Tanre 1997 method is used [21], these method identify dark pixels used for the visibility estimate, the recommended wavelength ranges for this model is (640 - 680) nm for the lower channel and (2100 - 2250) nm for the upper channel.

In this case, the SWIR2 band (2.21010) is used as the upper channel and the red band (0.6546) as the lower channel.

In summary, the parameters used for the proper configuration of the FLAASH atmospheric correction tool are:

1) *Scene center:* Given by the geographical coordinates of the center of the scene, in this case corresponding to Path 231 and row 72. (Latitude: -17'20"47.42 and Longitude: -63'41"28.03).
2) Sensor type: Multispectral - Landsat8 OLI.
3) Sensor altitude: 705 km.
4) Ground elevation: 0.384 km.
5) Pixel size: 30 m.
6) Date of image acquisition: See Table X.
7) Flight time: See Table X.
8) Atmospheric model: Tropical.
9) Aerosol model: Urban.
10) Visibility: See Table X.

b) Processing: The first step is to identify the environmental parameters that are extracted from the different spectral bands of the satellite images.

In Table XI, we can observe a list of the indexes calculated, their respective definitions [22], [23], [24] and their equations.

We proceeded to program an algorithm to extract the environmental parameters, for this the MATLAB software was used. The algorithm develops three major processes, as can be seen in Fig. 8.

TABLE. X. Data for the Atmospheric Correction

Variable	20/02	14/07	31/08	18/10	3/11
Temperature	27.5	25.4	22.4	31.4	28.4
Time (HH:MM:SS)	14:16:50	14:16:44	14:17:01	14:17:13	14:17:13
VV (average visibility)	11.3	11.4	12.9	6.9	12.9
Average relative humidity	73	60	34	52	67
Atmospheric Model	Tropical	Tropical	Tropical	Tropical	Tropical

TABLE. XI. ENVIRONMENTAL INDEXES

Name	Definition	Equation
NDSI	Normalized Difference Soil Index	$NDSI = \frac{SWIR - NIR}{SWIR + NIR}$
NDMI	Normalized Difference Moisture Index	$NDMI = \frac{NIR - SWIR_1}{NIR + SWIR_1}$
NDVI	Normalized Difference Vegetation Index	$NDVI = \frac{NIR - Red}{NIR + Red}$
NDWI	Normalized Difference Water Index (Content in leaves)	$NDWI = \frac{NIR - SWIR_1}{NIR + SWIR_1}$
MNDWI	Modified Normalized Difference Water Index	$MNDWI = \frac{Green - SWIR}{Green + SWIR}$

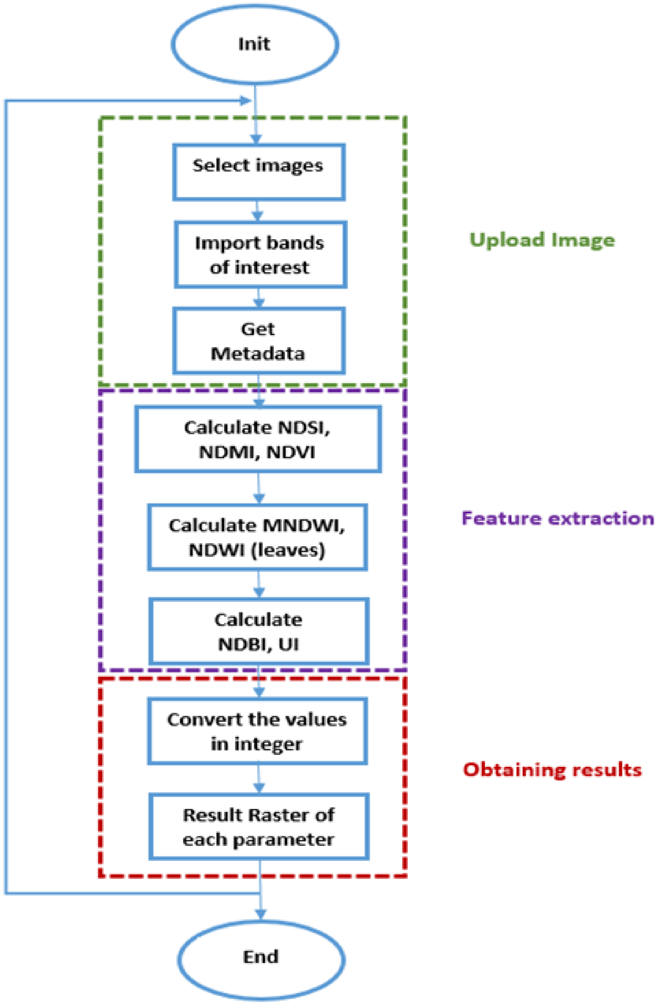


Fig. 8. Flow Diagram of the Processing Algorithm.

Upload the image: In this stage the satellite image must be selected, the bands of interest used for the calculations must be imported and the metadata must be read.

Feature extraction: In this step, different parameters must be calculated, for which different equations must be programmed using different spectral bands of the image, in order to obtain different indexes such as the NDVI, NDSI, among others.

Obtaining results: In this stage the raster resulting from the calculation of each of the parameters must be obtained, an evaluation of the results obtained must be carried out to corroborate that the obtained indexes are within the maximum ranges established in the theory, to confirm that the calculations made are correct.

IV. Analysis of the Relationship of all the Data Extracted

Among the obtained data, one has seven parameters or indices extracted from the satellite images, three parameters from the meteorological stations, and the data related to the cases of Chagas in the area under analysis. The parameters obtained from the satellite images and the data from the meteorological station, both dataset form a total of ten features.

The first analysis to be performed is the calculation of the correlation between the ten selected features.

In Table XII, one can see the Pearson correlation between the eight selected features where a good correlation is shown between the NDMI and the NDWI. The p-value of the correlations previously mentioned (see Table XIII) is 0.0001 in both cases; this value being less than 0.05 confirms the significance of both correlations.

Having ten features, one applies the principal component analysis (PCA) to reduce the space of variables to be analyzed.

After the PCA, it is observed that the workspace is reduced only to four principal components (see Table XIV).

Among the four main components shown in Table XV, they explain the 100% of the data collected. As shown in Table XII and Fig. 9, Component 1 explains more than 94% of the data.

Since Component 1 is the one that practically explains the data thoroughly, we look for the correlation of Component 1 with the epidemiological data that represent the number of cases of Chagas in the area under analysis.

In Table XVI, it can be seen that there is a high correlation (Pearson correlation) between Component 1 and epidemiological data. This absolute value of correlation is equal to 0.8469.

In Fig. 10, we can observe a correlation map between Chagas cases and the results found in the analysis of the principal components extracted from Table XIV.

The characteristics of component 1 obtained by PCA were combined to create heat maps where the blue areas represent high values and the orange areas correspond to low values.

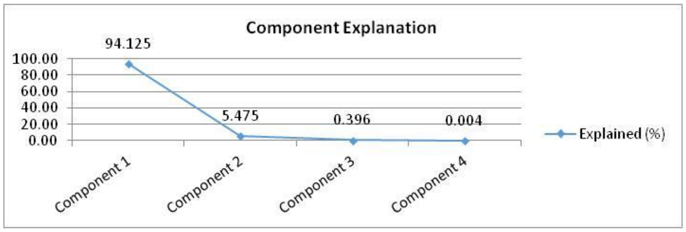


Fig. 9. Distribution of the Percentage of Explanation for the Four Principal Components after the PCA.

TABLE. XII. CORRELATION BETWEEN ALL TEN FEATURES FROM SATELLITE IMAGES AND METEOROLOGICAL STATION

	NDSI	**NDMI**	**NDVI**	**NDWI (Leaves)**	**MNDWI**	**Temperature**	**Relative humidity**	**Visibility**
NDSI	1,0000	-0,9976	-0,5035	-0,9976	-0,6463	-0,1595	0,4019	-0,0570
NDMI	-0,9976	1,0000	0,4551	1,0000	0,6818	0,1573	-0,4033	0,1024
NDVI	-0,5035	0,4551	1,0000	0,4551	-0,3268	-0,3267	-0,0563	-0,0513
NDWI (Leaves)	-0,9976	1,0000	0,4551	1,0000	0,6818	0,1573	-0,4033	0,1024
MNDWI	-0,6463	0,6818	-0,3268	0,6818	1,0000	0,4816	-0,3471	0,0405
Temperature	-0,1595	0,1573	-0,3267	0,1573	0,4816	1,0000	-0,7676	-0,7429
Relative humidity	0,4019	-0,4033	-0,0563	-0,4033	-0,3471	-0,7676	1,0000	0,4133
Visibility	-0,0570	0,1024	-0,0513	0,1024	0,0405	-0,7429	0,4133	1,0000

TABLE. XIII. P-VALUE FROM CORRELATION SHOWED IN TABLE XII

	NDSI	**NDMI**	**NDVI**	**NDWI (Leaves)**	**MNDWI**	**Temperature**	**Relative humidity**	**Visibility**
NDSI	1,0000	0,0001	0,3872	0,0001	0,2387	0,7978	0,5024	0,9274
NDMI	0,0001	1,0000	0,4412	0,0000	0,2049	0,8006	0,5008	0,8698
NDVI	0,3872	0,4412	1,0000	0,4412	0,5915	0,5915	0,9284	0,9347
NDWI (Leaves)	0,0001	0,0000	0,4412	1,0000	0,2049	0,8006	0,5008	0,8698
MNDWI	0,2387	0,2049	0,5915	0,2049	1,0000	0,4115	0,5671	0,9484
NDBI	0,0001	0,0000	0,4412	0,0000	0,2049	0,8006	0,5008	0,8698
UI	0,0003	0,0018	0,3208	0,0018	0,2921	0,7973	0,5077	0,9971
Temperature	0,7978	0,8006	0,5915	0,8006	0,4115	1,0000	0,1297	0,1503
Relative humidity	0,5024	0,5008	0,9284	0,5008	0,5671	0,1297	1,0000	0,4892
Visibility	0,9274	0,8698	0,9347	0,8698	0,9484	0,1503	0,4892	1,0000

TABLE. XIV. RESULT OF THE PCA. THE WORKSPACE IS REDUCED TO FOUR PRINCIPAL COMPONENTS

	Comp. 1	**Comp. 2**	**Comp. 3**	**Comp. 4**
Coefficients	-0,00346636	0,00350483	0,02501616	-0,47097695
	0,00344560	-0,00393762	-0,01981733	0,45440768
	0,00044936	-0,00374080	-0,05152317	0,04637818
	0,00344560	-0,00393762	-0,01981733	0,45440768
	0,00347139	0,00234426	0,03949607	0,60243097
	0,11338987	0,76616650	0,63035961	0,00760066
	0,99352569	-0,08817349	-0,07090754	-0,00778362
	-0,00119390	-0,63651515	0,76941084	0,00395777

TABLE. XV. PERCENTAGE OF EXPLANATION OF THE FOUR PRINCIPAL COMPONENT AFTER THE PCA

	Explained (%)
Component 1	94.125
Component 2	5.475
Component 3	0.396
Component 4	0.004

TABLE. XVI. CORRELATION BETWEEN THE COMPONENT 1 AND THE EPIDEMIOLOGICAL DATA

	Pearson
Correlation	-0.8469

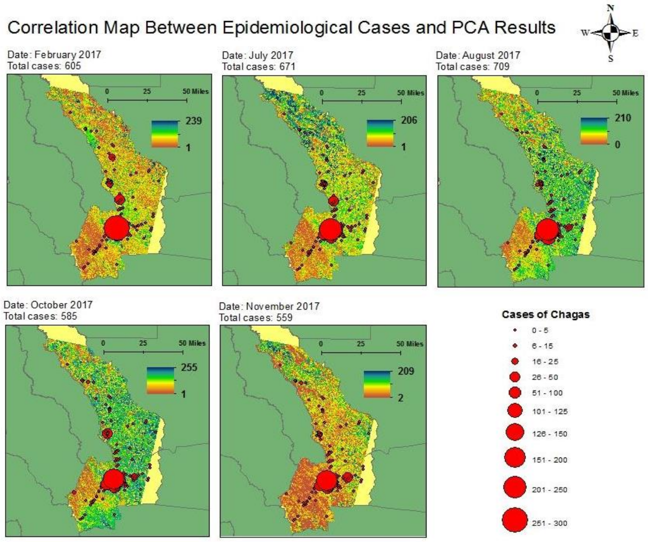


Fig. 10. Correlation Map between Epidemiological Cases and PCA.

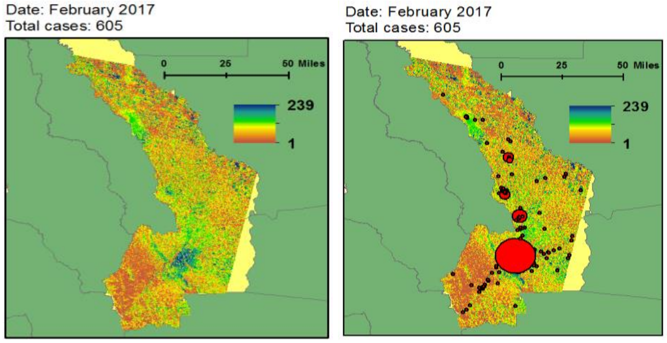


Fig. 11. Comparison between the Correlation Map and the Heat Map.

It can be observed in the five months of the study that the areas with the highest number of Chagas cases also present high and concentrated values of the characteristics of component 1, as can be seen in Fig. 11 that shows the results of February.

V. Conclusions

Five environmental parameters were extracted from the satellite images, this are the NDSI, NDMI, NDVI, NDWI, and MNDWI. These environmental parameters were complemented with weather data as temperature, relative humidity, and visibility extracted from meteorological station. Through the principal components analysis (PCA) of all the environmental variables, it was established that only one component is necessary to explain almost 94% of the data. The parameter that has a strong positive influence on the appearance of Chagas outbreaks according to the first component is the Relative Humidity. The parameter that has a strong positive influence according to the second component is the Temperature. The parameter that has a strong positive influence according to the third component is the Visibility. Finally, the parameter that has a strong positive influence on the appearance of Chagas outbreaks according to the fourth component is the MNDWI.

As a general conclusion of the present study, it has been possible to obtain a comprehensive understanding of the disease based on the use of geospatial technology with 84.69% of correlation. One studied the five months with larger number of cases, between them one establish that the months that present an increase of cases of Chagas in the study area are the months of July and August, this months present environmental conditions that have a strong influence on the occurrence of cases.

As one can see in the study of state of the art, different satellites were used for the environmental analysis related to the propagation of vectors. The contribution of this work is represented mainly in the demonstration of the use of Lansat 8 images for this type of application of epidemiological study specifically Chagas in Bolivia.

As a future work, we want to include data from multiple satellites to reduce the temporal resolution of the satellite used in the present study, besides it is desired to build a geographic information system that contains all the data extracted from the analysis of the satellite images, in addition it is planned to analyze multiple years and finally it is planned to develop an alert platform that can indicate the potential areas that may present Chagas cases according to their environmental conditions.

References

[1] Orellana-Halkyer N, Arriaza-Torres B. Enfermedad de Chagas en poblaciones prehistóricas del norte de Chile. Rev Chil Hist Nat 2010; 83: 531–541.

[2] Chagas CI de la IA de C de la TV y T de la E de. Definición de variables y criterios de riesgos para la caracterización epidemiológica e identificación de áreas prioritarias en el control y vigilancia de la transmisión vectorial de la Enfermedad de Chagas. Uniandes. Fac. de Ciencias, 2004.

[3] O. P. de la Salud, «Enfermedad de Chagas: guía para vigilancia, prevención, control y manejo clínico de la enfermedad de Chagas aguda transmitida por alimentos», Ser. Man. Téc. 12, 2009.

[4] Barbosa M das GV, Ferreira JMBB, Arcanjo ARL, et al. Chagas disease in the State of Amazonas: history, epidemiological evolution, risks of endemicity and future perspectives. Rev Soc Bras Med Trop 2015; 48: 27–33.

[5] C. H. Rotela, «Desarrollo de Modelos e Indicadores Remotos de Riesgo Epidemiológico de Dengue en Argentina.», abr. 2019.

[6] M. Neteler, D. Roiz, D. Rocchini, C. Castellani, y A. Rizzoli, «Terra and Aqua satellites track tiger mosquito invasion: modelling the potential distribution of Aedes albopictus in north-eastern Italy», *Int. J. Health Geogr.*, vol. 10, n.º 1, p. 49, ago. 2011.

[7] Lanfri S, Frutos N, Porcasi X, Rotela C, Peralta G, De Elia E, Lanfri M, Scavuzzo M. -Algoritmos para el Alerta Temprana de Dengue en un Ambiente Geomático. - Instituto de Altos Estudios Espaciales Mario Gulich, Comisión Nacional de Actividades Espaciales. Centro Espacial Teófilo Tabanera, Córdoba, Argentina - (89 - 104).

[8] Díaz ML, González CI. Enfermedad de Chagas agudo: transmisión oral de Trypanosoma Cruzi como una vía de transmisión re-emergente. Rev Univ Ind Santander Salud 2014; 46: 177–188.

[9] J. A. Pérez-Molina y I. Molina, «Chagas disease», Lancet Lond. Engl. vol. 391, n.o 10115, pp. 82-94, 06 2018.

[10] Uribe AG, Bernal GB. Ministro de salud y Protección social. 34.

[11] R. A. Kolliker-Frers, I. Insua, G. Razzitte, y F. Capani, «Chagas disease prevalence in pregnant women: migration and risk of congenital transmission», J. Infect. Dev. Ctries., vol. 10, n.o 09, pp. 895-901, sep. 2016.

[12] F. M. Sano, M. Sano, y O. P. de la Salud, «La enfermedad de Chagas. A la puerta de los 100 años del conocimiento de una endemia ancestral», A doença de Chagas. Perto dos 100 anos do conhecimento de uma endemia americana ancestral, 2007.

[13] C. Forsyth, «From Lemongrass to Ivermectin: Ethnomedical Management of Chagas Disease in Tropical Bolivia», Med. Anthropol., vol. 37, n.o 3, pp. 236-252, 2018.

[14] Ministry of Health Bolivia, "Boletín de vigilancia epidemiológica, Chagas Agudo", 2016.

[15] A. Rassi, A. Rassi, y J. Marcondes de Rezende, «American trypanosomiasis (Chagas disease)», Infect. Dis. Clin. North Am., vol. 26, n.o 2, pp. 275-291, jun. 2012.

[16] Roy DP, Kovalskyy V, Zhang HK, et al. Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. Remote Sens Environ 2016; 185: 57–70.

[17] Chander G, Markham BL, Helder DL. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. Remote Sens Environ 2009; 113: 893–903.

[18] Aguilar-Arias H, Mora-Zamora R, Vargas-Bolaños C. Metodología para la Corrección Atmosférica de Imágenes Aster, Rapideye, Spot 2 y Landsat 8 con el Módulo Flaash del Software ENVI. Rev Geográfica América Cent 2014; 2: 39–59.

[19] Tardy B, Rivalland V, Huc M, et al. A software tool for atmospheric correction and surface temperature estimation of landsat infrared thermal data. Remote Sens; 8. Epub ahead of print 2016. DOI: 10.3390/rs8090696.

[20] Kruse F. Comparison of (ATREM), (ACORN), and (FLAASH) Atmospheric Corrections using Low-Altitude (AVIRIS) Data of Boulder, Colorado. 2004.

[21] Remer LA, Tanré D, Kaufman YJ. Algorithm for remote sensing of tropospheric aerosol from MODIS: Collection 005.

[22] Deng Y, Wu C, Li M, et al. RNDSI: A ratio normalized difference soil index for remote sensing of urban/suburban environments. Int J Appl Earth Obs Geoinformation 2015; 39: 40–48.

[23] Xu H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. Int J Remote Sens 2006; 27: 3025–3033.

[24] Zha Y, Gao J, Ni S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. Int J Remote Sens 2003; 24: 583–594.

A Novel Secure Fingerprint-based Authentication System for Student's Examination System

Abdullah Alshbtat[1], Mohammad Alfraheed[3]
Department of Computer Science and Information
Technology, Faculty of Science
Tafila Technical University, Tafila, Jordan

Nabeel Zanoon[2]
Department of Applied Science
Al-Balqa Applied University
Aqaba, Jordan

***Abstract*—In the fingerprint image processing, various methods have been suggested as using band pass filter, Fouries transform filter and Fuzzy systems. In this paper, we present a useful and an applicable fingerprint security system for student's examination using image processing on such away and a well-organized algorithm is applied. As a university team work, we have recently tested this security procedure for different samples of students in our institution. The experimental results show a high level of accuracy is obtained. Due to the need to connect and manage the connection, we use the Ethernet card and the Arduino Uno card which they are combined together in such a way to do so. Moreover, the administrator runs a special website in the PC to assign ID to the scanned fingerprint. The calculation of the proposed system is carried out by uploading a suitable Adafruit fingerprint library to the used Audruino Uno card. Finally, the most important security point is that the PC has been used not only to send the developing software into the Uno card but also to disconnect the process electronically while the code is running.**

Keywords—Finger-print; examination system; image processing; bio informatics

I. Introduction

The authentication plays an important role in examination systems. The most authentication approach applied in those systems is identity- based authentication. However, the reality in a lot of developed examination system is that the identity-based authentication is not enough to verify the student identification.

One of authentication solutions is the biometric processing based on iris or fingerprint data. Because of cost related issues, fingerprint has been adopted as an automatic data analysis for identification. The main difference between the password-based authentication and fingerprint based authentication is that the first one cannot sometimes protect from unauthorized accessing to users data especially in examination based website applications. Using the fingerprint data does not mean that the examination system avoids using the password-based authentication. In addition to the password-based authentication, the examination system has to be boosted therefore by biometric process such fingerprint data.

Several advantages were found by using fingerprint. The simplicity of using fingerprint as a method to verify and prove the student identities helps allowing in entering the exam. In other hand, the high level of accuracy offered by each single unique human fingerprint helps to prevent any method for identity theft. The unlimited capacity of fingerprint sensors as advantage related to conversion each fingerprint to an image, where latterly it processed in the database. The portability of the fingerprint sensors considered as advantage because of the small device with the tiny components.

The fingerprint-based system offers usually a reliability of such systems with different challenges. Using an external database, however, affects the efficiency of fingerprint-based system. Furthermore, the commercial devices developed by the fingerprint have been constructed to store the fingerprint features for limited users. Those devices have been developed also as an individual unit with their own interface and database, once the user places the fingerprint confirmation message arisen on the output screen without other details (text or password). Another challenge related installation process where the commercial device has to be fixed in a place and connected to the facility networks, where the data transfer to and from the database server. This process increases the cost and chance of hacking. The security web services developed for the examination system represents another challenge. These services could be a promise way to improve the security of examination system once it addressed with biometric processing such as fingerprint.

In this work, a novel secure fingerprint-based authentication system has been developed to discuss these challenges in context of web services. The developed system has been developed to further ensuring for usability, confidentiality and portability of such fingerprint based system.

II. Related Works

Identity authentication considered a topic of interest in the recent years, while increasing the need for more reliable and useful identity authentication systems for security [1]. The old traditional authentication systems depend on using passwords or ID cards found to be less reliable [2]. Indeed, the traditional authentication systems are not able to recognize the original person from the cheater using the password [1].

The case of how to improve the security in conjunction with increase usability and decrease interventions still under work [3], so in order to reduce such negativity related traditional authentication systems biometrics was used.

Biometrics is the physiological and behavioural characteristics that can be measured in the human body and used to confirm the identity and differentiate it from others [4] such as fingerprint. The strength of such biometric methods come from the inability to be stolen or lost as well as difficult to be faked [3].

Managing authentication is a very important activity by fingerprints in order to ensure the integrity, accelerate the process, decrease error rate and fasting verification process, where in a study reviled that time needs for students attendance verification using manual process 23.66 second per student, more than 6.65 second using fingerprint [5].

The most common mechanism for biometrics authentication systems consist of two phases [6]. The enrolment phase used in collecting data and mathematically analyzed it using specific algorithms to perform a data base. The releases phase that interest in comparing and verifying the identity. The general scenario for scanning the figure by using sensor based on capturing two main fingerprint characteristics; the valleys and ridges [7].

Several reasons were considered as advantages for using fingerprint for authentication. The easy to apply is the most important one; the low cost of the uses device as well as don't need much power [6]. Although of the different advantages for fingerprint some disadvantages presented with the complexity in obtaining high quality images of finer patterns related to present of tear, dirty and cuts of finger which will affect the accuracy, time of response and reliability [6].

In order to enhance fingerprint authentication system, different approaches have been deployed in authentication system, where they in other hand have been fused in fingerprint systems for indoor localization [8] by analyzing each indoor algorithms to use the strengths and step down the weakness points to build the best systems.

In other hand, effective identity authentication should be available for wireless networking. So a new robust authentication algorithm based on the phase noise fingerprint of the physical-layer was built [9]. As well as a security authentication scheme of combined physical-layers fingerprints to ensure the survivability of the network from attacks [9].

The security associated with fingerprint-based system is an important issue to be taken inconsideration.

Fingerprint-based system has been addressed in different applications. It was designed for ATM accesses, computer network accesses, class room entering and building door looks [10]. Fingerprint authentication system presented also in mobiles and smart devices where some security insights on touch dynamic provided [3]. New technique by using 3 dimensions magnetic finger motion pattern based implicit authentication to provide highly accuracy was used in smart phones [11], [12].

For attendance checking, the used system built based on fingerprint technology bonded with GPS presented in smart phones to check user availability at anywhere [12]. In very large countries like India, fingerprint-based system was used as a voting system for all population by removing the geographical constrains [13].

Using fingerprint as an access control in information systems depends usually on user awareness and acceptance. User could choose one of two categories of fingerprint-based system on the measurements. Unimodal systems uses only one finger mostly the index (14), which proposed for low/medium security places. Multi modal systems uses two or more fingers for authentication, mostly using the index and middle one, where it consider preferable for medium/high security items [14]. These systems found to be best related to low error rates and high efficiency.

A multi-intance fingerprint based authentication system has been developed which consider more invincible to different problems encountered in previous systems using the crossing number technique [15]. The developed system considered highly efficient in verification the user with highly accuracy and low run time. Also, the system provides the flexibility to switch from multi-intance to unimodal in case of fault tolerance in order to preserve their independency [15].

Indeed, the fingerprint based authentication system developed to possesses the features of highly reliable and easy for secondary development, as well as having several advantages such highly secure, highly accurate, easy in use and being standardized make it applicable in different areas needs authentication such as educational institutions, factories, offices, security and access control systems [16].

III. Design of Proposed System

The fingerprint-based system has been developed in the context of examination system. Two phases have to be carried out; registration and verification.

A. Registration Phase

In this phase, the user (i.e. the student) is requested to add the fingerprint in the database. The fingerprint is previously processed to extract its features in which the fingerprints are uniquely distinguished from each other. A website interface is firstly assigning the user's Identifier UI (i.e. the university student number) to the scanned fingerprint. Once the administrator send on order to scan the fingerprint, a secure connection is established between the front interface and the proposed device of the fingerprint-based system. In addition, a random password-based text is generated and assigned to the user identifier. The proposed system is then activated to start the scanning process.

The user has to follow the prompt massages shown in the LCD screen. The secure established connection which automatically disconnected when the scanning process is activated. Another connection is established between the proposed device and the target record at the database. Both the fingerprint features and the random password are integrated into the user's record. The operation of the registration phase is given as follows:

- The user (i.e. student) U_i is given his identifier. The identifier represents the university student number STD.
- A random password is computed $R_i = Rand(U_i)$.
- U_i imprints his finger print FPion the sensor.
- Compute $ID_i=h(FP_i)$,where h(.) denotes one way hash function which is used to convert the FP_i to identifier.

- 5. Compute Rec_i= U_i+ID_i+R_i
- 6. Reterive Rec_i*from the DB, Rec_i*= (U_i, ID_i, R_i)
- 7. Rec_i* ?= Null
- 8. Store Rec_i at the Database DB, iff Rec_i* = Null

B. Verification Phase

Within this phase, the main target of the proposed system is offered. Two scenarios have been introduced to double check whether the user has been authorized to access the examination system or not. First, the student has to imprint his fingerprint via the proposed system. Then the password shown on the LCD screen; if the user has the access permission, otherwise the proposed system has to display access denied. In the second scenario, the proposed system has been developed to be directly connected with login page of the examination system. Once the student is successfully given the access permission, the exam page is automatically activated for the student. His details are also shown in the exam page. In order to ensure the security issue both scenarios have been developed to open a secure connection once the proposed system is enabled to connect with the database. The database is required to disconnect the connection after the response is sent to the proposed system.

The suggested scenarios are expected to be run. Both of them are initiated by the student who imprints his fingerprint on the sensor. The following operations are therefore carried out:

- Compute ID_i* = h (FP_i).
- Compute C = h (ID_i*+ T) where T is the current timestamp of the login process.
- Retrieve the Rec_i from the database, where Rec_i = (U_i, ID_i, R_i).
- If (ID_i ≠ ID_i*), the system reject the retrieved $Rec_{i.}$
- In first scenario, the system sends a prompt message M=(R_i) to displayed on the LCD screen.
- In second scenario, an activated order is sent to the login page order= (U_i, R_i) for authentication process.
- Compute C'=h(ID_i*+T') where T' is the current timestamp for receiving the activated order.
- If (C'-C) ≥ ∆T, where ∆T is the expected valid time interval for transmission delay, the login phase has to reject the activated order. This kind of timestamp is for double check and to protect the proposed system from attackers.

IV. System Description

The idea of the fingerprint-based examination system has been introduced while the computerized exam was running. Some of students were trying to do the computerized exam instead of his colleagues. Since the password was previously given to student, it is easy to exchange their passwords while the exam's advisor is distracted. Furthermore, the student could sometime impersonate his colleagues.

When it came to designing a fingerprint-based system, some non-functional requirements have been taken in our account. They are Simplicity, Probability, Security and Flexibility. The proposed system has been developed to be directly connected to the administrator's PC. An Ethernet card has been add to the system for managing the network connection. In addition to the fingerprint, LCD screen has been fixed in the proposed system, which is used as an output screen. These entire components are connected to the Arduino card which works as the system brain.

A. Hardware Implementation

As shown in Figure 1, the components of the proposed system have been installed and fixed with each other. These components have been connected into the Arduino Uno card via the appropriate Pins (i.e. Pin 1 to Pin 14). The Ethernet card has been fixed above the Uno Arduino card due to the need to control and manage the connection with the administrator's PC and the need to save more space. Making the proposed system more secure against the attackers, the PC has been used not only to send the developing software into the Uno card but also to disconnect electronically the connections (via Ethernet card) while the code of the proposed system is running. Moreover, the administrator runs special website in the PC to assign ID_i to the scanned fingerprint.

B. Software Implementation

The adafruit fingerprint sensor Library has been downloaded from the RET. The library has been uploaded to the Arduino card which, in turn, controls and manages the operation and calculation of the proposed system. The library has also developed to use the Ethernet card in the network connection. The Arduino card has been also provided by a developed code to show the user's identifier, the user's password and the prompt messages.

The main purpose of the proposed system is to control and manage the attendance procedure of students in the examination class room. Therefore, two hypertext preprocessor (PHP) pages have been developed as shown in Figure 2.

The login page has been developed to provide the proposed system by User's identifier (ID_i), User's name, User's Password (R_i), once the administrator presses the scan button, the page start the scanning process as shown in Figure 3.

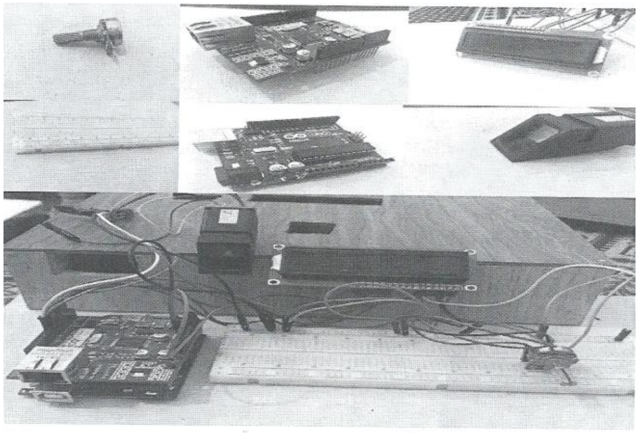


Fig. 1. Components of Fingerprint-Based System.

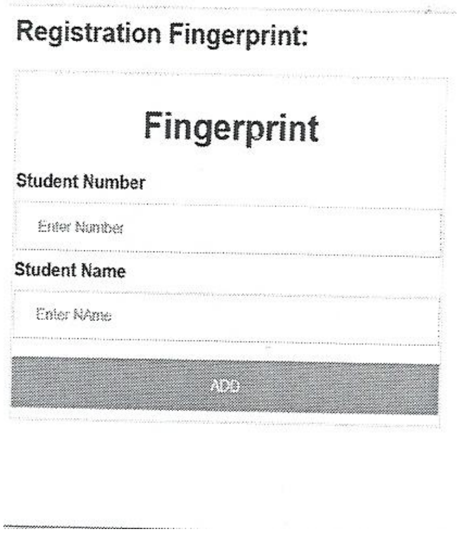


Fig. 2. Registration Page.

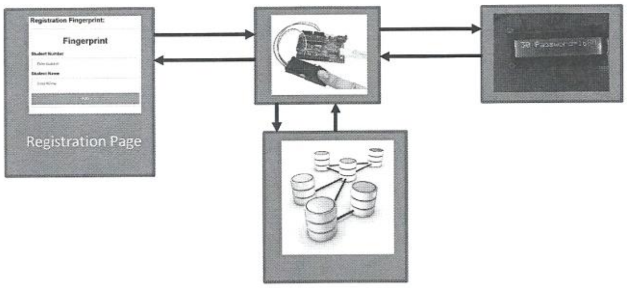


Fig. 3. Scanning Process.

The Login page as shown above has been developed to start the scan process again. However, the Login page sends an order to retrieve the user's identifier and user's password from the database via the proposed system. Therefore, other commands have been developed and uploaded into the Arduino card to communicate with other components of the developed system.

V. Network Security

To ensure a secure network connection, the Ethernet card has been used as a firewall. Moreover, the Ethernet card has been installed to be only connected with the Administrator's PC. Since the proposed system has been introduced as an authentication system for the examination system, the core of the proposed system will be an interested target for attackers. Once the PC is activated to send or receive data from the proposed system, the Ethernet card is enabling to build a secure connection between the proposed system and the PC. After deactivation is the communication is automatically and immediately dissolved. This mechanism aims to separate the proposed system from surrounding and keep it in touch only with the PC. In this way, the proposed system could ensure the network security by accepting the activation order given only by login page or registration page. In other hand, the link of the URL is stored in the SD memory where the server sends a request to retrieve it when needed.

VI. Result and Discussion

Here, the proposed system has been tested and the results has been monitored and collected. First of all, the functionality of the individual components (i.e. device of the proposed system) has been successfully verified while the proposed system was running either in registration and verification phase. Then the proposed system has been tested as a whole for any error. The test has been carried out using the code of Arduino and the initial parameters have been passed via the Ardiuno's interface. Consequently, the proposed system has be successfully constructed and run.

As for fingerprint testing, the proposed system has been run in different conditions. The testing has been carried using 30 users. The proposed system has been connected to the normal (i.e. Core i3). As presented in Table I below, during the testing one user has been asked to place the fingerprint partially on the fingerprint sensor. The proposed system failed to recognize the fingerprint's features and did not successfully read the fingerprint. In addition, the captured image of the fingerprint does not include the main features so the verification failed as (case 2). A female student interested in using beauty cream, therefore, the proposed system failed when the finger was very oily to scanned and verified (case 3). Usually the male students tend to do hard life duties, so in (case 4) the fingerprint was partially disappeared as shown in Figure 4 The proposed system failed detecting the fingerprint as well as the verifying. For the remaining 27 samples which presented with normal case in printing processes, they have been successfully tested by the proposed system.

In general, the proposed System seems to get successfully threading when the Finger is less Dirty, Oily or wet. Using the tested cases (i.e.30 students), the proposed system claims to have more than 99% accuracy rate. As for the failed cases (2,3,4), it can be solved by using another fingerprint from the user's hand. In case all user's fingerprint are failed to be registered in the proposed system, the pass can be traditionally using the user's password.

TABLE I. Results of Proposed System

Number of students	Case 1	Case 2	Case 3	Case 4
30	27	1	1	1

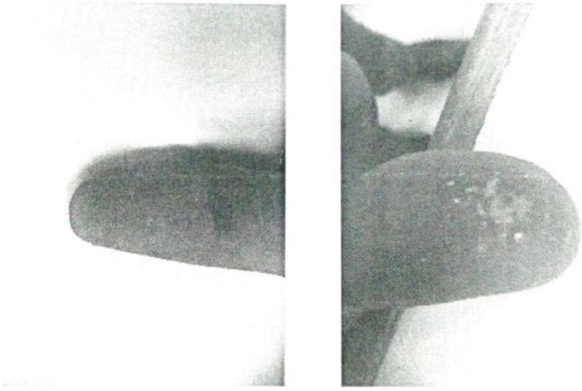


Fig. 4. Burned Finger.

In comparison with other different devices, the proposed system recognized itself from others by different characteristics. Portability which means the capability to move the device from one place to another is applicable in the proposed system so it easy to use indoor and outdoor places not like others used inside doors only [13],[17],[18].

The simplicity of the proposed device affects its ease of development by different institutions where the approximate manufacturing cost around 100 dollars. This simplicity makes it cheaper as well as it is compatible with the android system "open source", on the contrary with other different systems considered expensive and mainly used paid operating system [18], [19].

Connections with external networks done by the control PC (lap top) only which consider the first wall protection that increase the security for the proposed device. In other hand, other devices depend on connecting the fingerprint device to the external networks using network card, which could affect the security and make them high risk to breakthrough [19],[20]. Furthermore the proposed device depends on a temporary communication channels with the device to send and receive biometric data, after that the system disconnect immediately so become harder to hack.

VII. Conclusion

This research work discussed in detail fingerprint pre-processing, minutiae extraction and minutiae matching. This research work has been able to provide a physical security and authentication for students before entering the class room. The experimental result shows efficient registration and verification of subjects with accuracy over 98%.

Acknowledgment

Special thanks are given to our students Mohammad Abu Osbah and Almuthna Alsaide for their efforts in items installation. Moreover, we would thank the president of Tafila Technical University for his financial support.

References

[1] J. Zhou, G. Su, Ch. Jiang, Y. Deng, C. Li (2007). A face and fingerprint identity authentication system based on multi-rout detection. Neurocomputing, 70(1): 922-931.

[2] M. H. Barkadehi, M. Nilashi, O. Ibrahim, A. Z. Fardi, S. Samad (2018). Authentication systems: A literature review and classification. Telematics and Informatics, 35(5): 1491-1511.

[3] P. Sh. Teh, N. Zhang, A. B. J. Teoh, K. Chen (2016). A survey on touch dynamics authentication in mobile devices. Computer and Security, 59(1): 210-235.

[4] M. H. Hammad, A. Mohammed, M. E. Eldow (2015). Design an electronic system use the audio fingerprint to access virtual classroom using artificial neural networks. International conference on computer, communications and control technology (14CT), Kuching,1(1): 192-195.

[5] I. A. Justina (2015). Fingerprint-based authentication system for time and attendance management. British journal of mathematics and computer science, 5(6): 735-747.

[6] I. M. Alsaadi (2015). Physiological biometric authentication systems, advantages, disadvantages and future development: a review. International journal of scientific and technology research, 4(12): 285-289.

[7] A. Suganya, G. M. A. Sagayee (2015). A Delaunay pentangle-based fingerprint authentication system for preserving privacy using topology code. International journal of research in engineering and advanced technology, 2(6): 142-149.

[8] M. Chiputa, L. Xiangyang (2017). Real time Wi-Fi indoor positioning system based on RSSI measurements: A distributed load approach with the fusion of three positioning algorithms. Wireless personal communications: An international journal, 99(1): 67-83.

[9] C. Zhao, M. Huang, L.Huang, X. Du, M. Guizani (2017). A robust authentication scheme based on physical-layer phase noise fingerprint for emerging wireless networks. Computer networks, 128(1): 164-171.

[10] D. Sunehra (2014). Fingerprint based biometric ATM authentication system. International journal of engineering inventions, 3(11): 22-28.

[11] Y. Zhang, M. Yang, Z. Ling, Y. Liu, W. Wu (2018). Finger auth: 3D magnetic finger motion pattern based implicit authentication for mobile devices. Future generation computer system (ISSN 0167-7399).

[12] B. Soewito, F.L.Gaol, E. Simanjuntak, F. E. Gunawan (2015). Attendance system on android smart phone. International conference on control electronics, renewable energy and communications, Bandung, pp: 208-211.

[13] D. Khojare, V. Chaudhary, M. Malviya, Sh. Shukla (2018). FPKIVS-A stellar approach to voting systems in India. Advances in intelligent systems and computing, 653(1).

[14] S. Ribaric and N. Pavesic (2008). A finger based identification system. The 14th IEEE Mediterranean electrotechnical conference, Ajaccio, pp: 816-821.

[15] A. Llugbusi and A. O. Adetunmbi (2017). Development of a multi-intance fingerprint based authentication system. International conference on computing networking and informatics (ICCNI), Lagos, pp: 1-9.

[16] P. Sana, Sh. Prajakta, P. Kamini (2017). Fingerprint based exam hall authentication system using microcontroller. International journal of engineering researches and management studies, 4(2): 89-91.

[17] B. Molina, E. Olivares, C. E. Palau, M. Esteve (2018). Amultimodal fingerprint-based indoor positioning system for airports. IEEE Access, 6(1): 10092-10106.

[18] K. Chow, S. He, J. Tan, G. Chan (2019). Efficient locality classification for indoor fingerprint based systems. IEEE Transactions on mobile computing, 18(2): 290-304.

[19] J. J. Stephan, S. A. Abdullah, R. D. Resan (2017). Use fingerprint technology in developing country security. Annual conference on new trends in information and communications technology applications, Baghdad: 57-62.

[20] J. Baidya, T. Saha, R. Moyashir, R. Palit (2017). Design and implementation of a fingerprint based lock system for shared access. IEEE 7th annual computing and communication workshop and conference, Las Vegas: 1-6.

Ensemble and Deep-Learning Methods for Two-Class and Multi-Attack Anomaly Intrusion Detection: An Empirical Study

Adeyemo Victor Elijah[1], Azween Abdullah[2], NZ JhanJhi[3], Mahadevan Supramaniam[4], Balogun Abdullateef O[5]
School of Computing and IT, Taylor's University, Subang Jaya, Selangor, Malaysia[1, 2, 3]
Research and Innovation Management Centre, SEGI University, Malaysia[4]
Department of Computer Science, University of Ilorin, Ilorin, Kwara State, Nigeria[5]

***Abstract*—Cyber-security, as an emerging field of research, involves the development and management of techniques and technologies for protection of data, information and devices. Protection of network devices from attacks, threats and vulnerabilities both internally and externally had led to the development of ceaseless research into Network Intrusion Detection System (NIDS). Therefore, an empirical study was conducted on the effectiveness of deep learning and ensemble methods in NIDS, thereby contributing to knowledge by developing a NIDS through the implementation of machine and deep-learning algorithms in various forms on recent network datasets that contains more recent attacks types and attackers' behaviours (UNSW-NB15 dataset). This research involves the implementation of a deep-learning algorithm–Long Short-Term Memory (LSTM)–and two ensemble methods (a homogeneous method–using optimised bagged Random-Forest algorithm, and a heterogeneous method–an Averaged Probability method of Voting ensemble). The heterogeneous ensemble was based on four (4) standard classifiers with different computational characteristics (Naïve Bayes, kNN, RIPPER and Decision Tree). The respective model implementations were applied on the UNSW_NB15 datasets in two forms: as a two-classed attack dataset and as a multi-attack dataset. LSTM achieved a detection accuracy rate of 80% on the two-classed attack dataset and 72% detection accuracy rate on the multi-attack dataset. The homogeneous method had an accuracy rate of 98% and 87.4% on the two-class attack dataset and the multi-attack dataset, respectively. Moreover, the heterogeneous model had 97% and 85.23% detection accuracy rate on the two-class attack dataset and the multi-attack dataset, respectively.**

Keywords*—*Cyber-security; intrusion detection system; deep learning; ensemble methods; network attacks

I. Introduction

The proliferation of information and the technology used for enabling communication in everyday life has prompted the immense need for computer security [1]. The impact of Information and Communication Technology on economic growth, social wellbeing, private and public business growth, and national security is enormous as it provides the devices that propagate digital communications among hosts. The overall protection of these hosts, which exist as computers, network devices, network infrastructures, etc. [2], as well as data and information against cyber-attacks, worms, potential leakage and information theft is fundamental to cyber-security [3].

The level of research on the development of Intrusion Detection System (IDS) continues to increase as attacks abound and attackers continue to evolve in practice. As a result, IDSs must evolve to prevail over the dynamic malicious activities carried out over a network. The development of a Network Intrusion Detection System (NIDS) is critical for monitoring the network pattern behaviour of a computer networked system [4]. Typically, an IDS monitors network packets to facilitate the identification of attacks and are basically categorised as either misuse/signature or anomaly based. Signature based IDS matches attacks to previously known attacks, and anomaly-based IDS uses the created normal profile of a user to flag any profile that deviates from the user known behaviour [5].

Because of the unrelenting efforts of attackers to compromise a known network of computers and the new pattern of executing attacks and other malicious activities, the need for a robust, up-to-date IDS is imminent to adequately prevail against unknown attacks/threats or zero-day vulnerabilities.

As such, an empirical research study was conducted to develop an IDS that can address new types of attacks in our modern-day network using machine and deep learning algorithms. The contributions to knowledge produced during this research work are highlighted below:

1) The use of more recent and complex network data as input data, i.e. the UNSW-NB15 dataset, for the development of an IDS.

2) Two (2) methods of implementing ensemble learning methods for the development of an IDS;

3) Implementation of a deep-learning technique (LSTM) for building a NIDS;

4) Development of two (2) categories of NIDS, i.e., two-class (normal and attack labels) and multi-attack (ten class labels).

Moreover, it is the intent of this research work to answer the following research questions:

1) How effective is the ensemble learning method implementation of NIDS for detecting attacks, both in a two-class scenario and a multi-attack scenario?

2) How effective is the deep-learning implementation of NIDS for detecting attacks, both in a two-class scenario and a multi-attack scenario?

3) What peculiarities are found in two-class and multi-attack datasets and how do they affect the developed NIDS models?

II. Related Works

The research conducted by [6] presented a deep-learning method for developing a NIDS. The work proposed and implemented a Self-taught Learning (STL) deep-learning based technique on a NSL-KDD dataset. The STL model when evaluated based on training and test data achieved, in terms of percentage, 88.39% accuracy for 2-class and 79.10% accuracy for 5-class.

The work of [4] is a closely related work, wherein the authors developed a multi-classification NIDS using the UNSW-NB15 dataset and implemented an Online Average One Dependence Estimator and an online Naïve Bayes with 83.47% and 69.60% accuracy, respectively.

Another research work conducted by [7] reported the use of a deep neural network for development of a NIDS. The study implemented LSTM- Recurrent Neural Network (RNN) to identify network behaviour as normal or affected based on the past observations. KDDCup'99 was used as the dataset, and the work achieved a maximum value of 93% efficiency.

The research work carried out by [8] developed four different IDS models using the RNN algorithm and tested them on a NSL-KDD dataset (binary and 5-classes) to evaluate the models. The best model on a binary class achieved 98.1% accuracy using a 1-hidden layer BLSTM. For a 5-class, 87% accuracy was achieved using a 1-hidden layer BLSTM.

Using deep autoencoder (AE) after extracting features via statistical analysis methods, [9] developed an IDS that achieved 87% accuracy on NSL-KDD dataset.

The study of [10] focused on using machine learning methods for developing an IDS using J48, MLP and Bayes Network (BN) algorithms to achieve the overall best accuracy of 93% with J48, 91.9% accuracy using MLP and accuracy of 90.7% using BN on the KDD dataset.

III. Method

A. Dataset

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".

Most research studies on the development of IDS use the KDDCup'99 dataset; however, this dataset is gradually becoming (if not already) obsolete because it does not contain most new forms of attacks prevalent in modern networks of computers. Reflection of contemporary threats and the inclusion of normal network packets are two important features of a high-quality NIDS dataset. Because attackers execute dynamic attacks daily, it is thus necessary to make use of a recent dataset to uncover new malicious activities in a network [11]. Thus, UNSW-NB15 was used in this study. The UNSW-NB15 data was developed using the IXIA PerfectStorm tool in the Cyber Range laboratory of the Australian Centre for Cyber Security, which captured the sets of abnormal and modern-day normal network traffic. More details regarding the dataset creation are given in [2].

Table I provides insights into the datasets used in this study.

As depicted in Table I above, the dataset is comprised of 45 attributes, of which, two (2) are dependent variables. Two subsets of data are obtainable from the original dataset according to the dependent variables; one of these subsets was obtained to develop a two-class anomaly IDS, and the other was use dot develop a multi-attack anomaly IDS. The distribution of the attacks is contained in the attack_cat attribute, and the label attribute is comprised of normal and attack instances, denoted as 0 and 1, respectively.

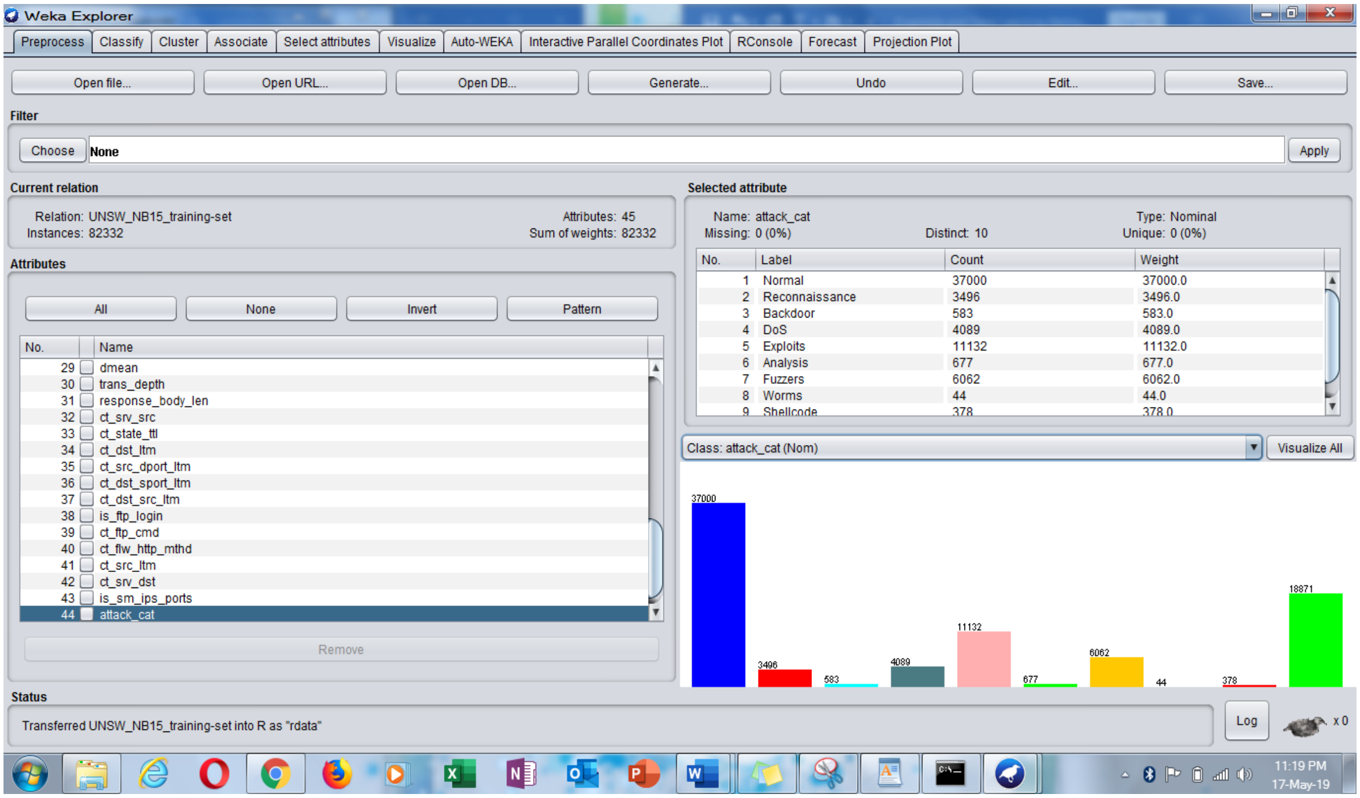
Regarding the features, Table II presents the details of both the independent and target variables.

Moreover, in light of data pre-processing and removal of redundant attributes, the first attribute indexed–id, serving as the index of the dataset, was removed because it is irrelevant, thus leaving two-class and multi-attack datasets with 43 attributes each.

Fig. 1 and Fig. 2 above depict the data distribution for both subsets of the original dataset. Fig. 1 depicts the ten (10) class labels of the multi-attack dataset presented in Table I; each of the labels is displayed using different colour. Fig. 2 shows the two-class labels as presented in Table I above, with blue colour representing the normal labels and red colour representing the attack labels.

TABLE. I. Dataset Description

Dataset Description			
No. of Attributes			45
No. of Independent Variables			43
No. Of Dependent Variables			2
Details of the First Dependent Variable		Name: label	
		Normal	Attack
		37,000	45,332
Details of the Second Dependent Variable		Name: attack_cat	
	Normal		37,000
	Reconnaissance		3, 496
	Backdoors		583
	DoS		4,089
	Exploits		11,132
	Analysis		677
	Fuzzers		6,062
	Worms		44
	Shellcode		378
	Generic		18,871



Filter: None

Current relation

Relation: UNSW_NB15_training-set
Instances: 82332
Attributes: 45
Sum of weights: 82332

Selected attribute

Name: attack_cat
Missing: 0 (0%)
Distinct: 10
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	Normal	37000	37000.0
2	Reconnaissance	3496	3496.0
3	Backdoor	583	583.0
4	DoS	4089	4089.0
5	Exploits	11132	11132.0
6	Analysis	677	677.0
7	Fuzzers	6062	6062.0
8	Worms	44	44.0
9	Shellcode	378	378.0

Attributes

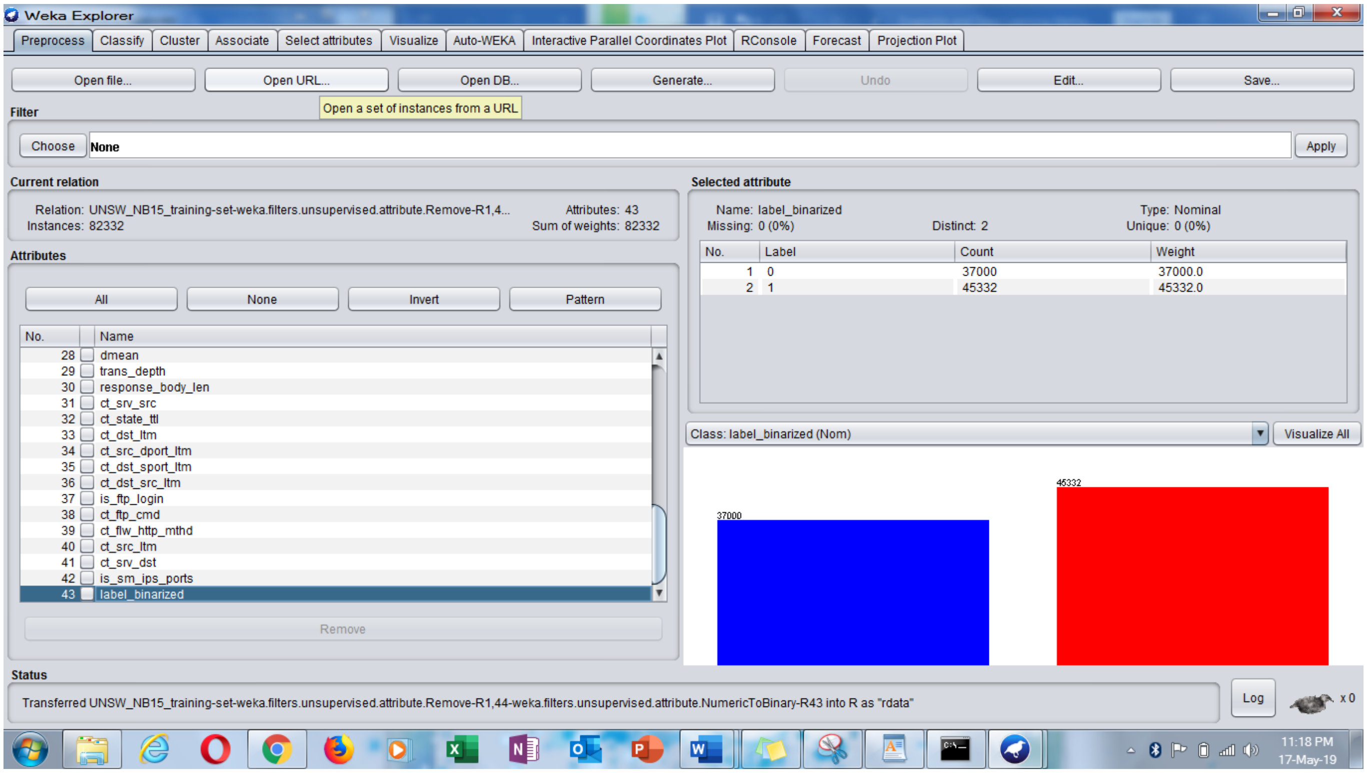
No.	Name
29	□ dmean
30	□ trans_depth
31	□ response_body_len
32	□ ct_srv_src
33	□ ct_state_ttl
34	□ ct_dst_ltm
35	□ ct_src_dport_ltm
36	□ ct_dst_sport_ltm
37	□ ct_dst_src_ltm
38	□ is_ftp_login
39	□ ct_ftp_cmd
40	□ ct_flw_http_mthd
41	□ ct_src_ltm
42	□ ct_srv_dst
43	□ is_sm_ips_ports
44	□ attack_cat

Class: attack_cat (Nom)

Status

Transferred UNSW_NB15_training-set into R as "rdata"

Fig. 1. Data Distribution in the Multi-Attack Dataset.



Filter: None

Current relation

Relation: UNSW_NB15_training-set-weka.filters.unsupervised.attribute.Remove-R1,4...
Instances: 82332
Attributes: 43
Sum of weights: 82332

Selected attribute

Name: label_binarized
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	0	37000	37000.0
2	1	45332	45332.0

Attributes

No.	Name
28	□ dmean
29	□ trans_depth
30	□ response_body_len
31	□ ct_srv_src
32	□ ct_state_ttl
33	□ ct_dst_ltm
34	□ ct_src_dport_ltm
35	□ ct_dst_sport_ltm
36	□ ct_dst_src_ltm
37	□ is_ftp_login
38	□ ct_ftp_cmd
39	□ ct_flw_http_mthd
40	□ ct_src_ltm
41	□ ct_srv_dst
42	□ is_sm_ips_ports
43	□ label_binarized

Class: label_binarized (Nom)

Status

Transferred UNSW_NB15_training-set-weka.filters.unsupervised.attribute.Remove-R1,44-weka.filters.unsupervised.attribute.NumericToBinary-R43 into R as "rdata"

Fig. 2. Data Distribution in the Two-Class Dataset.

TABLE. II. UNSW-NB15 ATTRIBUTES

No.	Features	No.	Features
1	id	23	dtrcpb
2	dur	24	dwin
3	Proto	25	tcprtt
4	Service	26	synack
5	State	27	ackdat
6	spkts	28	smean
7	dpkts	29	dmean
8	sbytes	30	trans_depth
9	dbytes	31	response_body_len
10	rate	32	ct_srv_src
11	sttl	33	ct_state_ttl
12	dttl	34	ct_dst_ltm
13	sload	35	ct_src_dport_ltm
14	dload	36	ct_dst_sport_ltm
15	sloss	37	ct_dst_src_ltm
16	dloss	38	is_ftp_login
17	sinpkt	39	ct_ftp_cmd
18	dinpkt	40	ct_flw_http_mthd
19	sjit	41	ct_src_ltm
20	djit	42	ct_srv_dst
21	swin	43	is_sm_ips_ports
22	stcpb	44	attack_cat
		45	label

B. *Implemented Models*

This empirical analysis implements three (3) different data mining methods to develop a robust NIDS using both datasets mentioned above. The approaches include: (i) Homogeneous Ensemble, (ii) Heterogeneous ensemble, and (iii) Deep Learning (DL) implementations.

An ensemble method [12] is the process of combining some different results, produced by contributing base learners, of predictive models via different combination methods to make a final prediction based on aggregated learning. This method is typically implemented via two phases: the first phase being the construction of various models, and the second phase involving the combination of the estimates obtained from the various models [13]. The ensemble method is said to be homogeneous when the contributing base learners are multiples of the same computational characteristics (family). Base learners in an ensemble model are standard classifiers. In this study, the homogeneous ensemble was implemented in the form of the Random-Forest (RF) algorithm. The Random-Forest algorithm is a bagging method that consists of a finite number of decision tree algorithms with the addition of a 'perturbation' of the classifier used for fitting the base learners. In particular, RF uses 'subset splitting'. The RF ensemble of trees makes use of only a random subset of the variables while building its trees; thus, the ensemble method is homogeneous.

Alternatively, a heterogeneous ensemble is the combination of various results of base learners that have different learning methods or computational characteristics, that is, the contributing base learners belong to different categories of classification algorithms. The standard classifiers for the heterogeneous ensemble considered in this study are described as follows: Bayes Theory (Naïve Bayes algorithm), Instance Learning (k Nearest Neighbour), Rule-based (RIPPER) and Tree methods (C4.5 Decision Tree). The voting combination method [14] [15] was adopted in this study for building the heterogeneous ensemble method. The voting method is a non-complicated method of combining several predictions of varied or different models, and it can be implemented in a variety of approaches, including majority vote, minority vote and average of probabilities. The average of probabilities method of voting [16] was selected for combining the results of each standard classifier because the averaged results of the models are used to provide the final prediction.

DL is an advanced implementation of a neural network. A neural network is the simulation of the human brain, that is, a model of connected neurons. A neural network is usually constructed to possess input, processing and output layers of neurons [17]. The processing layer, often referred to as the hidden layer, may contain one or more layers–a basic implementation of neural network is the Multilayer Perceptron (MLP) [18]. DL is an advancement on the MLP [19], but with more sophisticated and densely connected neurons that are capable of representing and extracting data in a more advanced form from data and mapping it into the output [20, 21]. The neural network implementations that are used for DL include but not limited to Convolutional Neural Network, RNN and Long Short-Term Memory (LSTM). In this study, the deep-learning method implemented was LSTM–a type of RNN. A typical LSTM [7] consists of a cell, an input, an output, and a forget gate, with which it captures the order dependence and recollection of values over random time intervals.

Using the three (3) different data mining methods discussed above, several predictive models were developed using the afore-mentioned datasets. Because it is known that model development is the next stage after the dataset and algorithm selection process and method identification phases, the percentage split model development process was used in this research work. The percentage split is the method of dividing a given dataset into two: the first part is used for executing a training phase-wherein the algorithms builds or fits their respective models, and the second part of the dataset is then used for testing–the phase whereby the fitted models are tested by making predictions using the independent variables of the disjoint test set. Thus, a certain percentage value is given to split the dataset into the training split and the test split. Moreover, having two datasets (two-class and multi-attack datasets), each selected algorithm was fitted on each dataset type, and the resulting models were tested on each corresponding test sets, thereby producing some sets of models that are categorised as (i) two-class attack anomaly IDS, and (ii) multi-attack anomaly IDS, each having three (3) separate models with respect to the applied method discussed above.

To summarise how the data mining methods were implemented and all robust NIDS models were all developed in this study, the proposed empirical framework is depicted in Fig. 3, and the experimental results produced are presented in table and charts and extensively discussed as seen in sections below.

C. Performance Evalutaion Metrics

Following the model development process stage, the developed models are evaluated. As such, the performances of models were evaluated based on the category they belonged to. The two-class anomaly IDS models were evaluated using the following metrics [17]: Detection rate, Area Under Curve (AUC), True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). The multi-attack anomaly IDS models were evaluated based on the following metrics [18]: Detection rate, Kappa value and Weighted (AUC, TP, FP and F-measure). The multi-attack models were evaluated using weighted values because of the multiple values of the class labels (ten in number), unlike the two-class anomaly IDS, which has just two classes (normal and attack)–a binary classification model.

The proposed empirical framework presented in Fig. 3 above consists of the Data Pre-Processing and Re-Labelling Module and the Method Module, which interacts with the Model Development Process Module in producing the two forms of IDS mentioned in this study. The Algorithm module consists of the selected algorithms for this study, and this module interacts with the Method module, which defined the data mining implementations. Last, the Metrics component evaluates the produced model based on its form, and the evaluation results are subsequently discussed.

Table III presents the parameter settings for each algorithm used in this study. All models were trained and tested using the percentage split strategy–80% was used for training and 20% was used for testing, and their performances were evaluated using various metrics as appropriate for the type of the developed IDS model.

Conclusively, all experiments were carried using Waikato Environment for Knowledge Analysis (WEKA) tool for data analysis, wherein results were all obtained and presented in relevant section of this paper.

TABLE. III. IMPLEMENTATION OF EACH ALGORITHMS

Algorithm	Parameter Settings
NB	useKernelEstimator = True; useSupervisedDiscretisation = False, batchsize = 100
kNN	windowSize = 0, batchsize = 100
RIPPER	usePruning = True, seed = 1, batchsize = 100; folds = 5, minNo = 2.0; optimisations = 2, checkErrorRate = True
C4.5 Decision Tree	batchsize = 100, binarySplits = False collapseTree = True; confidenceFactor = 0.25; minNumObj = 2; numFolds = 5; subtreeRaising = True, unpruned = False; seed = 1. .useLaplace = False; useMDLcorrection = True
RF	bagSizePercent = 100; batchSize = 100; breakTiesRandomly = False; maxDepth = 0; computeAtrributeImportance = False; numFeatures = 30; numIterations = 20; seed = 1
LSTM (two-class)	reluAlpha = 0.01, Updater = adam, OptimizationAlgorithm = SGD, learning rate = 0.1, dataset= standardise. While developing the two-class anomaly IDS, LSTM layer was configured as neurons = 128, activation function = ReLU, gate activation function= Sigmoid, dropout = 0.3; Output layer parameter was lossFunction = LossMCXENT, activation function = softmax
LSTM (Multi-attack)	activation function = softmax; gate activation function = ReLU

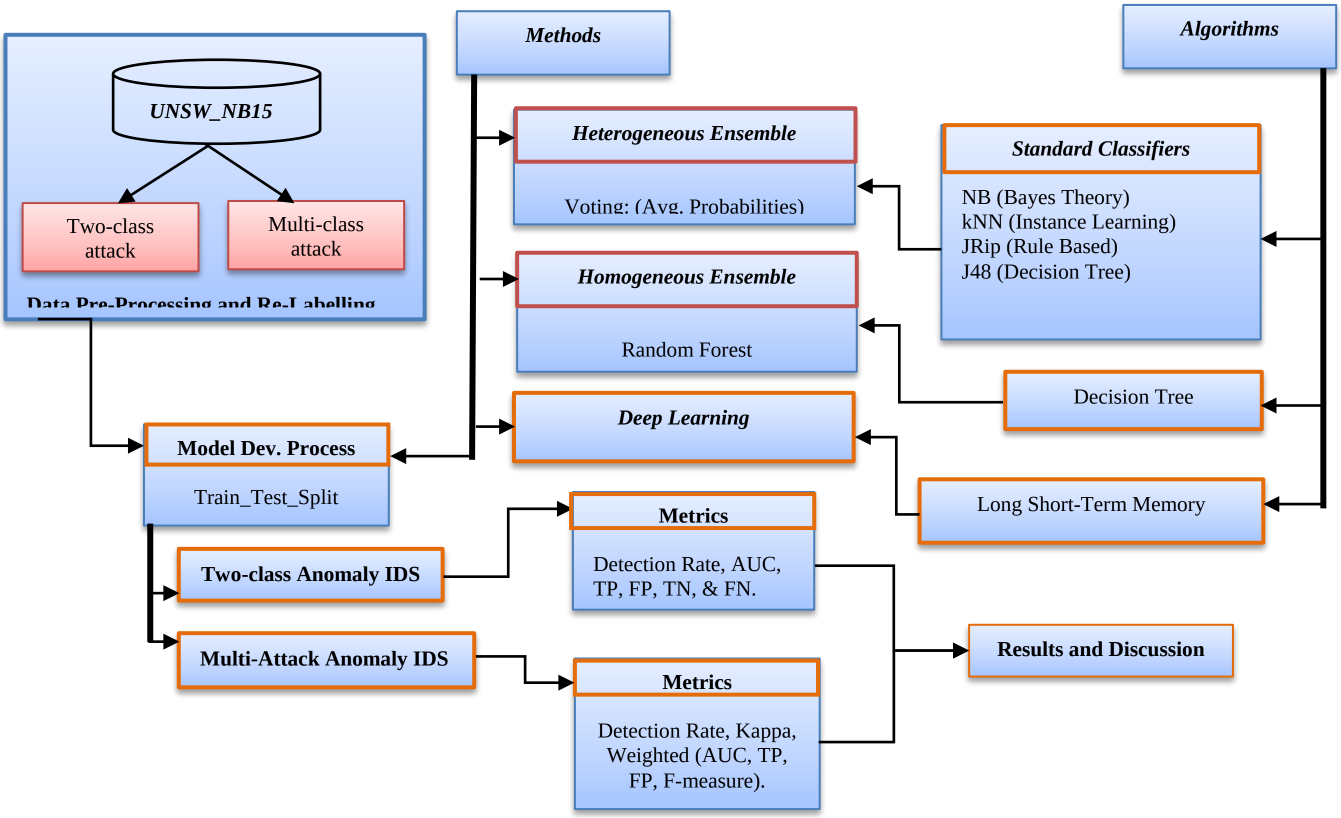


Fig. 3. Proposed Empirical Framework

IV. Results

Having implemented the proposed framework of this research, the reported results will be categorised into two according to the model development processes. Note that the test was conducted on 20% of the dataset, resulting in 16,466 instances. First, the two-class anomaly IDS is basically the prediction of whether a network packet is normal or an attack and is thus evaluated using the given metrics in Fig. 3. For the homogeneous method, Tables IV and V present the performance scores of the model and its corresponding confusion matrix, respectively.

From Table IV, the homogeneous ensemble had an overall detection rate of 97.96% with an AUC score of 0.997, indicating a very strong prediction model. The TP value of 0.98 indicates that the model classified 98% of normal packets as normal, and the TN value of 0.976 denotes that the attack packets were correctly classed as attack at the rate of 97.6%. The FP value of 0.024 denotes that just 2% of normal packets were classified as attack, and the FN value of 0.0158 indicates that approximately 1.58% of attack packets were predicted as normal. Likewise, Table V–the confusion matrix of the homogeneous ensemble, depicts the actual figures of the TP–7278 of 7395 normal instances classified as normal, FP–219 of 9071 attack instances misclassified as normal, TN–8852 of 9071 attack instances correctly classified as attack, and FN–117 of normal instances misclassified as attack.

For the heterogeneous ensemble, the voting cum average probabilities results for different techniques are shown in Tables VI and VII below.

TABLE. IV. Homogeneous Ensemble Model's Evaluation

Evaluation Metric	**Score**
AUC	0.997
TP Rate	0.984178
FP Rate	0.024143
TN Rate	0.975857
FN Rate	0.015822
Detection rate	0.979594

TABLE. V. Homogeneous Model Confusion Matrix

	Normal	***Attack***
Normal	7278	117
Attack	219	8852

TABLE. VI. Heterogeneous Ensemble Model's Evaluation

Evaluation Metric	**Score**
AUC	0.994
TP Rate	0.984043272
FP Rate	0.042994157
TN Rate	0.957005843
FN Rate	0.015956728
Detection rate	0.969148549

From Table VI, the heterogeneous ensemble had an overall detection rate of 96.92% with an AUC score of 0.994, indicating yet another very strong prediction model. The TP value of 0.98 indicates that the model classified 98% of normal packets as normal, and the TN value of 0.957 denotes that the attack packets were correctly classed as attack at the rate of 95.7%. The FP value of 0.43 denotes that approximately 5% of normal packets were classified as attack, and the FN value of 0.016 indicates that approximately 1.6% of attack packets were predicted as normal. Likewise, Table VII–the confusion matrix of the heterogeneous ensemble, depicts the actual figures of the TP–7277 of 7395 normal instances classified as normal, FP–390 of 9071 attack instances were misclassified as normal instances, TN–8681 of 9071 attack instances correctly classified as attack and FN–118 of 7395 normal instances misclassified as attack.

Last in this category, the results of deep-learning method for developing a two-class anomaly IDS as implemented with the specified parameters described in the previous section are shown in Tables VIII and IX.

Table VIII shows that the deep leaning model had an overall detection rate of 80.72% with an AUC score of 0.926, i.e. the deep-learning model is a competitive predictive model. The TP value of 0.57 indicates that the model classified 57% of normal packets as normal–a fair result as compared to other models in this category; it has a strong TN value of 0.998, indicating that the attack packets were correctly classed as attack at the rate of 99.8%-the best TN value in this category. The model had a FP value of 0.002, denoting an insignificant number of misclassified normal instances, and the FN value of 0.426 indicates that approximately 42.6% of attack packets were predicted as normal. Likewise, Table IX – the confusion matrix of the heterogeneous ensemble, depicts the actual figures of the TP–4239 of 7395 normal instances classified as normal, FP–19 of 9071 attack instances misclassified as normal, TN–9052 of 9071 attack instances correctly classified as attack and FN–3156 of normal instances misclassified as attack.

TABLE. VII. Heterogeneous Model Confusion Matrix

	Normal	**Attack**
Normal	7277	118
Attack	390	8681

TABLE. VIII. Deep-Learning Model's evaluation

Evaluation Metric	**Score**
AUC	0.926
TP Rate	0.573225
FP Rate	0.002095
TN Rate	0.997905
FN Rate	0.426775
Detection rate	0.807178

TABLE. IX. Deep-Learning Confusion Matrix

	Normal	**Attack**
Normal	4239	3156
Attack	19	9052

Critical evaluation of the models in this category reveals that, despite all models performing well using the AUC metric, the deep-learning model is weak in the detection of normal packets and will generate more false flagging of normal packets, thereby degrading the network monitoring in real time. Moreover, although the homogeneous and heterogeneous models competed fairly with each other, as they are both robust models for the detection of normal and attack packets, the homogeneous ensemble model is the best model in terms of lower FP and higher AUC values.

The second category is the multi-attack anomaly IDS, which is the classification of packets into normal and nine different types of attacks–a typical multi-classification problem, as discussed in previous section. The models are evaluated as depicted in Fig. 3. For the homogeneous ensemble method in this category, Table X reveals various performances scores.

Table X reveals the model's ability to detect whether a packet belongs to any of the ten (10) classes at 87.39%. This model had a kappa value of 0.8 and a weighted AUC of 0.98. The weighted TP value is 87.4%, and the weighted FP value is 2.5%. The model also had a weighted F-measure value of 0.87.

Similarly, in Table XI, this model detection rate was 85.23% but with a weighted AUC of 0.98, a weighted TP value of 0.852–85.2% correct classification of each class label instances, a weighted FP value of 0.031, a weighted F-measure of 0.855, and a kappa value of 0.79.

Last in this category, the deep-learning model of multi-attack anomaly IDS was also evaluated; its scores are represented in Table XII.

The deep-learning model yielded an ability to detect and predict the class of any packet at 72%. This result is achieved at a weighted AUC value of 0.868, a weighted F-measure score of 0.659, and a kappa value of 0.57. This model is capable of correctly detecting each class instance at the weighted TP value of 72.3, and it had a weighted FP value of 0.17.

TABLE. X. HOMOGENEOUS MODEL'S EVALUATION

Evaluation Metric	Score
Weighted AUC	0.98
Weighted TP Rate	0.874
Weighted FP Rate	0.025
Weighted F-Measure	0.87
Kappa Statistics	0.8227
Detection rate	87.3861

TABLE. XI. HETEROGENEOUS MODEL'S EVALUATION

Evaluation Metric	Score
Weighted AUC	0.982
Weighted TP Rate	0.852
Weighted FP Rate	0.031
Weighted F-Measure	0.855
Kappa Statistics	0.7928
Detection rate	85.2302

TABLE. XII. DEEP-LEARNING MODEL'S EVALUATION OF MULTI-ATTACK ANOMALY IDS

Evaluation Metric	Score
Weighted AUC	0.868
Weighted TP Rate	0.723
Weighted FP Rate	0.171
Weighted F-Measure	0.659
Kappa Statistics	0.57
Detection rate	72.26

In this multi-attack category, the homogeneous ensemble method also achieved the best performance, with a weighted F-measure of 0.87, a kappa value of 0.82, and an overall detection rate of 87%. Although the heterogeneous had a weighted AUC of 0.982, it is the second best in this category. Last, the deep-learning model competed fairly well with the other models, with its weighted AUC of 0.868; however, it had a low kappa value of 0.57 and a low detection rate of 72.26. Moreover, the confusion matrix for each model reveals the classification and misclassification of the instances accordingly. The deep-learning model was found to be unable to detect many attack classes, whereas the homogeneous model was adequately robust.

A summary of the detection rate of all models for both categories is presented in Table XIII.

Table XIII concisely presents the detection rates for all the above-described models, as is pictorially depicted in Fig. 4.

TABLE. XIII. SUMMARY OF THE RESULTS

Models	Methods	Detection Rate (%)
Two-class Anomaly IDS	Homogeneous Ensemble	97.96
	Heterogeneous Ensemble	96.92
	Deep Learning	80.72
Multi-attack Anomaly IDS	Homogeneous Ensemble	87.39
	Heterogeneous Ensemble	85.23
	Deep Learning	72.26

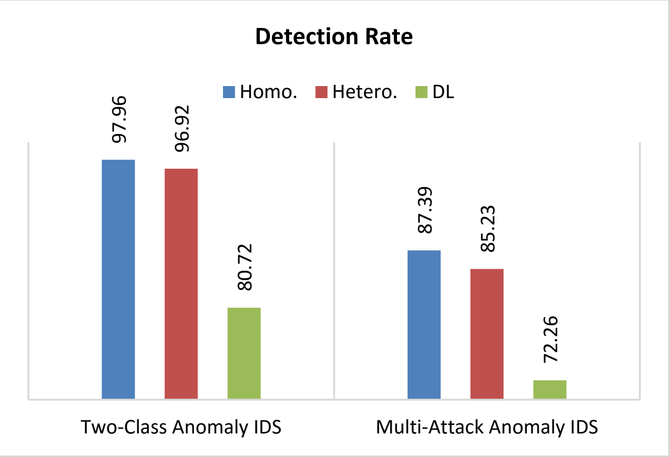


Fig. 4. Pictorial Representation of the Detection Rates of the Models.

V. Discussion

Based on the implementation of the various methods of machine learning and deep-learning algorithms in the development of several IDS models and making use of a modern-day dataset, it is possible to generalise the results. First, this study supports the fact that machine learning and DL are competent and effective technique for developing IDS in various capacities, such as two-class and multi-attack anomaly detection. This work also revealed that simple implementation of a machine learning algorithm is required and is a much more effective with less computational cost and complexity in the development of IDS regarding the strong predictive prowess of the homogeneous ensemble method compared to the heterogeneous (though it fiercely competed) and the deep-learning methods. Moreover, it can be generally stated that the detection rate for a two-class IDS is higher than that of multi-attack IDS because of the number of classes the machine learning will learn to make correct predictions and also the nature of data, wherein imbalance is peculiar to the multi-attack dataset, whereas the two-class dataset is mostly balanced.

Generally, the best model produced by this research work for detecting either a normal or attack packet (two-class anomaly IDS) operates at the rate of 97.96% and the best model for the multi-attack (ten-classes) anomaly IDS has the detection rate of 87.39%. In direct comparison with the recent work of [4], which actually outperformed much past research models, their work produced an overall detection rate of 83.47% for their online AODE model and a 69.60% detection rate for their online Naïve Bayes model, both of which were outperformed by the best (87.39%) and second-best (85.23%) detection rate of the multi-attack anomaly IDS models developed in this research work.

Comparatively, the research work conducted by [6] produced NIDS of 88.39% accuracy for a two class attack which was outperformed by two of the three NIDS of this study developed for 2-class attack detection (with 97.96% and 96.92% detection rate produced in this study), and also while their work produced a 79.10% accuracy for 5-class, the NIDS developed in this study produced two out of the three NIDS with 87.39% and 85.23% detection rate for 10-classes. Also, their STL model yielded a 75.76% f-measure value for the 5-class NIDS while this study produced 87% for homogeneous ensemble and 85% for heterogenous ensemble for a 10-class NIDS.

Additionally, the study of [10] developed IDS using J48, MLP and Bayes Network (BN) algorithms to achieve the overall best accuracy of 93% with J48, 91.9% accuracy using MLP and accuracy of 90.7% using BN on the KDD dataset. The homogenous ensemble NIDS developed in this study outperformed their work with a detection rate of 97.96% as well as the heterogenous ensemble NIDS with the detection rate of 96.92%

Having implemented several machine learning and deep-learning algorithms and several techniques for combining models, the application of feature selection technique to best select features from the available ones in the dataset is recommended as future work and also in practice to produce an optimal model with less cost and computational complexity. Moreover, the deep-learning method requires further investigation because there is need for improvement in both two-class and multi-attack anomaly IDSs.

VI. Conclusion

This research work revealed answers to several research questions. In response to the first question, the NIDS developed using machine learning is highly effective with a homogeneous ensemble implementation achieving a detection rate as high as 98% in a two-class scenario and 87% in a multi-attack scenario, and its heterogeneous counterpart is effective for NIDS with a detection rate of 97% in a two-class scenario and 85% in a multi-attack scenario.

In response to the second research question, the empirical research work revealed that a deep-learning implementation can be effective at as low as 80% detection rate in a two-class scenario and can effectively detect various types of attacks and normal packets in a multi-attack scenario at 72%.

Answering the third research question, it was discovered that two-class datasets have a balanced distribution unlike the multi-attack which is greatly imbalanced. These peculiarities affected the developed models as the developed models better fitted the balanced dataset than the imbalanced dataset.

The results of this research work also revealed that it is easier to identify two classes of network packet than ten (10) different classes belonging to a network packet.

This research work also revealed the weakness of DL, as it cannot produce a competitive model if its configuration is not sophisticated, i.e., is comprised of a high number of layers, which in turn increases computational complexity and cost.

A dataset consisting of 43 attributes is usually considered as a high-dimensional dataset that requires a feature pre-processing stage, wherein redundant, irrelevant and (in some cases) highly correlated attributes are removed to develop a robust model that neither over fits or under fits the dataset. This stage is executed by applying a feature selection technique, which includes filter and wrapper methods; however, this stage was not conducted in this research work and will be considered in future work. Additionally, while developing NIDS using two-class dataset, it was discovered that the dataset was imbalance. Thus, class balancing is also considered as future work.

The development and deployment of the developed NIDS models for real time detection of attack is considered as an important future work.

References

[1] M. Ahmed, A. N. Mahmood, and J. Hu, “A survey of network anomaly detection techniques,” J. Netw. Comput. Appl., pp. 1–13, 2015.

[2] N. Moustafa and J. Slay, “UNSW-NB15: A Comprehensive Data set for Network Intrusion Detection systems,” Proc. ofthe Mil. Commun. Inf. Syst. Conf., pp. 1–6, 2015.

[3] A. O. Balogun, A. M. Balogun, V. E. Adeyemo, and P. O. Sadiku, “A Network Intrusion Detection System : Enhanced Classification via Clustering,” Comput. Inf. Syst. Dev. Informatics Allied Res. J., vol. 6, no. 4, pp. 53–58, 2015.

[4] M. Nawir, A. Amir, N. Yaakob, and O. N. G. B. I. Lynn, "Multi-Classification of Unsw-Nb15 Dataset for," vol. 96, no. 15, pp. 5094–5104, 2018.

[5] S. M. Thaler, Automation for information security using machine learning. 2019.

[6] Q. Niyaz, W. Sun, A. Y. Javaid, and M. Alam, "A Deep Learning Approach for Network Intrusion Detection System," 2016.

[7] M. Ponkarthika and V. R. Saraswathy, "Network Intrusion Detection Using Deep Neural Networks," vol. 2, no. 2, pp. 665–673, 2018.

[8] A. Elsherif, "Automatic Intrusion Detection System Using Deep Recurrent Neural Network Paradigm," 2018.

[9] C. Ieracitano, A. Adeel, M. Gogate, K. Dashtipour, and C. R. Aug, "Statistical Analysis Driven Optimized Deep," Int. Conf. Brain Inspired Cogn. Syst. Springer, Cham., pp. 759–769, 2018.

[10] M. Alkasassbeh and M. Almseidin, "Machine Learning Methods for Network Intrusion Detection," no. October, 2018.

[11] G. Li and Z. Yan, "Data Fusion for Network Intrusion Detection : A Review," vol. 2018, 2018.

[12] G. Seni and J. F. Elder, Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. 2010.

[13] P. Illy, G. Kaddoum, C. M. Moreira, K. Kaur, and S. Garg, "Securing Fog-to-Things Environment Using Intrusion Detection System Based On Ensemble Learning," no. April, pp. 15–18, 2019.

[14] M. Sabzevari and G. Mart, "Vote-boosting ensembles," Pattern Recognit., 2018.

[15] D. Murphree et al., "Ensemble Learning Approaches to Predicting Complications of Blood Transfusion," in IEEE Eng Med Biol Soc., 2016, pp. 1–11.

[16] I. H. Witten, E. Frank, and M. A. Hall, Data Mining - Practical Machine Learning Tools and Techniques. 2011.

[17] Y. Xin et al., "Machine Learning and Deep Learning Methods for Cybersecurity," IEEE Access, vol. 6, pp. 35365–35381, 2018.

[18] M. A. Mabayoje, A. O. Balogun, A. O. Ameen, and V. E. Adeyemo, "Influence of Feature Selection on Multi - Layer Perceptron Classifier for Intrusion Detection System," Comput. Inf. Syst. Dev. Informatics Allied Res. J., vol. 7, no. 4, pp. 87–94, 2016.

[19] F. Feng, X. Liu, B. Yong, R. Zhou, and Q. Zhou, "Anomaly detection in ad-hoc networks based on deep learning model: A plug and play device," Ad Hoc Networks, 2018.

[20] D. Papamartzivanos and G. Kambourakis, "Introducing Deep Learning Self-Adaptive Misuse Network Intrusion Detection Systems," IEEE Access, vol. 7, 2019.

[21] SH Kok, Azween Abdullah, NZ Jhanjhi, Mahadevan Supramaniam, "A Review of Intrusion Detection System Using Machine Learning Approach", in International Journal of Engineering and Research, Jan 2019.

Timed-Arc Petri-Nets based Agent Communication for Real-Time Multi-Agent Systems

Awais Qasim[1], Sidra Kanwal[2], Adnan Khalid[3], Syed Asad Raza Kazmi[4], Jawad Hassan[5]
Department of Computer Science, Government College University, Lahore Pakistan

***Abstract*—This research focuses on Timed-Arc Petri-nets-based agent communication in real-time multi-agent systems. The Agent Communication Language is a standard language for the agents to communicate. The objective is to combine Timed-Arc Petri-nets and FIPA Performatives in real-time multi-agent systems. FIPA standards provide a richer framework for the interaction of agents and makes it easier to develop a well-defined system. It also ensures the management by precisely specifying the agent's interaction. Though FIPA protocol has already been described with the help of Petri-nets but this specification lacks the timing aspect that is a dire need for real-time multi-agent systems. The main objective of this research is to provide a method of modeling existing FIPA performatives by combining Timed-Arc Petri-nets in real-time multi-agent systems. We have used properties, such as liveness, deadlock and reachability for the formal verification of the proposed modeling technique.**

***Keywords*—Formal verification; FIPA; multi-agent systems; timed-arc petri nets; real-time systems**

I. Introduction

Modeling of agents for the transmission of messages is much needed as agents interact with one another to achieve goals. Cooperation in Multi-Agent Systems (MAS) is mainly achieved through interacting agents. For effective communication these interacting agents require some interaction protocol. The main purpose of the interaction protocol is to provide a set of well defined rules for the communication of agents. The primary function of an agent is to handle the dynamic situations. There is no agent that possesses information of the whole system rather its decision making is dependent upon limited view of the complete system [1]. Real-time agents can be depicted within the deadline. MAS's have been formally specified using petri-nets but no such work has been done for the formal modeling of agent communication protocols in real-time environment. By using the Real-Time Multi-Agent Systems (RTMAS), the interaction of agents is bounded within the proposed context. Protocols become the cause of message flow among these RTMAS's and specify the sequence of messages, number of messages and updation. Foundation for Intelligent Physical Agent (FIPA) performatives provide outline for the existence of agents, their actions and architecture. They also elaborate the authentication of the agents. Agent Communication Language (ACL) is a proposed standard language for agent communications like FIPA-ACL. FIPA has 20 performatives i-e request, inform, accept proposal, etc that describe the interpretation of messages. These messages are actually the actions (communicative actions or CAs). Interaction protocols are the standards that oversee the interaction between agents. It permits the description of expressly sequence of dialogue between agent's communication. Interaction Protocols are utilized to characterize set of messages transmitted between agents and portray how collaborative agents response on messages. It is normal to make models from basic conventions into complicated protocol. These interaction protocols have been demonstrated in MAS but not presented in RTMAS. FIPA performatives provide an outline for the existence and action of agents. Existing FIPA protocols have been described through Petri-nets but the timing aspect was not specified which was a limitation. The main idea of this research is to formally model the existing FIPA performatives by Timed Arc Petri-Nets (TAPNs) in real-time multi-agent systems. RTMAS is formally determined and checked with the time limitations. In a real time system, there are some actions, which have specified deadlines and depicts how long agents will wait for reply or perform next action by using FIPA performatives.

Agent's communication has been modeled formally in the past but not for real-time environment. The communication between agents is modeled by using Agent Petri-nets (APN) and it is undeniable that integration between protocols and APN greatly facilitates the development of a system which leads to correct interaction between agents through appropriate specification of the exchange of messages. The time aspect has not been handled in [2] before performing any target. The work of [3], [4], [5] leads to formal specification and verification of interactive real-time software agents (RT Agents). Agents work independently and handle the uncertain scenarios. Visually expressive broader structure and modeling approach i.e. TAPN have been used for specification and representation of Stock Market System (SMS). It is based on RTMAS. The Model is verified by Timed Computational Tree Logic (TCTL) fragments AF, AG, EG and EF. In this paper, KQML register conversation and simple negotiation interaction conversation are modeled through CPN. The work done in [6] describes popular ACL like FIPA to formally model the organization of MASs and clarifies the analysis about FIPA ACL semantics. FIPA specification has been used for guidance. ACL specifications have been introduced with the example of online stock brokering to secure the interaction between agents. This new model of ACL has expressiveness and reusability. In [7] nested petri-nets have been used to model multi-agent systems. In [8] the overview of FIPA-ACL and protocols has been given. FIPA-ACL is based on speech act theory as communicative act. With the passage of time, a lot of improvements have been made but still, none of the protocols can be treated as the complete in itself. In [9] an agent-based framework in an unconstrained platform has been described. This research highlights agents and multi-agents system as the state-of-the-art and distributed environments respectively. The work of [10] emphasizes the modeling of the vehicle framework that can go through crossing points with no or less delay. This approach determines how the model emphasizes the traffic movement (transitions) by reachability graph within time constraints. Time-based constraints for MAS are presented in [11]. It provides clear ways to accomplish the MAS compliance. Vehicle to vehicle communication is presented by using Petri-nets. Vehicle to vehicle communication makes an efficient exchange of messages between cars and also matches the ID of cars for verification. It also works for the modeling and stimulation of vehicles to vehicle communication of discrete event system, Petri-net is used as a powerful tool. In [12] a more far-reaching display and re-enactment approach is introduced that records for the MAS-related conventions as portrayed in the FIPA particular; together with a co-reproduction stage for the examination of MAS. The process of checking a smart workflow management framework and postpone forecast is presented [13], [14]. The work of [15] describes the formal specification approach for the presentation of communicative agents. It explores the internal state of agents and behavior of interactive agents. In [16] model checking techniques are discussed, which are utilized as a part of tool TAPAAL for reachability analysis using inhibitor arcs. Reachability is utilized for the dead state. It finds that there is a state which is not reachable from some other state. The research of [17] presents a liveness based analysis for Timed-Arc Petri-Nets with weights and arcs i.e. inhibitor arcs, transport arcs and age invariants. This research highlights agents and MAS as state of the art and distributed environments, respectively. Face detector and tracker agents interact through contract net protocol. This system tracks the agents which path has been chosen and time spent. Events triggered sample data consensus has been proposed in [18], [19].

Even triggering condition is intermittently examined after constant sampling instant for distributed MAS with directed graph. Consensus of distributed MAS can be transformed into stability of system with time. Then a sufficient condition on the consensus of Multi-Agent system is derived. The management of communication between agents is presented in [20]. This model is illustrated with the help of Petri-nets and result is validated with coordination between agents. In [21] demonstrates that response to the presence of discrete-time and continuous time strategies occur at the same time or getting input or wait with decision until next occasion is activated and creates algorithm for discrete Timed-arc Petri-net games. A work process based that concentrates on the foundational issues of soundness is strongly based on timed-constrained soundness. Through subclass of bounded nets, we can efficiently verify the design [22]. For the effective processing, agents search for more agents by using KQML. Contract net protocol is utilized on the interaction of agents. If an agents bounds with one contract, it is illegible to take new one until the completion of previous [23].

According to our knowledge, formal specification of MAS's interaction in a real-time environment is a novel approach. Although formal modeling of MASs have been done in the past but it is limited to domain functionality of the complete system either at the micro or at macro level. In MAS goals can be defined at any single agent level or at the complete system level but any single goal may contain several agent performatives. Formal modeling of agents interaction allows the designer to verify the system's correctness at the design time. In this research, FIPA performatives are formally specified in TAPN. TAPN is a framework for visual representation of sequence of events in time. This is used to describe the modeling approach. Arcs are used to represent the time specification i-e inhibitor arc, transport arc and invariants. In the formal specification of RTMAS, agents interact with one another to achieve their goals within time constraints. This element is required for their correct functioning. Protocols define how much long period of time the agents would wait for the concerned interacting agents and also for updation. TAPAAL model checker is used for formal verification of FIPA performatives in RTMASs. TAPAAL is a graphical representation and verification tool of TAPN. It is also used for the verification of different queries specified to ensure the correct functioning of the system's model. The model is verified with the fragments of TCTL whose fragments are AF, AG, EG and EF.

II. Preliminaries

A few terminologies and computational models have been described in this section that will be used in the rest of the discussion for the specification of the problem under analysis.

A. Timed-Arc Petri-Net (TAPN)

TAPN is an established technique for the formal modeling of multi-agents systems in real-time environments. The timing aspect considers unequivocal treatment of real-time. TAPN provides a mechanism to formally verify the properties of interest of the system to ensure its correctness. The TAPN definition has been proposed in [24]. A TAPN is a 6-tuple (P, T, IA, OA, I, Type) where

P: is finite set of places.

T: is a finite set of transitions.

IA: P*T is finite set of input arcs.

OA: T*P is finite set of output arcs.

I: represents the age interval of places.

Type represents the type of arc normal, inhibitor, transport arc.

TAPNs are basically an extension of the standard petri-nets in which age of the tokens is utilized to incorporate timing aspect of the system. The time intervals defined on arcs are used to restrict the progress of the system by only allowing those tokens whose age falls in the interval. Arc defines a place to transition and transition to place. Transition firing and enabling depend on the age of token. Transition enabling and firing is not possible when the time interval is expired that is mentioned on a certain arc. There are three types of arcs such as transport arc, inhibitor arc and normal arc. Transport arc produces the same time which is consumed on input arc. Inhibitor arc restricts the certain age of token on the place. Normal arc produces the age zero token although consumed of any age token. TAPN is used for the verification of reachability, boundedness and liveness properties.

B. TAPAAL

TAPAAL is the graphical representation and powerful tool for modeling of Timed-Arc Petri-nets. TAPAAL is used for simulation and verification of TAPN. It is graphical, modeling editor for Timed-Arc Petri-net. It provides its own engine for verification [25]. For verification different properties are used such as Reachability, boundedness, and liveness. These properties are defined under the Timed Computational Tree Logic (TCTL). TCTL is the actual logic which is utilized for the determination of properties about structure. The fragments of TCTL are AF, AG, FG, and EF.

C. FIPA Performatives

FIPA developed its standards in 1995 for agent's communication. Speech act theory provides a base for FIPA ACL. Modeling of agents is useful for the transmission of messages. Agents interact with one another for the achievement of some goals. Agent communication language is a proposed standard language for agent communications like FIPA-ACL. In [26] FIPA provides a specification for agent Communication. FIPA-ACL suggests the parameters for effective communication among the agents within the scenario. All performatives of FIPA contain message structure i-e sender, receiver and content of the message. The message expresses the meaning of agents. It consists of the content and action of the communication. If an agent communicates with another agent, a suitable performative is used. If the agent does not understand some message and is unable to the process the sent message, it can reply with the "not understand" performative. In [26] FIPA has proposed 20 performatives that cover all the maximum range of aspect of agents expected communication. FIPA-ACL is now a basic standard for determining the encoding and exchange of messages among agents. In [27] a set of performatives is given for ACLs that specifies how these communicating actions should be executed in a concurrent and reactive way with respect to a given logical semantics. It has assumptions about how the recipient should react to the message.

III. Proposed TAPN based FIPA Performatives

In our designed model of performatives, workflow shows within time highlights, enables transitions, sends time interval and manages task completion within deadlines. Our model is the combination of FIPA performatives and Timed-Arc Petri-Nets in which token sends the message with time constraints. We use places as agents which generate tokens. Tokens convey an age and the age is exactly equal to the time interval that is defined on arc. Arc defines the area and limit within which transition takes place. Transition firing and enabling depend on the age of the token. In our proposed solution, each performative must be sent within seven seconds. When we map these performatives to a real-time multi-agent system, each performative takes predefined time but the starting time and ending time of transition may change according to the RTMAS. To save space only two out of twenty performatives have been specified in this paper.

A. Request

Using request performative, one agent requests another agent to take some action. For the simulation of this act, we have used two places and six transitions. The agent1 place has an initial marking 0.0 token that represents the age of the token. Agent1 requests to Agent2 as transition firing <Request_A1_A2> from Agent1 to <Request_A1_A2> and <Request_A1_A2> to A2. Time interval on arc restricts the token firing time that must be fired within [StRequest, EnRequest] including delay. Agent2 accepts the request by passing the message of Agree as transition firing <Agree_A2_A1> and can also refuse the request as transition firing <Refuse_A2_A1>. A1 receives the response of Agent2. In case Agent2 agrees to the request of Agent1, token is fired on Agent1 place with the age 0 for the response of agree and another token is fired on Agent2 place to continue the process with same age that has been consumed at the fired transition. For this purpose, we have used transport arc.

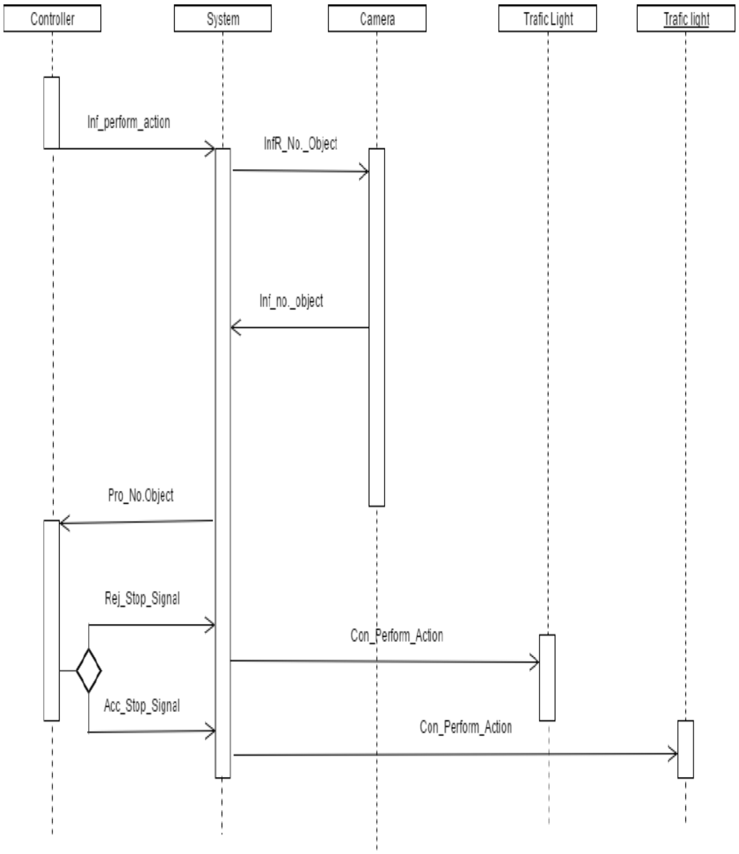


Fig. 3. Proposed Sequence of Traffic Light System.

In response of Agree, Agent2 replies in the form of <InformDone_A2_A1> or within [StInfDone , EnInfDone] if the task is executed within the deadline. Agent2 can reply in the form of <InfoRef_A2_A1> within [StInfoRef , EnInfoRef] in case of detailed reply. If the agent fails to fulfill the requirement within the specified deadline, reply <Failure_A2_A1> comes within [StFailure , EnFailure]. Finally, the task is executed within time limit with the options of EnFailure, EnInfoRef or EnInfDone. Simulation history of the two agents is represented in Fig. 1 and the results of temporal constraints are shown in Table I.

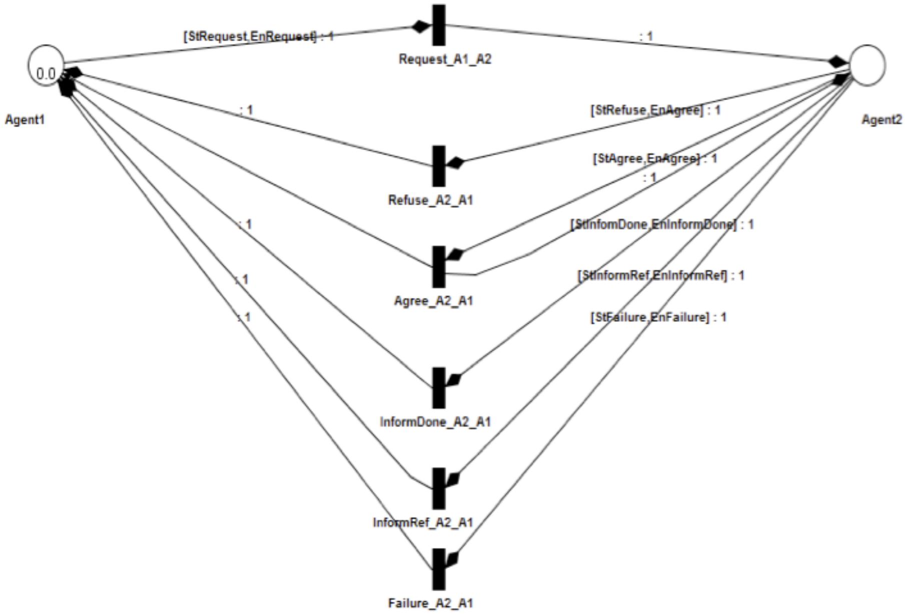


Fig. 1. TAPN based Request Performative for two agents.

TABLE I. TRANSITION AND TIMING DURATION OF TWO AGENTS IN REQUEST PERFORMATIVE.

Place	Transition	Total Time Duration	Start Time	End Time
Agent1	Request_A1_A2	7	0	7
Agent2	Refuse_A2_A1	7	8	15
Agent2	Agree_A2_A1	7	8	15
Agent2	InformDone_A2_A1	7	16	23
Agent2	InfoRef_A2_A1	7	16	23
Agent2	Failure_A2_A1	7	16	23

TABLE II. TRANSITION AND TIMING DURATION OF ONE TO MANY AGENTS INTERACTION IN CFP PERFORMATIVE.

Place	Transition	Total Time Duration	Start Time	End Time
Agent1	CFP_A1_A2A3	7	0	7
Agent2	Proposal_A2_A1	7	8	15
Agent2	Refuse_A2_A1	7	8	15
Agent2	NotUnderstand_A2_A1	7	8	15
Agent3	Proposal_A3_A1	7	8	15
Agent3	Refuse_A3_A1	7	8	15
Agent3	NotUnderstand_A3_A1	7	8	15
Agent1	RejectProposal_A1_A2	7	16	23
Agent1	RejectProposal_A1_A2	7	16	23
Agent1	AcceptProposal_A1_A2	7	24	31
Agent1	AcceptProposal_A1_A3	7	24	31
Agent2	InfoDone_A2_A1	7	32	39
Agent2	InfoRef_A2_A1	7	32	39
Agent2	Failure_A2_A1	7	32	39
Agent3	InfoDone_A3_A1	7	32	39
Agent3	InfoRef_A3_A1	7	32	39
Agent3	Failure_A3_A1	7	32	39

B. Call for Proposal

Call for proposal (CFP) is used to start communication between agents. FIPA CFP performative augments approval and denial of preceding form of communicating agents. In CFP performative, one agent acts as a manager, that requires a certain task to be accomplished effectively within a specified time. In CFP performative, the manager sends the call for proposal to the contractors. The contractors send the reply in one of the three forms that are refused, not understanding and proposal within the time limit. After getting the response from the contractors, the manager approves one of the proposals and sends rejection (reject proposal) to the remaining agents. The selected contractor then apprises the manager of task completion. In CFP performative, we have used three places and fifteen transitions. Agent1 place has 0.0 token that represents the age of the token. Agent1 CFP to Agent2 and Agent3 is represented as transition firing <CFP_A1_A2A3>. Time interval on arc restricts the token firing time that must be fired within [StCFP, EnCFP] including delay. Agent2 accepts the message by passing the proposal as transition firing <Proposal_A2_A1> and refuses transition firing <Refuse_A2_A1>. Agent1 can accept the proposal as transition firing <AcceptProposal_A1_A2> within [StAccept , EnAccept]. For accept Proposal, we have used transport arc which means that the age of the token remains same as the age of token at the firing time. Agent1 can reject the proposal as transition firing <RejectProposal_A1_A2> with time constraints [StReject, EnReject]. In case of reject proposal, the token is transferred to Agent2 or Agent3. The agent replies further within time constraint if the proposal is accepted by Agent1. After Acceptance of proposal, Agent2 replies in the form of <InformDone_A2_A1> or <InformDone_A3_A1> within [StInfDone, EnInfDone] if the task is completed within deadline. Agent2 or Agent3 can reply in the form of <InfoRef_A2_A1> <InforRef_A3_A1> within [StInfoRef, EnInfoRef] if detailed reply. If agent fails to fulfil the proposal requirements within fixed deadline then it replies as <Failure_A2_A1> or <Failure_A3_A1> within [StFailure,

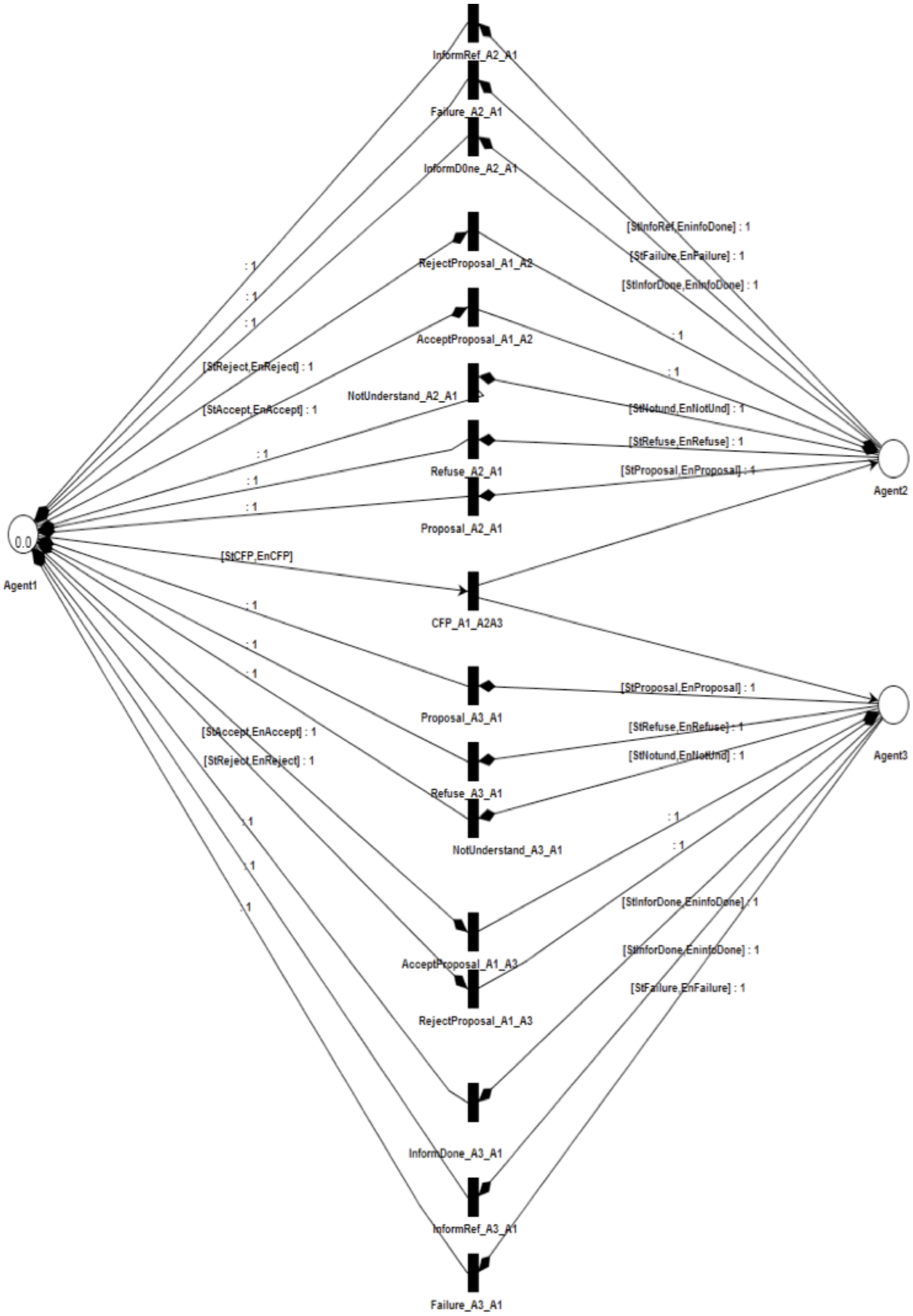


Fig. 2. TAPN based CFP Protocol for one to many agents interaction.

EnFailure]. The complete process of communication for CFP performative is represented in Fig. 2 and the results of temporal constraints are shown in Table II.

IV. Verification of the Proposed TAPN based FIPA Performatives

We have verified the proposed solution by using the properties of liveness, reachability and boundedness. The results of these properties are shown below stating whether the property is satisfied or not.

A. Request

The interaction between two or many agents for the request protocol has been verified here. In Table III verification results of the properties of interest of the system specified in TCTL fragments are shown. Table IV and Table V shows the result for K-boundedness of the system.

TABLE III. Properties for Request Performative and TCTL fragments for TAPAAL.

Query	Formula	Result
Are both agents reachable?	EF (Request1to1.Agent2 = 1)	Satisfied
Are all agents accessible?	EF (Request.Agent1 =1 and Request.Agent2 = 1 and Request.Agent3 =1)	Satisfied

TABLE IV. K-boundedness of Request Performative "Are both agents reachable?".

Count	Transition
1	Request1to1.Request_A1_A2
0	Request1to1.Refuse_A2_A1
1	Request1to1.Agree_A2_A1
1	Request1to1.InformDone_A2_A1
1	Request1to1.InformRef_A2_A1
1	Request1to1.Failure_A2_A

TABLE V. K-boundedness of Request Performative "Are all agents accessible?".

Count	Transition
4	Request.Request_A1_A3
4	Request.Request_A1_A2
3	Request.Refuse_A3_A1
3	Request.Agree_A3_A1
3	Request.InformDone_A3_A1
3	Request.InformRef_A3_A1
3	Request.Failure_A3_A1
1	Request.Refuse_A2_A1
1	Request.InformDone_A2_A1
1	Request.InformRef_A2_A1
1	Request.Failure_A2_A1
1	Request.Agree_A2_A1

TABLE VI. Properties for CFP Performative and TCTL formula.

Query	Formula	Result
Is agent2 globally reachable?	EG (CFP.Agent2 = 1)	Not Satisfied
Are both agent2 and agent3 reachable?	EF ((CFP.Agent1 = 2 and CFP.Agent2 = 0) or (((!(CFP.Agent1=2) and CFP.Agent4=1) and CFP.Agent=1 and !(CFP.Agent1 = 2)) and !(CFP.Agent3 = 2))	Satisfied
All these states reachable?	EF ((CFP.Agent1 = 2 and CFP.Agent2 = 0) or (((!(CFP.Agent1=2) and CFP.Agent4=1) and CFP.Agent=1 and !(CFP.Agent1 = 2)) and !(CFP.Agent3 = 2)))	Satisfied

TABLE VII. K-boundedness of CFP Performative "All these states reachable?"

Count	Transition	Place	Maximum Token
8	CFP.InformDone_A1_A2A3	CFP.Agent1	2
1	CFP.Proposal_A3_A1	CFP.Agent3	1
1	CFP.Refuse_A3_A1	CFP.Agent2	1
1	CFP.Proposal_A2_A1		
1	CFP.NotUnderstand_A3_A1		
0	CFP.AcceptProposal_A3_A1		
0	CFP.RejectProposal_A3_A1		
0	CFP.InformDone_A3_A1		
0	CFP.InformRef_A3_A1		
0	CFP.Failure_A2_A1		
0	CFP.AcceptProposal_A1_A2		
0	CFP.RejectProposal_A1_A2		
0	CFP.InformDone_A2_A1		
0	CFP.InformRef_A2_A1		
0	CFP.Failure_A2_A1		
0	CFP.InformRef_A3_A1		
0	CFP.NotUnderstand_A3_A1		

TABLE VIII. K-boundedness of CFP Performative "Are both agent2 and agent3 reachable?"

Count	Transition	Place	Maximum Token
1	CFP.CFP_A1_A2A3	CFP.Agent1	1
0	CFP.Proposal_A3_A1	CFP.Agent3	1
0	CFP.Refuse_A3_A1	CFP.Agent2	1
0	CFP.AcceptProposal_A1_A3		
0	CFP.RejectProposal_A1_A3		
0	CFP.InformDone_A3_A1		
0	CFP.InformRef_A3_A1		
0	CFP.Failure_A3_A1		
0	CFP.Proposal_A2_A1		
0	CFP.Refuse_A2_A1		
0	CFP.AcceptProposal_A1_A2		
0	CFP.RejectProposal_A1_A2		
0	CFP.InformD0ne_A2_A1		
0	CFP.Failure_A2_A1		
0	CFP.InformRef_A2_A1		
0	CFP.Aceept		
0	CFP.Accept		
0	CFP.NotUnderstand_A3_A1		
0	CFP.NotUnderstand_A2_A1		

TABLE IX. Properties for TLS and TCTL Fragments for TAPAAL.

Query	Formula	Result
Are all systems reachable?	EF (TLS.System1 = 1 and TLS.System2 = 1 and TLS.System3 = 1 and TLS.System4 = 1)	Satisfied
Are all cameras work properly?	EF (TLS.Camera1 = 1 and TLS.Camera2 = 1 and TLS.Camera4 = 1 and TLS.Camera3 = 1)	Satisfied

TABLE X. EXECUTION COUNT FOR ALL THE SIGNALS IN TAPAAL.

Count	Transition
31	TLS.Acc_Op_Sig1
31	TLS.Acc_Op_Sig4
31	TLS.Acc_Op_Sig3
31	TLS.Acc_Op_Sig2
26	TLS.Rej_St_Sig2
26	TLS.Rej_St_Sig1
26	TLS.Rej_St_Sig
26	TLS.Rej_St_Sig3
21	TLS.Inf_Per_Act1
21	TLS.Inf_Per_Act3
21	TLS.Inf_Per_Act4
21	TLS.Inf_Per_Act2
14	TLS.Con_Per_ActG1
13	TLS.Con_Per_ActR1
12	TLS.Pro_N_Obj1
12	TLS.InfR_N_Obj1
8	TLS.Pro_N_Sig3
8	TLS.Con_Per_ActG3
8	TLS.InfR_N_Obj3
8	TLS.Con_Per_ActR3
7	TLS.Pro_N_Obj4
7	TLS.Con_Per_ActR4
7	TLS.Con_Per_ActG4
7	TLS.InfR_N_Obj4
6	TLS.Pro_N_Obj2
6	TLS.InfR_N_Obj2
6	TLS.Con_Per_ActG2
6	TLS.Con_Per_ActR2
2	TLS.Go1
2	TLS.Stop1
2	TLS.Inf_N_Obj1
1	TLS.Inf_N_Obj2
1	TLS.Go2
1	TLS.Stop2
1	TLS.Stop4
1	TLS.Go4
1	TLS.Inf_N_Obj4
1	TLS.Go3
1	TLS.Inf_N_Obj3
1	TLS.Stop3

TABLE XI. EXECUTION COUNT FOR TRAFFIC LIGHT SYSTEMS IN TAPAAL.

Count	Transition
1122	TLS.Con_Per_ActG1
967	TLS.Con_Per_ActG3
891	TLS.Con_Per_ActG4
846	TLS.Con_Per_ActG2
725	TLS.Acc_Op_Sig1
725	TLS.Acc_Op_Sig4
725	TLS.Acc_Op_Sig3
725	TLS.Acc_Op_Sig2
673	TLS.Go1
528	TLS.Go3
465	TLS.Go4
431	TLS.Go2
421	TLS.Con_Per_ActR1
388	TLS.Con_Per_ActR3
372	TLS.Con_Per_ActR4
367	TLS.Stop1
364	TLS.Con_Per_ActR
299	TLS.Stop3
272	TLS.Stop4
256	TLS.Stop2
165	TLS.Rej_St_Sig2
165	TLS.Rej_St_Sig1
165	TLS.Rej_St_Sig
165	TLS.Rej_ST_Sig3
159	TLS.Pro_N_Obj1
159	TLS.InfR_N_Obj1
154	TLS.Pro_N_Sig3
154	TLS.InfR_N_Obj3
151	TLS.Pro_N_Obj4
151	TLS.InfR_N_Obj4
150	TLS.Pro_N_Obj2
150	TLS.InfR_N_Obj2
135	TLS.Inf_N_Obj1
122	TLS.Inf_N_Obj3
118	TLS.Inf_N_Obj4
115	TLS.Inf_N_Obj2
35	TLS.Inf_Per_Act1
35	TLS.Inf_Per_Act3
35	TLS.Inf_Per_Act4
35	TLS.Inf_Per_Act2

B. *Call for Proposal*

The interaction between two or many agents for the CFP protocol has been verified here. In Table VI, verification results of the properties of interest of the system specified in TCTL fragments are shown. Table VII and Table VIII shows the result for K-boundedness of the system.

V. APPLICATION OF THE PROPOSED APPROACH

In this section, we have formally modeled FIPA performatives in RTMAS using TAPN to demonstrate the application of the proposed approach to real-time applications. We have formally specified and verified the Traffic Light System (TLS) based on RTMAS using TAPN. In TLS, the traffic actions are governed by TAPN in to control the lights intersections. The proposed TAPN based application is an eight line traffic system. The signals are used in a cyclic arrangement of on and off. The presented technique can be used to manage the rush of traffic, balance traffic flow, and traffic safety. The application has been explained through a sequence transactions and TAPN.

A. *Traffic Light System*

The proposed application is based on eight lines. A signal gets on and off in a cyclic pattern. Cameras are fixed to detect vehicles and inform to the system about the number of vehicles. We use the following agents which are controller, System, Camera, <YG_Traffic_Light> <G_Traffic_Light>, <YR_Traffic_Light>, <R_Traffic_Light>. Complete working of the system is shown in Fig. 3.

A large number of vehicles are managed by traffic lights so their efficient and smooth working is very important in maintaining traffic flow and to prevent accidents. In our proposed application, we have attempted to map FIPA performatives for traffic light system with timing aspect. We have used an eight lines traffic system comprising the following agents; controller, system, camera, <YR_T_Light>, <YG_T_Light>. TAPN based specification of the system is shown in Fig. 4. The controller agent manages the system and gives instructions to other systems to perform the task. System manages the cameras that are fixed above the lines with the traffic signal. The controller gives instructions to the system to perform action. For this scenario, we have used 'inform' performative that works in the controller to <Inf_Per_Act> and <Inf_Per_Act> to system. <Inf_Per_Act> is the same for all systems that is <Inf_Per_Act1>, <Inf_Per_Act2>, <Inf_Per_Act3> <Inf_Per_Act4>. After getting instructions from the controller, the system gives instructions to its own camera that is fixed along with traffic lights. For sending the instruction from the system to the camera <InfR_N_obj1>, <InfR_N_obj2>, <InfR_N_obj3> and <InfR_N_obj4> are used for camera1, camera2, camera3, and camera4, respectively. Sensor

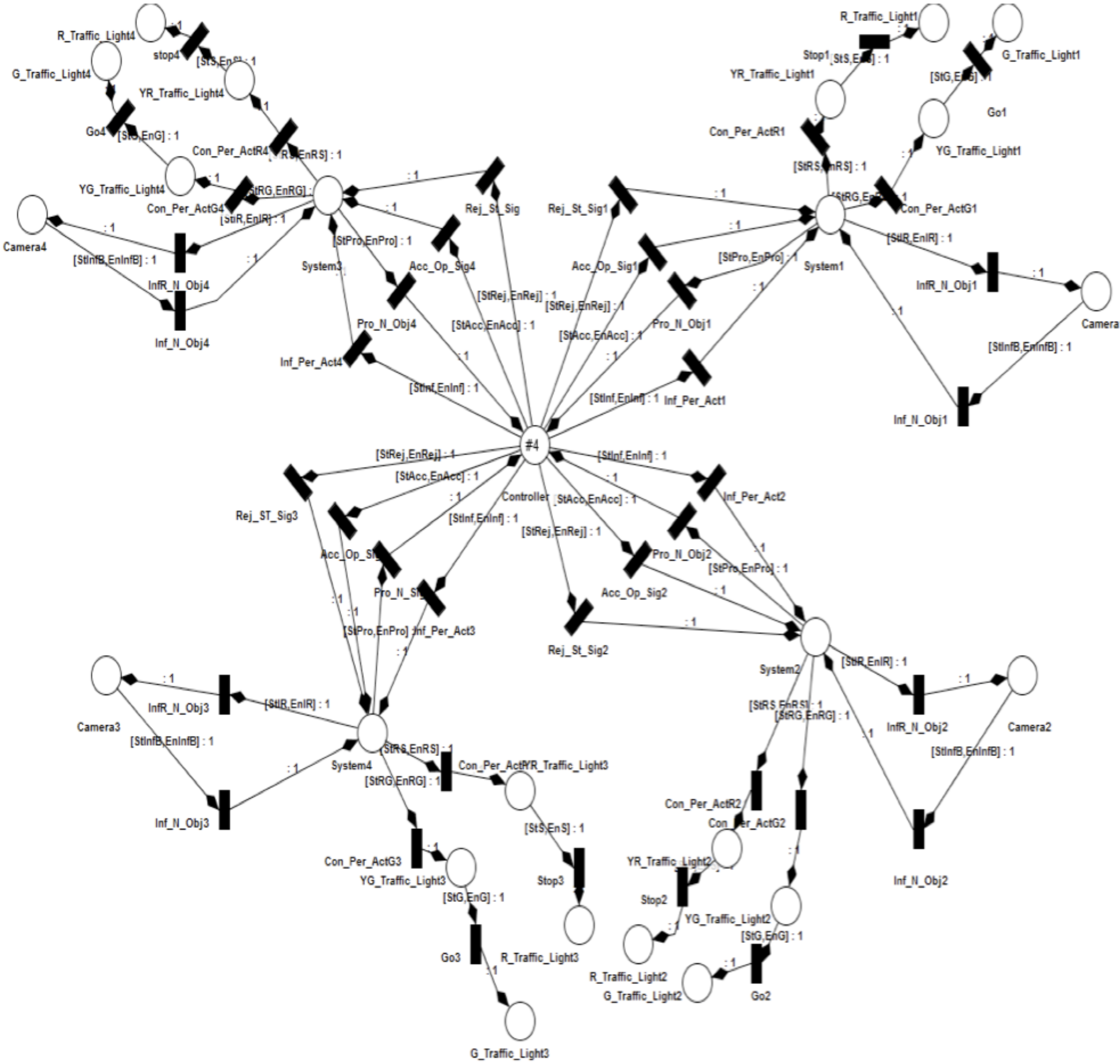


Fig. 4. TAPN based Traffic Light System.

senses the presence of vehicles and gives information to the system using <Inf_N_Obj> transition that is completed in [StInfO,EnInfO]. Now each system tells the controller about the number of vehicles using the proposal interaction protocol. Transition varies from system to <Pro_N_Obj> and <Pro_N_Obj> to the controller. The controller checks each system proposal and finds the maximum number of vehicles. It rejects the proposal of system that does not have the maximum number of vehicles. Next, the system sends message to yellow traffic light agent that is <YR_Traffic_Light> to get ready and the <YR_Traffic_Light> is turned to <R_Traffic_light> to perform the action of stop. The controller sends the acceptance message to the system that consists of the maximum number of vehicles and the System sends a message to <YG_Traffic_Light> to get ready. Then <YG_Traffic_Light> is shifted to <G_Traffic_light> using go transition within [StG, EnG]. In Table IX, verification results of the properties of interest of the TLS specified in TCTL fragments are shown. Table X and Table XI shows the transition count of different arcs during simulation. The higher number of counts represents more traffic at that signal.

TAPN has been used for the verification of boundedness, reachability and liveness properties. Formal verification gives the correctness of the system. Boundedness ensures the maximum and the minimum number of tokens that each place holds. Reachability determines the sequence from the first node of marking to the second node. In our proposed application, all the states are reachable. Liveness determines that the application is executable. All places can contain any number of tokens throughout the life cycle. Our system is deadlock free because there is no such place where deadlock can occur. The dead-lock free system implies that the system is live and all places are working properly. All these properties ensure the correctness of our system. We have also achieved the inter operability and a well-defined process by using FIPA-ACL standards in our application.

VI. Conclusion

In this research, we have formally modeled FIPA performatives in RTMAS using TAPN for agent communication. Communication of agents is a significant characteristic of RTMAS and is useful for message interaction in real-time multi-agent systems. Previous work has focused on formal modeling of domain functionality of multi-agent systems and not on the agent's interaction level. The formal specification and verification of these multi-agent systems in which the agents interact with one another to accomplish their objectives with time constraints ensures their reliability. TAPAAL has been used for the verification of the properties of interest of the system specified through AF, AG, EG and EF fragments of TCTL. The research provides future directions for formal modeling of standardized agent's interaction with timing constraints. The approach ensures that the system is deadlock free and live. For the future, we will work on FIPA performatives with Timed Colored Petri-nets in RTMAS.

References

[1] N. R. Jennings, K. Sycara, and M. Wooldridge, "A roadmap of agent research and development," *Autonomous agents and multi-agent systems*, vol. 1, no. 1, pp. 7–38, 1998.

[2] B. Marzougui and K. Barkaoui, "Interaction protocols in multi-agent systems based on agent petri nets model," *Int J Adv Comput Sci Appl*, vol. 4, no. 7, 2013.

[3] A. Qasim, S. A. R. Kazmi, and I. Fakhir, "Formal specification and verification of real-time multi-agent systems using timed-arc petri nets," *Adv. Elect. Comput. Eng.*, vol. 15, no. 3, pp. 73–78, 2015.

[4] A. Qasim and S. A. R. Kazmi, "Mape-k interfaces for formal modeling of real-time self-adaptive multi-agent systems," *IEEE Access*, vol. 4, pp. 4946–4958, 2016.

[5] I. Obaid, S. A. R. Kazmi, and A. Qasim, "Modeling and verification of payment system in e-banking," *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, vol. 8, no. 8, pp. 195–201, 2017.

[6] J. Pitt and A. Mamdani, "Some remarks on the semantics of fipa's agent communication language," *Autonomous Agents and Multi-Agent Systems*, vol. 2, no. 4, pp. 333–356, 1999.

[7] L. Chang, X. He, and S. M. Shatz, "A methodology for modeling multi-agent systems using nested petri nets," *International Journal of Software Engineering and Knowledge Engineering*, vol. 22, no. 07, pp. 891–925, 2012.

[8] D. Juneja, A. Jagga, and A. Singh, "A review of fipa standardized agent communication language and interaction protocols," *Journal of Network Communications and Emerging Technologies*, vol. 5, no. 2, pp. 179–191, 2015.

[9] C. Zaghetto, L. H. M. Aguiar, A. Zaghetto, C. G. Ralha, and F. de Barros Vidal, "Agent-based framework to individual tracking in unconstrained environments," *Expert Systems with Applications*, vol. 87, pp. 118–128, 2017.

[10] Y.-S. Huang, Y.-S. Weng, and M. Zhou, "Design of traffic safety control systems for emergency vehicle preemption using timed petri nets," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2113–2120, 2015.

[11] D. Calvaresi, M. Marinoni, A. Sturm, M. Schumacher, and G. Buttazzo, "The challenge of real-time multi-agent systems for enabling iot and cps," in *Proceedings of the international conference on web intelligence*. ACM, 2017, pp. 356–364.

[12] C. Shum, W. H. Lau, T. Wong, T. Mao, S. Chung, C. Tse, K. F. Tsang, and L. L. Lai, "Modeling and simulating communications of multiagent systems in smart grid," in *2016 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 2016, pp. 405–410.

[13] A. Pla, P. Gay, J. Meléndez, and B. López, "Petri net-based process monitoring: a workflow management system for process modelling and monitoring," *Journal of Intelligent Manufacturing*, vol. 25, no. 3, pp. 539–554, 2014.

[14] S. Khosravifar, "Modeling multi agent communication activities with petri nets," *International Journal of Information and Education Technology*, vol. 3, no. 3, p. 310, 2013.

[15] B. Marzougui, K. Hassine, and K. Barkaoui, "A new formalism for modeling a multi agent systems: Agent petri nets," *Journal of Software Engineering and Applications*, vol. 3, no. 12, p. 1118, 2010.

[16] J. F. Jensen, T. Nielsen, L. K. Oestergaard, and J. Srba, "Tapaal and reachability analysis of p/t nets," in *Transactions on Petri Nets and Other Models of Concurrency XI*. Springer, 2016, pp. 307–318.

[17] M. Andersen, H. G. Larsen, J. Srba, M. G. Sørensen, and J. H. Taankvist, "Verification of liveness properties on closed timed-arc petri nets," in *International Doctoral Workshop on Mathematical and Engineering Methods in Computer Science*. Springer, 2012, pp. 69–81.

[18] G. Guo, L. Ding, and Q.-L. Han, "A distributed event-triggered transmission strategy for sampled-data consensus of multi-agent systems," *Automatica*, vol. 50, no. 5, pp. 1489–1496, 2014.

[19] G. S. Seyboth, D. V. Dimarogonas, and K. H. Johansson, "Event-based broadcasting for multi-agent average consensus," *Automatica*, vol. 49, no. 1, pp. 245–252, 2013.

[20] W. Louhichi, B. Marzougui, and K. Hassine, "Formal model for coordination in multi-agents system based petri net agent," in *2017 International Conference on Smart, Monitored and Controlled Cities (SM2C)*. IEEE, 2017, pp. 134–137.

[21] P. G. Jensen, K. G. Larsen, and J. Srba, "Discrete and continuous strategies for timed-arc petri net games," *International Journal on Software Tools for Technology Transfer*, vol. 20, no. 5, pp. 529–546, 2018.

[22] J. A. Mateo, J. Srba, and M. G. Sørensen, "Soundness of timed-arc workflow nets," in *International Conference on Applications and Theory of Petri Nets and Concurrency*. Springer, 2014, pp. 51–70.

[23] A. Kaur and S. Jain, "Kqml-from scenario to technology," *International Journal of Advanced Studies in Computers, Science and Engineering*, vol. 7, no. 3, pp. 30–34, 2018.

[24] W. M. Zuberek, "Timed petri nets definitions, properties, and applications," *Microelectronics Reliability*, vol. 31, no. 4, pp. 627–644, 1991.

[25] J. Byg, K. Y. Jørgensen, and J. Srba, "Tapaal: Editor, simulator and verifier of timed-arc petri nets," in *International Symposium on Automated Technology for Verification and Analysis*. Springer, 2009, pp. 84–89.

[26] P. D. O'Brien and R. C. Nicol, "Fipa—towards a standard for software agents," *BT Technology Journal*, vol. 16, no. 3, pp. 51–59, 1998.

[27] N. Dragoni and M. Gaspari, "Performative patterns for designing verifiable acls," in *International Workshop on Cooperative Information Agents*. Springer, 2006, pp. 375–387.

Scale and Resolution Invariant Spin Images for 3D Object Recognition

Jihad H'roura[1], Aissam Bekkari[2],
Driss Mammass[3], Ali Bouzit[4]
IRF-SIC Laboratory
Ibn Zohr University
Agadir, Morocco

Patrice Méniel[5]
ARTEHIS laboratory,
University of Bourgogne Franche-Comté
Dijon, France

Alamin Mansouri[6], Michaël Roy[7]
Le2i Laboratory
University of Bourgogne Franche-Comté
Auxerre, France

***Abstract*—Until the last decades, researchers taught that teaching a computer how to recognize a bunny, for example, in a complex scene is almost impossible. Today, computer vision system do it with a high score of accuracy. To bring the real world to the computer vision system, real objects are represented as 3D models (point clouds, meshes), which adds extra constraints that should be processed to ensure a good recognition, for example the resolution of the mesh. In this work, based on the state of the art method called Spin Image, we introduce our contribution to recognize 3D objects. Our motivation is to ensure a good recognition under different conditions such as rotation, translation and mainly scaling, resolution changes, occlusions and clutters. To that end we have analyzed the spin image algorithm to propose an extended version robust to scale and resolution changes, knowing that spin images fails to recognize 3D objects in that case. The key idea is to approach the representation of spin images of the same object under different conditions by the mean of normalization, either these conditions result in linear or non-linear correlation between images. Our contribution, unlike spin image algorithm, allows to recognize objects with different resolutions and scale. Plus it shows a good robustness to occlusions up to 60% and clutters up to 50%, tested on two datasets: Stanford and ArcheoZoo3D.**

***Keywords*—3D object; recognition; spin image; resolution; scaling**

I. Introduction

New information and communication technologies have emerged in 1990s and have grown exponentially in power. The digital revolution which has been developing since its birth at the end of 20th century, has affected different sectors throughout the world. This revolution has led to the emergence of new type of data, resulting in new and broader databases, particularly 3D data. This requires technological advances in image processing or overall computer vision. Given the very wide spectrum of industrial, military and medical applications that can be considered, this field in its turn has developed very quickly. In the context of this digital revolution, notably in cognitive sciences, scientists in computer vision have redirected their efforts to put into place a variety of interactive applications with 3D real world, like 3D object recognition. To that aim, a better understanding of how the human visual system works is necessary. A first classical hypothesis assumes that, to recognize an object, the human brain starts by extracting features from objects captured by eyes. Then, depending on his previous knowledge, he elaborates a matching process. Nevertheless, with the development of neuro-sciences, scientists assume that data in human brain travels in the neural networks where each node performs a separate task, to achieve the visual cortex where the recognition is performed based on its previous trained data. So, inspiring from this discovery, researchers in computer vision give another perspective called deep learning. Concerning classical hypothesis, different approaches have been proposed depending on the explored levels of the object and also extracted features. If the targeted level is global and tends to describe the overall shape of the object, we talk about global approaches. In the other hand, if the method focuses on extracting only local features, they are called local methods. Proposed approaches, either local or global aim to ensure the robustness to different condition 3D objects in real scenes can be through for example, rotation, translation, geometric deformations, occlusions, clutters, scaling, etc. In this respect, when it comes to occlusions and clutters, local approaches are known to be more efficient. Other strengths of this category is that they are popular to not requiring any segmentation and the pose estimation is simpler. However, the fact that local approaches are founded on local neighborhood, which is highly affected by the resolution changes, make them less discriminating. In addition, a verification, step is always needed to eliminate incorrect correspondences and the spatial information is missed. Concerning global methods, they are more discriminating since they provide a global description of the shape of the object. Besides, by only computing the nearest neighbor of the descriptor, we can perform matching, which makes it easier. In the opposite to local methods, they do not handle occlusions and clutters, the pose estimation is more complicated and they usually require a segmentation as a pre-processing. In this paper we introduce a novel local shape based approach approach for 3D object recognition, crafted to deal with resolution and scaling changes of the object in occluded and cluttered scenes. Our contribution, called Invariant to scale and resolution spin images (ISRSI), is based on a state of the art method called spin images. Spin images fails when the resolution and the scale of objects change. By performing a normalization step and defining efficiently the required parameters, we succeed to make this descriptor invariant to scale and resolution changes. Our contribution has shown good robustness to occlusions up to 60% and clutters up to 50%. The paper is laid out as follow. We briefly quote some related works in Section 2. Then, in Section 3 we describe the background method. The Section 4 is dedicated to introduce more details about our contribution. While experiments are conducted in Section 5. And finally, a conclusion is given in

Section 6.

II. Related Works

3D free form objects recognition is a very challenging task due to the presence of different conditions revealed in the real world to take into account, like occlusions, clutters and other transformations such as scaling, rotation and translation. Besides, the 3D reconstruction of real objects adds more constraints mainly mesh resolution changes. To that end, researchers have proposed different range of methods. The stat of the art introduces different survey on 3D object recognition approaches [1] and [2]. One can classify those methods to shape based approaches, local shape-based approaches, topological approaches and view-based approaches. Global shape-based approaches: As their name indicates it, they aim to describe the coarse shape of the 3D model. In this direction, Osada et al. [3] represent an object as a shape distribution by elaborating five functions based of the choice of a random set of points. Authors have shown that their approach is invariant to geometric transformations. Another approach have been proposed by Paquet et al. in [4] that can be used in the same time for 2D and 3D objects, have shown good robustness to resolution, translation and rotation. Local shape based approaches: are also known sometimes as key point based methods. In this branch we find a multiscale approach proposed by Nouri et al. [5]. They use patches with adaptive size to detect salient regions on the surface of a 3D model. Tang et al. [6] have proposed a local descriptor based on geometric centroids. Another method have been introduced by Maes et al. [7] as an extension of SIFT descriptor [8] to the 3D domain. Spin image descriptor [15] is another approach that aims to explore the local distribution of vertices on the surface area of the object to create a set of 2D images considered as the descriptor of the object. View based approaches: or in other words 3D/2D approaches describe object based on its projections in a 2D space. For example, Xiang et al. [9] have introduced a new descriptor called 3DVP for 3D voxel pattern encodes the object by a triplet (appearance, 3D shape, occlusions). In another contribution in [10] authors compute different features for object's views, such as 2D Fourier descriptor, 2D Zernike moments and 2D Krawtchouk moments. Topological approaches: we cite here for example the contribution of Pickup et al. [11]. It consists of constructing the skeleton based on Au et al's technique [12]. Then 3D pose normalization is performed using the canonical form of the skeleton of the object. And finally, utilizing Yan et al's approach [13] a deformation of the mesh is fulfilled in order to match the canonical transformation of its skeleton. Another approach aims to improve Reeb Graph of an object has been presented by Thierny et al. in [14] following three steps: 1) Extraction of salient vertices. 2) Emphasis of the overall shape of the object using an application function. 3) Refinement of Reeb graph into topological skeleton by the mean of constrictions.

III. Background: Spin Image Algorithm

Spin image descriptor is an algorithm that has been first introduced by Johnson et al. [15]. The 3D mesh model is described by a set of its 2D projections on a well-defined 2D local coordinate systems. In order to define a local coordinate system, authors first define an oriented point O as the center of this local basis. The oriented point in its turn is defined by a vertex $p(x, y, z)$ and a normal surface n. The normal surface is the plan tangent to the vertex p and perpendicular to its vector normal n. Then to define the two cylindrical coordinates α and β are computed for each other vertex x on the surface mesh such as:

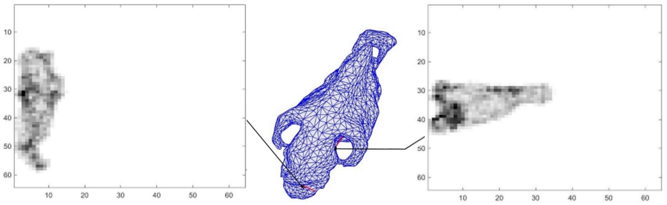


Fig. 1. Two spin images from two oriented points on the surface mesh of skull model.

$$\alpha = \sqrt{||x-p||^2 - (\vec{n}.(x-p))^2} \tag{1}$$

$$\beta = \vec{n}.(x-p) \tag{2}$$

So for each vertex a corresponding spin image is obtained using this projection function below:

$$S_O : R^3 \mapsto R^2$$

$$S_O(x) \mapsto (\alpha, \beta) = (\sqrt{||x-p||^2 - (\vec{n}.(x-p))^2}, \vec{n}.(x-p)) \tag{3}$$

During projection of vertices, authors have specified three parameters to take into account. First, we have bin size b which specifies the size of bins used to accumulate points projected. Then the angle support ϕ, it is the angle between the normal vector of each vertex to project and the normal vector of oriented point. Lastly, is the width W of the spin image. Equations (4) and (5) shows the relation between those three parameters.

$$i = \left\lfloor \frac{\frac{W}{2} - \beta}{b} \right\rfloor \qquad j = \left\lfloor \frac{\alpha}{b} \right\rfloor \tag{4}$$

$$a = \alpha - ib \qquad b = \beta - jb \tag{5}$$

Fig. 1 illustrates two spin images from two oriented point on the surface mesh of a horse's skull from ArcheoZoo3D database.

For the purpose of performing a matching between two objects, authors have put into place a surface matching algorithm following different stages. We summarize the different phases in the pipeline below. See Fig. 2. For detailed description readers can refer to [16].

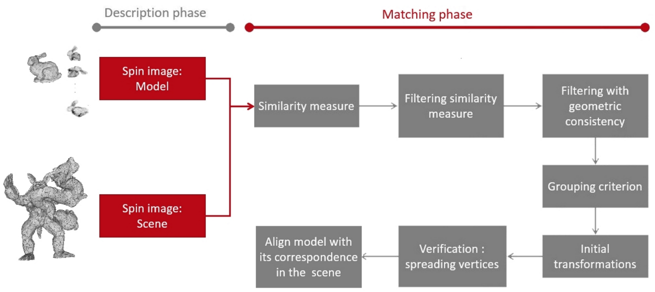

Description phase
Matching phase
Spin image: Model
Spin image: Scene
Similarity measure
Filtering similarity measure
Filtering with geometric consistency
Grouping criterion
Initial transformations
Verification : spreading vertices
Align model with its correspondence in the scene

Fig. 2. Pipeline of spin images matching.

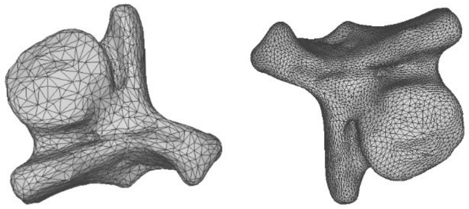


Fig. 3. An example of 3D mesh of caudal with two different resolutions.

IV. Scale and Resolution Invariant Spin Images: SRISI

A. Invariance to Resolution

At the end of the eighties, efforts to reproduce three dimensional world have borne fruit and the first 3D scanning systems, based on imaging triangulation, were installed for industrial applications. After decades, high definition 3D scanners are at the forefront of archeology field, for a wide range of items, small-sized artifacts such as coins, teeth and bones, fragments and scripts up to significant figures, statues and small buildings. Thus, bringing history back to the life. According to a 3D scanning pipeline [17], a complete 3D model of the object is provided, in general in the form of 3D meshes. Here comes other challenges relied to the representation of the object to take into consideration, in order to insure a good recognition. One of these parameters is the resolution of the mesh, which is defined here as the lengths of edges of the mesh, or precisely, the median of the lengths of all edges of the mesh. So the same object can be represented with different resolutions, implicitly different number of vertices, see Fig. 3.

As we have cited above, the spin image algorithm is known to be robust to occlusions and clutters, but when it comes to spot the same object with different resolution in the scene, or when we change the scale, the process of recognition fails. For the purpose of making spin image algorithm robust to resolution changes, we need first to understand what the impact of resolution changes is on the description phase that makes this algorithm fails and at which level of the matching phase the process crashes?

To that aim, let us back up a moment and talk about the creation of spin images, mainly, how we generate a spin map for each oriented point. The generation of spin images is controlled by three parameters. The first parameter is the bin size, which is defined as a multiple of the resolution of the surface mesh. Then the support angle, that controls the vertices to be projected based on the angle between their normal vectors and the normal vector of the oriented point. The third parameter is W the width of the spin image. So, when the resolution changes, the number of vertices is not the same and their space partitioning is different, which leads to a difference in the set of normal vectors to be managed. All these changes can be clearly seen on the results of equations (3), (4) and (5). In Fig. 4, we show the difference between two spin maps and their two corresponding spin images of the same object bunny with different resolution.

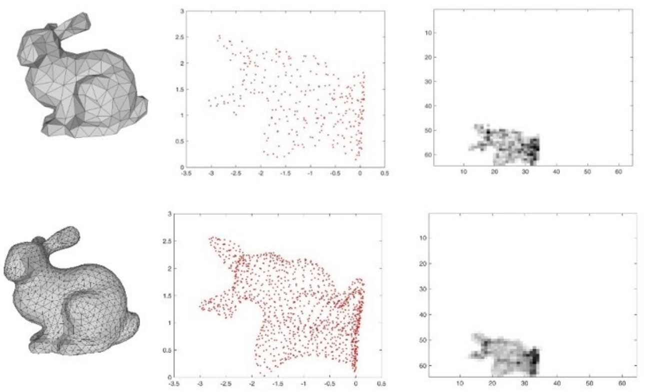


Fig. 4. Two spin maps and their two spin images of bunny with different resolution.

After visualizing results obtained of resolution changes during the extraction of the descriptor, we need to understand now how it impacts the matching phase. Let us first analyze the first step of matching algorithm: the computation of the similarity measure. In order to find for each model image the one that is most similar to it in the scene, authors have defined a similarity measure eq. (7) based on the correlation coefficient eq. (6).

$$R(P,Q)=\frac{N\sum p_i q_i-\sum p_i\sum q_i}{\sqrt{(N\sum p_i^2-(\sum p_i)^2)(N\sum q_i^2-(\sum q_i)^2)}} \tag{6}$$

$$C(P,Q)=(\arctan(R(P,Q)))^2-\lambda(\frac{1}{N-3}) \tag{7}$$

The correlation coefficient provides a measure of the intensity and direction of the linear relationship between two variables. Further, this metric is useful in measuring linear relationships. But when the relationship between two images is nonlinear, this measure may give somewhat misleading information. Since resolution changes cause weak linearity or even nonlinearity between images, the algorithm using the correlation coefficient doesn't provide good matches. We establish a correlation diagram to illustrate the impact of resolution changes on the relation between intensities of two spin images, Fig. 5 shows the results.

Besides, when we take a look at the values of intensities of model spin images and scene spin images, we can see clearly how different the ranges are, in Fig. 6 we provide an example

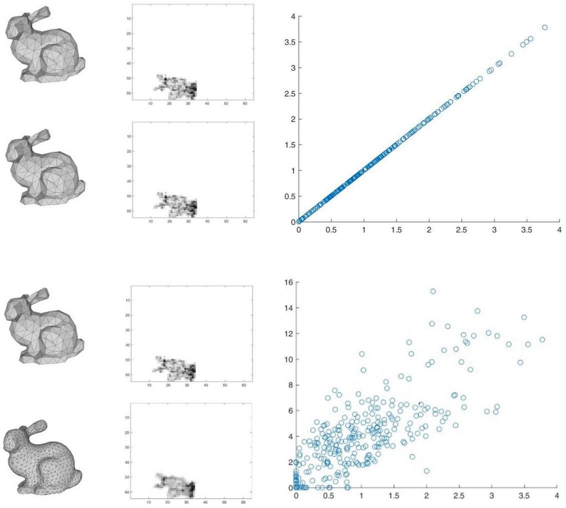
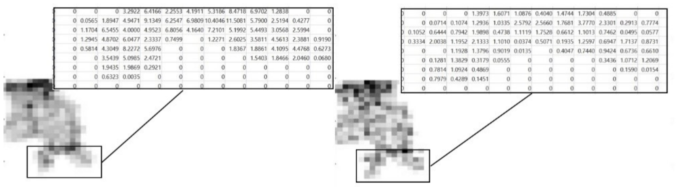


Fig. 5. An illustration of the impact of resolution changes on relation between images using the correlation diagram. Up the correlation between spin images of the same object with the same resolution is linear. Down, the difference in resolution results in a non-linearity of correlation.



0	0	0	3.2922	6.4166	2.2553	4.1911	5.3186	8.4718	6.9702	1.2838	0	0
0	0.0565	1.8947	4.9471	9.1349	6.2547	6.9809	10.4046	11.5081	5.7900	2.5194	0.4277	0
0	1.1704	6.5455	4.0000	4.9523	6.8056	4.1640	7.2101	5.1992	5.4493	3.0568	2.5994	0
0	1.2945	4.8702	6.0477	2.3337	0.7499	0	1.2271	2.6025	3.5811	4.5613	2.3881	0.9190
0	0.5814	4.3049	8.2272	5.6976	0	0	0	1.8367	1.8861	4.1095	4.4768	0.6273
0	0	3.5439	5.0985	2.4721	0	0	0	0	1.5403	1.8466	2.0460	0.0680
0	0	1.9435	1.9869	0.2921	0	0	0	0	0	0	0	0
0	0	0.6323	0.0035	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0

0	0	0	1.3973	1.6071	1.0876	0.4040	1.4744	1.7304	0.4885	0	0
0	0.0714	0.1074	1.2936	1.0335	2.5792	2.5660	1.7681	3.7770	2.3301	0.2913	0.7774
0.1052	0.6444	0.7942	1.9898	0.4738	1.1119	1.7528	0.6612	1.1013	0.7462	0.0495	0.0577
0.3334	2.0038	1.1952	2.1333	1.1010	0.0374	0.5071	0.1935	1.2597	0.6947	1.7137	0.8731
0	0	1.1928	1.3796	0.9019	0.0135	0	0.4047	0.7440	0.9424	0.6736	0.6610
0	0.1281	1.3829	0.3179	0.0555	0	0	0	0	0.3436	1.0712	1.2069
0	0.7814	1.0924	0.4869	0	0	0	0	0	0	0.1590	0.0154
0	0.7979	0.4289	0.1451	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

Fig. 6. An example of the difference between two spin images resulted from bunny under different resolution. Left: the correspondent spin image of vertex 100 from bunny with resolution 0.3 and number of vertices equal to 302. Right: the correspondent spin image of vertex 100 from bunny with resolution 0.15 and number of vertices equal to 1202.

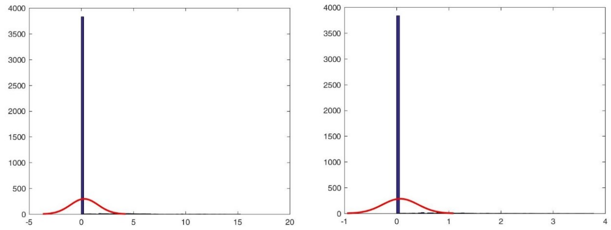


Fig. 7. The corresponding histograms of the two spin images from bunny under different resolutions. In the histogram left, range of intensity values varies between -5 and almost 20. Right, values are between -1 and almost 4.

to show the difference between intensities of two spin images that are meant to be similar.

In Fig. 7, we illustrate their corresponding histograms to show clearly the difference between ranges.

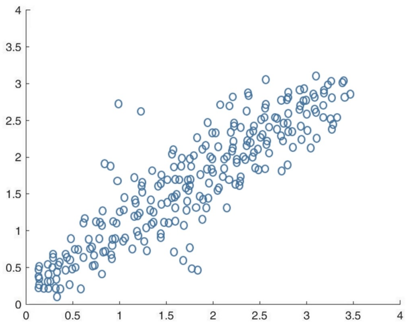


Fig. 8. correlation diagram of two spin images after normalization.

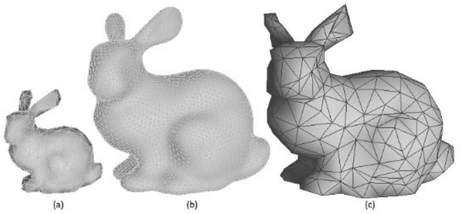


Fig. 9. Two different scenarios of scaling of bunny: (a) The original object. (b) Scaling and resolution scaling changes. (c) Scaling changes only.

As the histogram is of essential importance in terms of characterizing the global appearance of a given image one needs to represent the values of compared histograms in the same range in order perform an effective comparison. As known, the min-max normalization approach is the simplest normalization technique in which we fit the data, in a predefined range, as it is very common and usually more efficient. To normalize the data in the boundary of [A,B], the min-max normalization is defined as:

$$x_i - normalized = \frac{(x_i - min(x)) * (B - A)}{max(x) - min(x)} + A \qquad (8)$$

So, the idea here is to bring the two spin images to the same range [0,1], in order to normalize bin values for all spin images to be able to compute the correlation coefficient efficiently.

In Fig. 8, we show the impact of normalization in the correlation diagram of two spin images with different resolution after normalization.

Here we can see clearly that the two images are more correlated.

B. Invariance to Scale

One other drawback of spin image algorithm is the scaling. We have two scenarios about scaling. The first one concerns the object with the same number of vertices, but the scale is different, see Fig. 9(c) and the second one is when the same object is represented with different resolution and scale in the scene, see Fig. 9(b).

The first case is simpler. As the scaling here does not change the normal vectors of vertices, the number of vertices to project controlled by the parameter A (Angle support) is the same on each spin map. Since the image width is fixed for both spin image model and scene, to deal with changes which influences the accumulation of points in each bin of the spin image, the bin size of the scene spin image should be set to

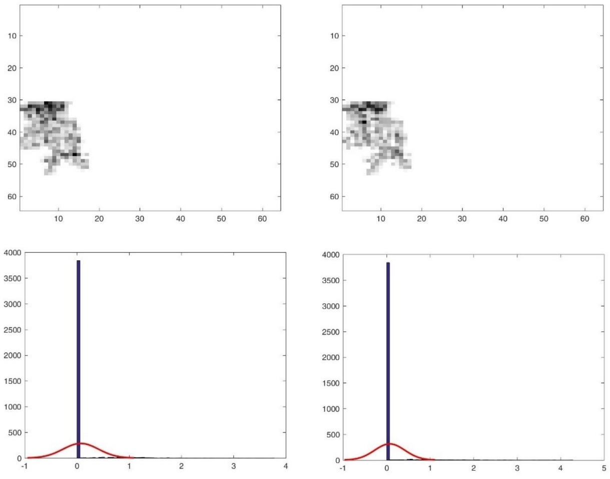


Fig. 10. Two spin images and their corresponding histograms of bunny and its scaled version with no resolution changes. On left the spin image of bunny and its histogram. On the right side the spin image of the scaled version and its histogram.

a multiple of scaling factor λ and bin size of the model spin image:

$$b_s = \lambda b_m \tag{9}$$

Which helps to get spin images with the same intensity values of spin images when we have no scaling.

In Fig. 10, we illustrate an example of a spin image of the object and its scaled version. As in practice we don't usually know the scaling factor, the bin size of the scene is determined empirically as a multiple of a multiple of the bin size of the model. This will reduce the effect of discrete location and individualization effect of points on the surface scene.

In order to show the importance of bin size for spin image matching, we experiment the effect of bin size on match distance, which is defined as the median of all distances computed between each computed correspondences during the phase of the similarity measure. A good match is established, which means correct correspondences are computed when the match distance is low. Results are shown in Fig. 11.

The second case which is more complicated is when both resolution and the scale are different. Combining the resolution changes and scaling has the same effect as changing only resolution. In that case also both the spatial distribution and the number of vertices are different. To explain with more details during the description phase different vertices on the surface mesh of the scene are falling into different bins. This is due to the difference in the number of vertices which leads to the difference of their spatial location from the ones of the model. Consequently, the normal vectors are also different, which has an impact on the choice of vertices based on the angle support α. Therefore, spin images that are meant to be similar will be dissimilar. In Fig. 12, we provide an example of this case showing the spin image of the same vertex for a bunny model with three different resolutions.

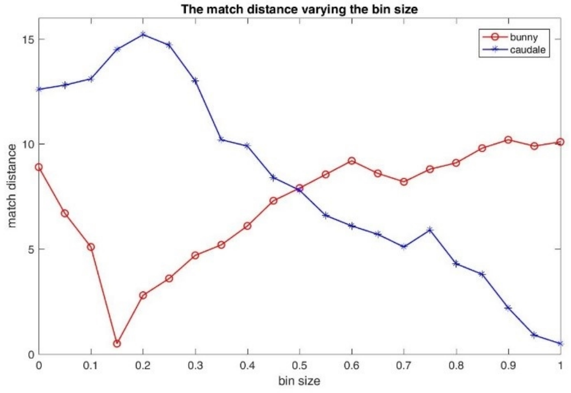


Fig. 11. The impact of varying bin size on the distance match.

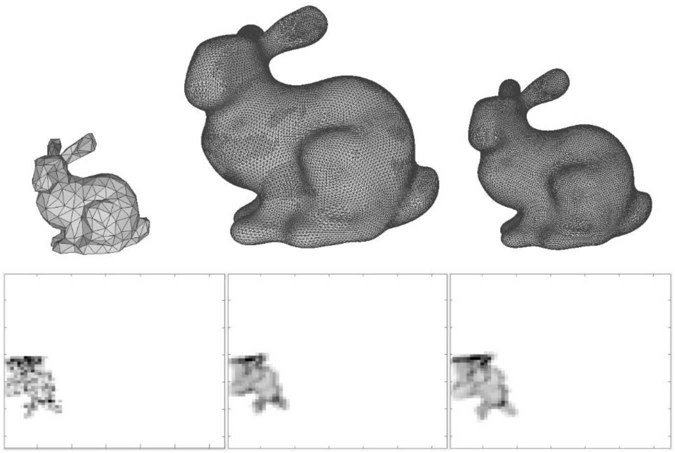


Fig. 12. Scaling and resolution changes and its impact on the generation of spin images.

To overcome this issue, we proceed in the same way as we did for resolution changes. Since the difference in image width is not handled by correlation coefficient, we start by fixing the image width for both models and scenes. Then to reduce the effect of discretization and in order to represent the shapes in spin images in the same scale level, we set the bin size of the scene to a multiple of the bin size of the model. Then we perform a normalization to bring intensity values to the same range to compute the correlation coefficient efficiently. To validate what we have explained above and the choice of bin size empirically we evaluate a plot of match correlation. We mention here that the match correlation measure is the median of the histogram of the correlation coefficient between spin-images computed for all point matches. When correlation is high, the correspondences are correctly computed. See Fig. 13.

V. Experimental Results

The current section provides an evaluation of our suggested approach SRISI in comparison with the spin image algorithm. For this purpose, we perform a wide range of tests utilizing models from two datasets: ArcheoZoo3D and Stanford's 3D scanning repository. First, Section A briefly presents our

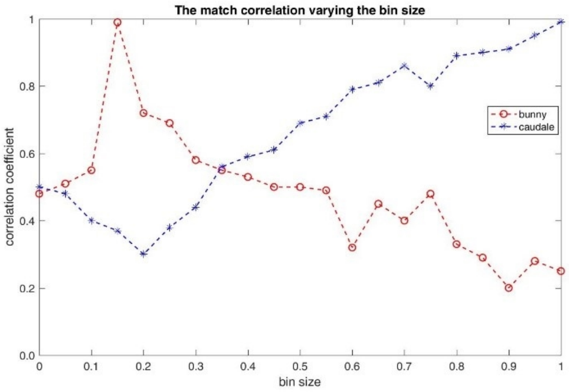


Fig. 13. The influence of varying the bin size on the match correlation for bunny and caudal.

database. Afterwards, in Section B, we provide detailed technical information on the implementation environment. Next, the experiment carried out is revealed in Section C. In the same section we measure the precision and recall to evaluate the performance of our contribution. And then we compare it to the standard algorithm with a discussion of strengths and shortcomings of our contribution.

A. Datasets

In this works we have validated our approach on two datasets. The first one is Stanford 3D scanning repository. A well known repository that provides some dense polygonal models publically. The second database is Archeozoo3D. It gathers 3D scans of horse's bones. Before recognition we have processed objects to remove all unreferenced vertices. Then we construct proper triangulated surfaces with screened Poisson surface method to remove holes. We sampled all objects to have the same resolution.

B. Implementation

In order to put the algorithm of spin images into action, we have based our implementation on the information provided in the thesis work [16]. We have implemented the whole phases of the algorithm from descriptor extraction to verification passing by the matching in Matlab. Concerning models in the two databases, they have been processed, whether for creating scenes, normalizing vectors or applying transformations, etc. with the aide of Meshlab, blender and using the "Toolbox Graph" of Peyre [1] in Matlab. About environmental information, our experiments were carried out on a computer with 2.50 GHz Intel i7 processor and 16GB of memory.

C. Results and Discussion

The purpose if this current section is to provide an evaluation of our proposed method SRISI in comparison with the original one SI. In order to provide a robust evaluation, the state of the art presents different metrics. We have chosen two of the most important ones utilized in the information retrieval domain, Precision and Recall. The mathematical formula for each one is given in equations (10) and (11).

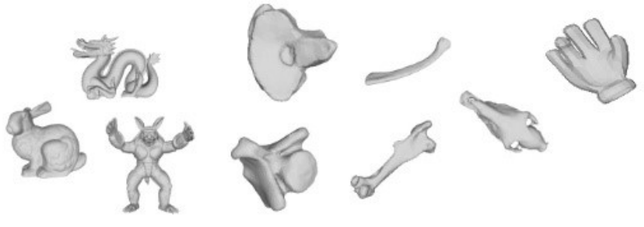


Fig. 14. Objects used to run tests. three first objects from 3D Stanford repository. Five second objects from ArcheoZoo3D database. And lastly glove model.

$$Precision = \frac{tp}{tp + fp} \tag{10}$$

$$Recall = \frac{tp}{tp + fn} \tag{11}$$

With tp: True positives is the number of times an existing object in different scenes with different conditions is correctly recognized. fp : False positives indicates the number of times a non-existing object in the scenes is mentioned to be recognized. It is to say that the algorithm finds correspondences on the scene, so the model is aligned with another object. Finally, fn: false negatives, when a model exists in the scene, but the algorithm fails to recognize them, in our case, it fails to find any correspondences. To test the validity of our approach, we used three objects from the Stanford repository, five objects from the ArcheoZoo3D database and one other object called glove modeled by Alexander Masliukivaky. The objects are listed in Fig. 14.

At first all objects have the same resolution. Resolution here refers to the median of the lengths of the edges between the vertices. The tests were done first for each isolated object. We initially change only the resolution and keep the scale fixed, then apply the transformations (translation, rotations) as well as truncating parts of the objects. We next carry out tests in reverse. We fixed the resolution and changed the scale. Lastly, we change both resolution and scale. In the second time we test the robustness of our method to occlusions and clutters. To do that, we have created 30 scenes from 4 objects of Stanford datasets, then we have changed the resolution of scenes two times, which results in 90 trials for each model. Then we have repeated the same process for Archeozoo3D datastet. So roughly, concerning SRISI we get 360 trials for Stanford and 360 for Archeozoo3D. For SI, as mentioned earlier, the algorithm does not find any correspondence. For the results presented in this work, image width is set to 64, the resolutions of models is set to 0.3, the bin size is 0.15 and the angle support equal 180. To show the effect of occlusions and clutters on our method, we will compute the recognition rate in terms of occlusions and clutters. To do this, for each scene of the 30 scenes created from the Stanford database, we run the recognition test. This will allow us to deduce the true positives, false positives and false negatives. Then we calculate

[1] http://www.mathworks.com/matlabcentral/fileexchange/5355-toolbox-graph

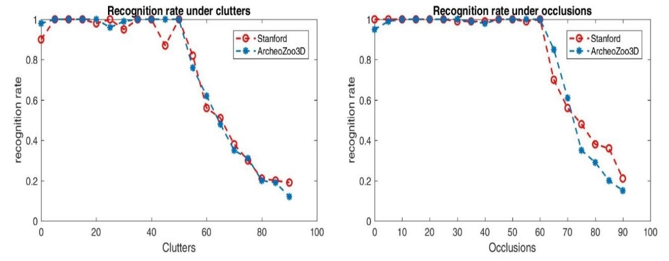


Fig. 15. Recognition rate under occlusions (left) and clutters (right) for both Stanford and Archeozoo3D.

for each test the occlusions and clutters given by the equations below.

$$\text{Occlusion} = 1 - \frac{\text{model surface match area}}{\text{total model surface area}} \quad (12)$$

$$\text{Clutter} = \frac{\text{clutter point in relevant volume}}{\text{total points in relevant volume}} \quad (13)$$

The surface area of a mesh is defined as the sum of the areas of its all faces. Clutter points are vertices in the scene surface mesh that does not belong to the model surface patch. Then, we repeat the same procedure for the thirteen scenes created from objects of Archeozoo3D database. Results for both databases are plotted bellow. See Fig. 15.

Examining the scatterplots in Fig. 15 we observe that the recognition rate is highly affected by occlusions. For both databases, from an amount of occlusions equal to 60%, the true positives rate starts to drop and in counterpart, at almost the same amount of occlusions true negatives and false positives increase. This is expressed by the failure of the algorithm to recognize object correctly in the scene for occlusions beyond 60%. In Fig. 15 (left), scatterplots show that clutters also influence the recognition rate. Up to 50% the algorithm still succeed at recognizing objects, but higher than this threshold, the recognition failures dominate. We assess the performance of our contribution in both Stanford and ArcheoZoo3D datasets by computing the precision and recall and comparing it to Spin Image algorithm and SHOT (Signature of Histograms of Orientations) [18]. In the table below we illustrate results obtained.

TABLE I. Performance of our contribution SRISI in comparison with some state-of-the-art methods.

Dataset	Method	Precision	Recall	Inv. to Scale	Inv. to Resolution
ArcheoZoo3D	SI	0.00	0.00	NO	NO
	SRISI	0.61	0.51	YES	YES
	SHOT	0.50	0.49	NO	NO
Stanford 3D	SI	0.00	0.00	NO	NO
	SRISI	0.67	0.59	YES	YES
	SHOT	0.58	0.50	NO	NO

From results in Table I, we see that our contribution achieves good results in term of precision and recall for both datasets, while the original algorithm fails to recognize objects when we change resolution and scale. The changes of scale and resolution of objects results in changes of spatial location of vertices and changes in the number of vertices, consequently, changes of normal vectors. As the creation of spin images is based on the projection of vertices on the surface mesh and also, this projection is controlled by angles between normal vector of the oriented point and normal vectors of other points, the accumulation of points in spin images becomes different between two same objects with different resolution and scale, resulting thus in a non-linearity between images so in difference intensity values. Knowing also that the correlation coefficient can only perform a good similarity between two spin images if only they have the same width and the transformation between them is linear. Defining efficiently good parameters for spin image generation, by setting the bin size of the scene to a multiple of the bin size of the model, then choosing to fix also the spin images width in order to represent shapes at the same scale level and finally bringing the intensity values to the same range by normalizing spin images before the matching algorithm, we have made the recognition of objects successful when resolution and scale change.

VI. Conclusion

In this paper we have introduced an approach robust to resolution and scale changes based on spin image algorithm. By understanding on the one hand that the issue of spin image algorithm was mainly related to the accumulation of points projected on the image, which leads to a non-linear transformation on the spin images to be compared and on the other hand the correlation coefficient will not properly calculate the similarity in this case, as well as the in-depth study of the influence of the parameters on the creation of spin images, we have succeeded to make the spin image descriptor robust to scale and resolution changes, with an occlusion rate up to 60% and 50% for clutters. We have integrated normalization into the matching pipeline and chose the right parameters by fixing the size of the spin images and setting the bin size of the scene into a multiple of the bin size of the model. Our future work aims to to pursue the field of artificial intelligence. To this end we are interested to integrate descriptor in a neural network framework to automate and improve the recognition results under different conditions.

References

[1] J. H'roura et al., "3D objects descriptors methods: Overview and trends," in Advanced Technologies for Signal and Image Processing (ATSIP), 2017 International Conference on, 2017, pp. 1–9.

[2] P. Loncomilla, J. Ruiz-del-Solar and L. Martinez, "Object recognition using local invariant features for robotic applications: A survey," Pattern Recognit., vol. 60, pp. 499–514, 2016.

[3] R. Osada, T. Funkhouser, B. Chazelle and D. Dobkin, "Shape Distributions," vol. 21, no. 4, pp. 807–832, 2002.

[4] E. Paquet, M. Rioux, O. Ontario and K. I. A. Or "Nefertiti : a Query by Content Software for Three-Dimensional Models Databases Management," pp. 345–352, 1997.

[5] A. Nouri, C. Charrier and O. Lézoray, "Multi-scale saliency of 3D colored meshes," in Image Processing (ICIP), 2015 IEEE International Conference on, 2015, pp. 2820–2824.

[6] K. Tang, P. Song and X. Chen, "3D object recognition in cluttered scenes with robust shape description and correspondence selection," IEEE Access, vol. 5, pp. 1833–1845, 2017.

[7] C. Maes, T. Fabry, J. Keustermans, D. Smeets, P. Suetens and D. Vandermeulen, "Feature detection on 3D face surfaces for pose normalisation and recognition," in Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, 2010, pp. 1–6.

[8] D. G. Lowe, “Object recognition from local scale-invariant features,” in Computer vision, 1999. The proceedings of the seventh IEEE international conference on, 1999, vol. 2, pp. 1150–1157.

[9] Y. Xiang, W. Choi, Y. Lin and S. Savarese, “Data-driven 3d voxel patterns for object category recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1903–1911.

[10] X. Li, I. Guskov and A. Arbor, “3D object recognition from range images using pyramid matching,” pp. 1–6, 2007.

[11] D. Pickup, X. Sun, P. L. Rosin and R. R. Martin, “Skeleton-based canonical forms for non-rigid 3D shape retrieval,” Comput. Vis. media, vol. 2, no. 3, pp. 231–243, 2016.

[12] O. K.-C. Au, C.-L. Tai, H.-K. Chu, D. Cohen-Or and T.-Y. Lee, “Skeleton extraction by mesh contraction,” ACM Trans. Graph., vol. 27, no. 3, p. 44, 2008.

[13] W. Jin, C. Yan, L. Ma, H. Ye and H. Wang, “Joint extended fractional Fourier transform correlator,” Opt. Commun., vol. 268, no. 1, pp. 34–37, 2006.

[14] J. Tierny, J. Vandeborre and M. Daoudi, “Analyse topologique et géométrique de maillages 3D pour l ’ extraction de squelette,” pp. 1–8, 2006.

[15] A. E. Johnson and M. Hebert, “Using spin images for efficient object recognition in cluttered 3D scenes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, no. 5, pp. 433–449, 1999.

[16] A. E. Johnson, “Spin-Images: A Representation for 3-D Surface Matching,” Carnegie Mellon University, 1997.

[17] F. Bernardini and H. Rushmeier, “for um The 3D Model Acquisition Pipeline,” vol. 21, no. 2, pp. 149–172, 2002.

[18] F. Tombari, S. Salti, L. Di Stefano, Unique signatures of histograms for local surface description, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 6313 LNCS (2010) $356 - 369doi : 10.1007/978 - 3 - 642 - 15558 - 1_26$.

A Novel Approach for Ontology-Driven Information Retrieving Chatbot for Fashion Brands

Aisha Nazir[1], Muhammad Yaseen Khan[2], Tafseer Ahmed[3], Syed Imran Jami[4], Shaukat Wasi[5]
Center for Language Computing, Department of Computer Science, Mohammad Ali Jinnah University, Karachi, Pakistan[1,2,3,4,5]
Department of Computer Science, Edwardes College, Peshawar, Pakistan[1]

***Abstract*—Chatbots or conversational agents are the most projecting and widely employed artificial assistants on online social media. These bots converse with the humans in audio, visual, or textual formats. It is quite intelligible that users are keen interested in the swift and relatedly correct information for their hunt in pursuit of desired product, such that their precious time is not wasted through surfing multiple websites and business portals. In this paper, we present a novel incremental approach for building a chatbot for fashion brands based on the semantic web. We organized a dataset of 5,000 question and answers of top-10 brands in the fashion domain, which covers the information about new arrivals, sales, packages, discounts, exchange/return policies, etc. We have also developed a dialogue interface for querying the system. The results generated against the queries are thoroughly evaluated on the criteria of time, context, history, duration, turns, significance, relevance, and fall back questions.**

***Keywords*—Artifical intelligence; semantic web; chatbots; fashion; ontology**

I. Introduction

The invention of the Internet has met almost everything in the world. It has played a significant part in showcasing and growth of many businesses in many aspects [1], [2]. In the context of the current era, online social media has made a considerable impact on the businesses [3]. In the same regard, the fashion industry, especially fashion brands that offer voguish couture and apparels, is growing at a breathtaking rate as they provide creative and versatile garments all around the globe. People prioritize these brands upon their choices and interests, which do have correlations with the events and seasons.

It is a challenging task to recommend an appropriate brand according to users' requirements and interests. For doing so, there are many conversational agents available on the official websites of clothing brands, but they deal with only one brand that resides it. What if a customer wants multiple options of the same kind of different brands at one place? Yes, online social media has answered this particular question to some extent, but surfing the Internet to acquire desired results is very time-consuming and exasperating effort. Thus, in comparison to the searching and sorting based tools, people need some promising alternatives [4].

To overcome user's requirements to complete their task with no time, it is mandatory to understand how human thinks about a particular thing, in connection with this, it is also imperative to understand that how do we make computers to do it like humans. Turing first floated this idea, "Can a machine think?" [5], [6], and in pursuit of the answer to this question, we can say, the whole field of artificial intelligence (AI) evolved. In this era, cognitive science deeply observes the human's mind and its working, which leads to outstanding success in the field of AI in the form of artificial assistant aka Chabot. Businesses started to use these chatbots to facilitate customers. Hence, the techniques and research aspects of AI chatbots have become an exciting field in the AI community. These AI assistants/chatbots have revolutionized by understanding customer queries in different languages and appropriately responding the meaningful information.

The main aim for these chatbots is to provides immediate, meaningful, informative, context-oriented responses to assist customers for the asked questions. The AI Chatbots retrieve information through different approaches. In modern-day practices, these approaches use various information repository structures like conventional (relational) and modern (NoSQL) database systems, ontologies, AIML, etc. to model querying system.

In this paper, we present an ontology-driven chatbot model that facilitates those customers who need the latest information about brands facilities such as packages, discounts, sales, prices, varieties, online shopping, home delivery services, etc. The proposed Chatbot covers all necessary and general information relevant to clothing brands like dress designs, fabric stuff, the material used in the product, accessories, and services like home delivery, return, exchange, discounts, sales, and, etc. Through our model, customers will get all type of information for their complex queries at one platform. For example, customer can ask like: "Which brand provides clutches in blue" and "What is the delivery time of Khaadi in Pakistan?", etc. In this respect, we create an ontology-based on the set of 5000 questions and answers considering the top-10 clothing brands of Pakistan, namely,

- Asim Jofa
- HSY Studio
- Al-Karam Studio
- Sana Safina
- Ethnic
- Thredz
- Gul Ahmed
- Khaadi
- J. (Junaid Jamshed)
- Nishat Linen

We are hopeful that the proposed model is adept for many other but similar domains, all over the world.

II. Background

This section presents the related information about the approaches considered for the development of chatbots.

Amongst the employed techniques, in most cases, the developers rely on IR techniques. This is good because IR based chatbots have the edge over others as they produce an informative and fluent response as they select responses from pre-generated conversation repositories. However, also, it can be a little bad because IR based methods may give blunt answers. A significant of the Semantic Web (SW) is seen in the development of computational tools and applications in the last decade. To understand the metaphor what is SW, we can think of a philosophy that integrates and links the data (technically termed as concepts) based on relationships and standard features in a web. Consequently, "Ontology is a formal, explicit description of concepts in a domain" [7]. Moreover, according to Abdul-Kader et al., in SW concepts are interconnected relationally and hierarchically by computing relations between concepts like synonyms and hyponyms [8]. This concept is introduced in computing sciences by Tim Berners-Lee in 2001 [9]. In many areas of applications, SW has been proved equally better in comparison to its counterparts, and it is notably exercised in many organizations. For example, giants in media like BBC and New York Times have developed their repository structures by linking data concepts [10]. Web companies like Google, Yahoo, Microsoft, and Facebook are connecting millions of entities based on graphs and linked-data concepts. In this regard, providers of DBMS have begun to provide native support of SW [11]. Thus, for example, we can see the SQL based ontology is used to maintain the history of the conversation.

Many researchers have created the ontology for the fashion domain, like Bollacker et al. worked out a fashion ontology which gives (fashion) advice on the basis on human features, fashion and manufacturing concepts [12]. In an approximately similar way, Vogiatzis et al. proposed a technique that recommends garments by incorporating knowledge on all aspects of fashion like material, colors, body, and facial features [13]. The authors [12], [13] have used OWL in their experiments. In the same way, Ajmani et al. [14] adopted the technique of using probabilistic multimedia ontology for creating a personalized fashion recommendation system through which the analysis on visual properties of garments has been performed according to latest fashion trends.

Al-Zubaide et al. [15] presented a query interpreter and responder "ontbot" that technically transforms ontology into a relational database query, before responding. Further, the chat is driven by natural language processing techniques (NLP) to extract keywords from the user query. Likewise, Rao et al. [6] construct a three-stage experimental system in which they take question string as a JavaScript Object Notation (JSON) and apply NLP keyword extraction techniques. These keywords were further matched with the ontology-based relational repository and ranked by using term-frequency and inverse-document-frequency (TF-IDF) technique. The answer/document with at the highest rank was presented to the customer.

Pathan et al. [16] build an e-commerce website based unobtrusive chatbot that simulates an intelligent conversation by pattern matching of customer response based on the given context. A reductionist approach was adopted to accumulate data and elicit further information from the customers who navigate through the product catalogs during dialogue. Similarly, Gupta et al. [4] have shown, the usage of dynamic end-user inputs by adopting frequently asked question (FAQs), in their system chat approach a conventional manner and intelligently overlaid hyperlinks to help customer to redirect them towards the desired results.

Augello et al. [17] have shown the exploitation of knowledge base (KB) in a twofold manner: firstly, they engineer ontology in an AIML format that is used for the creating dynamic answers as a result of inference, and in the latter part, the ontology is automatically populated offline on the basis of AIML categories. They have also maintained that they practiced ALICE for the conversation that follows pattern matching rules (employing NLP techniques) and returns dynamic answers instead of a list of links.

III. Methodology

This section presents the details of the proposed methodology. Initially, we have covered the description of the Semantic Web and the associated concepts therein, followed by step-by-step details and discussion on the proposed methodology.

Development of chatbots based on ontologies is seen as one of the promising practices in the world, where queries are answered by matching keywords in queries and retrieving appropriate responses placed on semantic representations. Whereas, IR based chatbots have an edge of producing informative and fluent responses, in a multi-turn conversation context [18], by seeking the responses from pre-generated conversation repository.

In the proposed system, we employed a novel incremental approach for domain-oriented ontology engineering. In this regard, a wide range of development tools have been utilized; such as we use Protégé [19] for ontology engineering of the domain of "clothing brands". Protégé is considered as one of the best tools for ontology engineering in the entire world, which also enables us to export ontologies in various other language formats like Resource Description Framework (RDF) schema [20], and Web Ontology Language (OWL) [21], [22]. Similarly, we used VOWL [23] and OntoGraf [24] plugins for the visualization of taxonomy, and SPARQL for querying system and data retrieval [25]. Besides it, we as lo worked on Jena [26] which is a Java-based library for the development of SW applications [27].

A. Ontology Engineering Process

Ontology engineering process spans through different phases as it is shown in Fig. 1. In the following, we are going to explain these phases one by one.

1) Dataset Preparation: Data is collected manually via different procedures, which involve survey questionnaires, and interviews with the official representatives. We also sought information available on official websites of mentioned clothing brands and used various scraping tools to get the text of posts and comments available on Facebook pages, respectively. These Facebook fan pages are the big source of extraction of information of meaningful and reliable conversation among

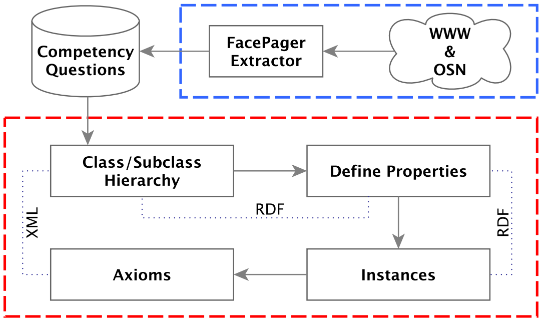


Fig. 1. The scheme of ontology engineering process (OEP). Block enclosed in green box is the initial phase of data gathering, while the block in red box defines the OEP.

users and different teams. Hence, we have employed a Google Chrome extension, namely, Scrapper[1], for scraping the information from the web-portals. The other tool used, in our course of the experiment, for extracting the text of posts and comments from Facebook pages, namely, FacebookPager[2]. The FacePager takes Facebook page key as an input and retrieves all posts, comments, pictures, videos, and user reviews available on the given page that you can export in comma-separated view format.

As a result, the collected corpus was consisting of unstructured and inconsistent data. Further, there were two more issues: it was redundant and not very much meaningful. Thus, the data is filtered and processed to make it useful as per the requirements of the SW based applications. The count of the parallel corpus, in two-way communication, is 5,000 sentences.

2) Competency Questions: After the acquisition of data, the first step towards ontology development is to lemmatize the scope of ontology through the competency questions. These are the vital questions for which an ontology has to answer. Moreover, these questions are the primary source of setting the precincts of ontology domain, and helpful to identify the terms that are further converted into the system of class and subclass hierarchy. In the proposed domain, for example, the competency questions can be: “Does Asim Jofa provide exchange/return facility?”, “Which brand provides accessories?”, or “Which brand offers 50% discount?”, etc.

3) Concepts and Classes: Classes are the basic building block of ontology, which can be interpreted as a set of specific individuals [28]. In OWL, these are also called concepts or entities having some distinctive characteristics. These classes are formed in a hierarchical system of super-class and subclass. However, these classes can be disjoint. In such case, the individuals of these classes are not common. This class and sub-class hierarchy is also known as Taxonomy [7]. Thus every super class exhibits the most general characteristics of all nested sub-classes, and in contrast, the sub-classes does opposite. For example, in our experiment, “brand/vendor” is the most general class which has nested sub-classes like “accessories”, “cloth variety”, “brand type”, “dress category”, etc. Likewise, the class “accessories” is further nested by two more classes i.e., “male accessories” and “female accessories”. While, the example of disjoint class can be “facilities” and “location/area”, these classes have different instances which do not overlap each other. The detailed class hierarchy is presented in Fig. 2.

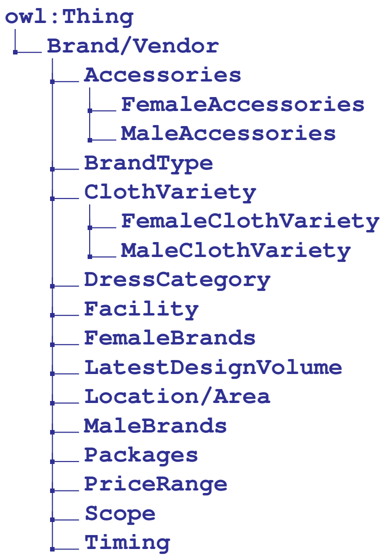


Fig. 2. Class Hierarchy of Fashion Brand Ontology

4) Properties, Attributes, or Predicates: Since these set of classes are not self-explanatory, therefore, we have to define the mappings inside/among classes [29]. The OWL properties describe the relationship between classes which can be of two kinds, namely, object properties, and data properties. Details of these properties are given below.

Object Properties. These properties are ones who establish a link between two individuals. This is also known as intrinsic properties [7]. Technically, as a rule of thumb, any property whose range is a class is an object property. Protégé provides numerous predicates that remove ambiguity from the taxonomy. Fig. 3 shows the object properties of the fashion brand ontology.

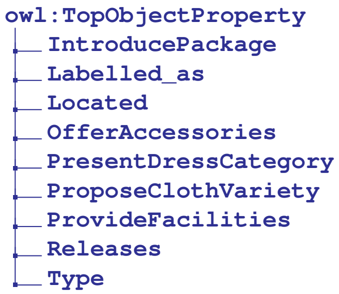
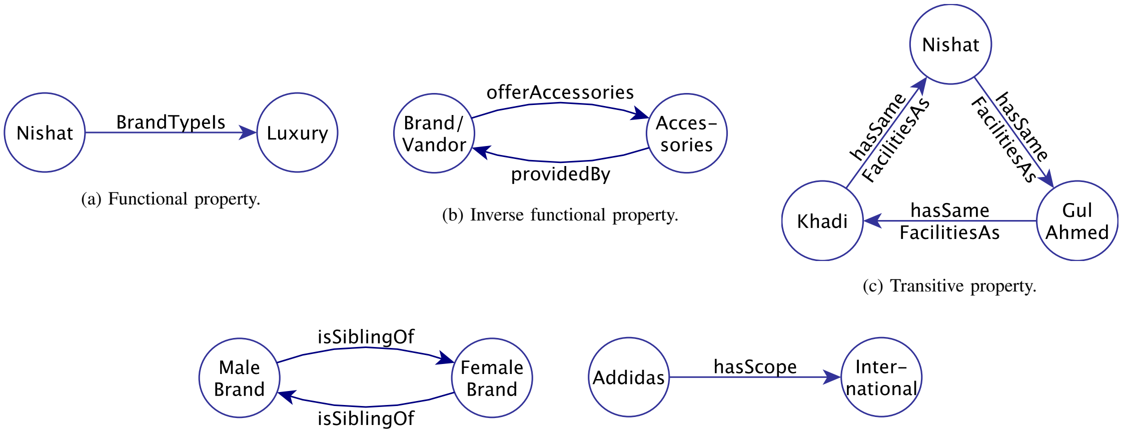


Fig. 3. Object Properties.

The core characteristics of object properties which show the global cardinality constraints on properties are: Functional, Inverse functional, Transitive, Symmetric, Asymmetric, Reflexive, and Irreflexive properties. Intuitively, *functional*

[1] https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecaccepngjd

[2] https://github.com/strohne/Facepager/wiki/About-Facepager



(a) Functional property.

(b) Inverse functional property.

(c) Transitive property.

(d) Symmetric property.

(e) Asymmetric property.

Fig. 4. One of many examples of object properties in proposed methodology.

property is a property which postulates that for any given individual there must be at most one out going relationship [22], [28], [30]; *inverse property* asserts that for any individual there should be at most one incoming relationship, through the property, which can uniquely identify the subject [30], see Fig. 4b, where a brand/vandor offers accessorises, which can identify the provider inversely. However, if there are many things related to one individual, through the functional or inverse-functional property, then the property characteristic will be inconsistent [22], [28]. The *transitive property* can be defined as the property which shows transitive implications among individuals, such that if an individual a is similar to individual b, and b is similar to individual c than we can say the individual a and c are also similar, through a transitive relation [30]–[33]. Fig. 4c depicts an example w.r.t to the current research. The *symmetric property* asserts that a given individual has itself an inverse function; whereas *asymmetric property* lacks this characteristic [30], [32]. We can see, as an example of symmetric property, if individual a and individual b are related to each via some property, then b should be related to a through the same property; while in the same setting, for the asymmetric case, b does not relate to a along the same property. Fig. 4d and 4e show the examples of symmetric and asymmetric properties respectively. Lastly, the *reflexive property* relates everything to itself, whereas, the *irreflexive property* means no individual can be related to itself by some role [22], [32]. Few examples of these object properties are illustrated in Fig. 4, and detailed mapping of these properties is shown in Table I.

TABLE I. EXAMPLES OF OBJECT PROPERTIES

Property	Name	Example
Symmetric	IsSibblingOf	(maleBrand, femaleBrand)
		(maleAccessories, femaleAccessories
		(maleClothingVariety, femaleClothingVariety)
Transitive	SamaFacilitiesAs	(Khaddi,Nishat,GulAhmed)
	SamaClothVarietyAs	(Khaddi,Nishat,GulAhmed)
	SamePriceRangeAs	(Nishat,JunaidJamshed,GulAhmed)
	SamePacksgesAs	(Thredz,Levise,Nishat)
	SameAccessoryAs	(Bonanza,GulAhmed,Nishat)
	SameClothingStuff	(Nishat,Khaddi,Bonanza)
	SameDressCategoryAs	(Bonanaza,GulAhmed,Khaadi)

Data Properties. We can briefly define a data property as a property that relates individuals to data-type values [33], in other words, any property whose range is any literal or data-type value is known as a data property. Extrinsic properties: like name, has string data-type. Table II shows the details of data properties with domain and ranges accordingly.

TABLE II. EXAMPLES OF DATA PROPERTIES

Property	Name	Domain	Range
Functional	Is_a	Levise	maleBrand
	Is_a	Bareeze	femaleBrand
	BrandTypeIs	Nishat	Luxury
	HasStuffQuality	Khaddi	Moderate
	AccessoryProvidedBy	Scarfs	JunaidJamshaid
	AccessoryProvidedBy	Belt	Levise
	OfferDressCategor	SanaSafina	Bridal
	LabledAs	Khaddi	International
Asymmetric	HasScope	Adidas	International
	IsTypeOf	Bareeze	Luxury
	LocatedAt	Khaddi	Saddar
	OfferedDiscount	Bonanza	float
Inverse	OfferAccessories	Brand/Vendor	Accessories
	ProvidedBy	Accessories	Brand/Vendor
	LabledAs	Brand/Vendor	Scope
	IsScopeOf	Scope	Brand/Vendor
	PresesntDressCategory	Brand/Vendor	DressCategory
	DressCategoryOfferedBy	DressCategory	Brand/Vendor
	ProposeClothVariety	Brand/Vendor	ClothVariety
	ClothVarietyOfferedBy	ClothVariety	Brand/Vendor

5) Instances: An instance is an individual/object that certainly belongs to a class. One key feature of OWL ontology is: it does not use the unique name assumption (UNA), so we can explicitly define that two individuals are the same or different. A class may have multiple instances. We can manually define the characteristics of each instance separately. For example, class Brand has instances like Khaddi, AsimJofa, Nishat, Al-Karam, and many others.

6) Axioms: After building class taxonomy and establishing links among classes and individuals: the following step is carried out to the semantics unambiguous. It is done so to ensure the validation and consistency of ontology; and as a procedure, we convert hierarchy into first-order logic, hence forming the "axioms" that is represented as $\langle \mathbf{C}, \mathbf{R}, \mathbf{I}, \mathbf{A} \rangle$, where $\mathbf{C}$ represents classes, $\mathbf{R}$ represents relations therein, $\mathbf{I}$ shows

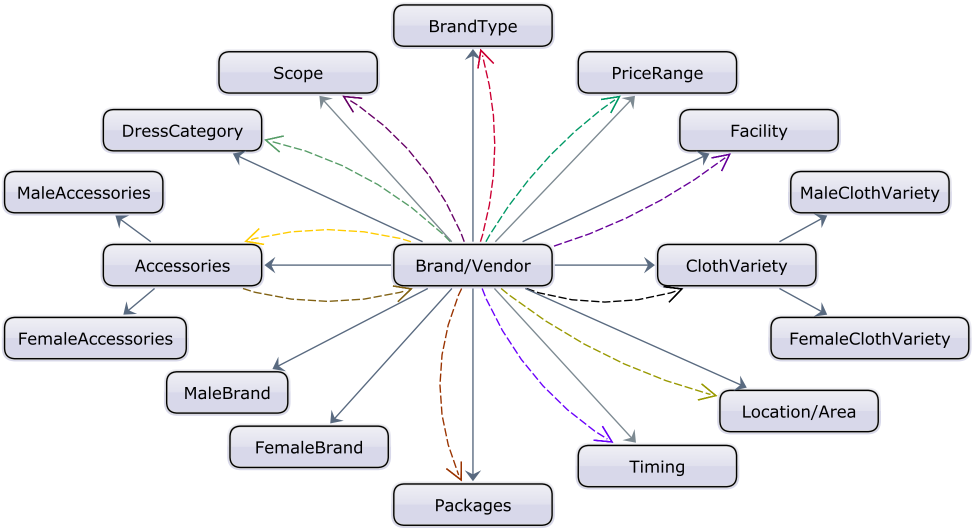
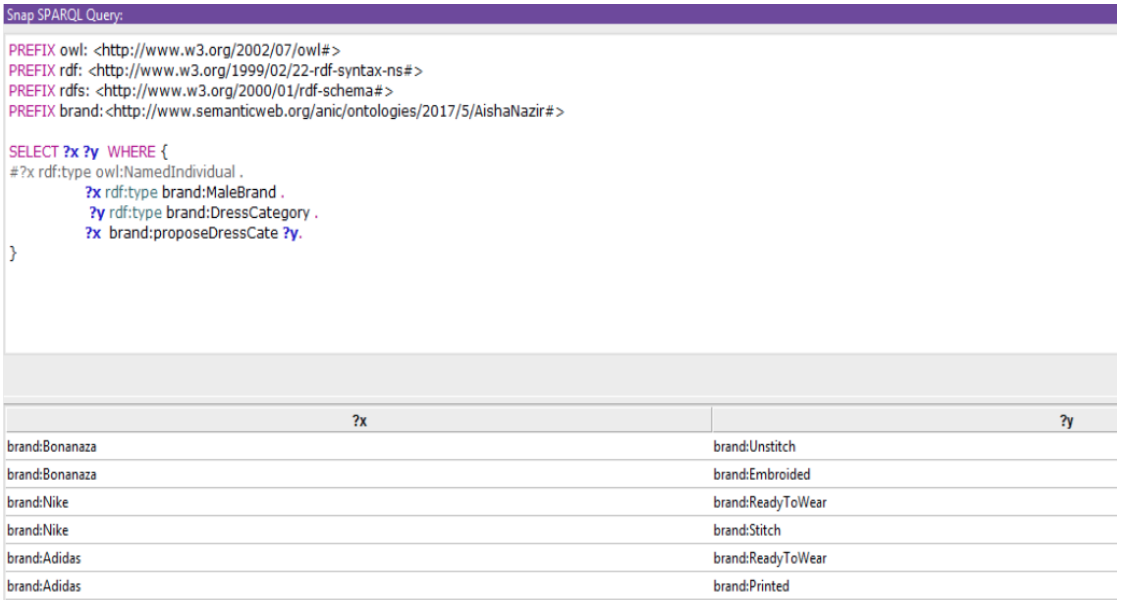


Fig. 5. The OntoGraph representation of fashion brand ontology.



```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX brand:<http://www.semanticweb.org/anic/ontologies/2017/5/AishaNazir#>

SELECT ?x ?y  WHERE {
#?x rdf:type owl:NamedIndividual .
            ?x rdf:type brand:MaleBrand .
             ?y rdf:type brand:DressCategory .
            ?x  brand:proposeDressCate ?y.
}
```

?x	?y
brand:Bonanaza	brand:Unstitch
brand:Bonanaza	brand:Embroided
brand:Nike	brand:ReadyToWear
brand:Nike	brand:Stitch
brand:Adidas	brand:ReadyToWear
brand:Adidas	brand:Printed

Fig. 6. SPARQL query in Protégé.

their instances, and **A** shows axiom [26]. Protégé provides 'Reasoner' support to formulate and manipulate logical formulae [33]. Through Reasoner, we can check the consistency of ontology; as well as add inference to semantic web application [8]. Further, in order to visualize ontology with its conceptualization, graphical representation was generated using OntoGraf and VOWL plugins [29]. Fig. 5 shows the visualization of the proposed ontology.

On completing this phase, we are done with the ontology engineering process; thus, in next sections, ontology integration and chatbot designing phases are discussed in detail.

B. Rule Engineering

We worked out different scenarios based rules at the backend of the chatbot. These rules are defined to make the conversation in a flow, and the system to be more efficient to produce context-oriented responses. In this regard, IR approaches are commonly practiced, in connection with the combination of rules. Thus, based on rules, a history-oriented and well-aligned conversation leads to generate more accurate and logical responses. Basic programming structures like conditional statements and repeating structures are employed.

C. Integration and Querying with SPARQL

In this phase, we deploy ontology in an environment where it can easily retrieve data by establishing a connection between the chatbot interface and itself. "Jena"[3] is a Java-based library, specifically use to support semantic web-based applications [26]. Semantic query language (SPARQL) is used to manipulate semantic web repository [25]. SPARQL operate on the triple store, and itself entails triple pattern. Protégé

[3] https://jena.apache.org/

TABLE III. RESULT OF CONVERSATION EVALUATION

	Accuracy	Relatedness	Dynamic	Non-sequitur	structured	Context-oriented	Follow-ups
Expert 1	3	3	2	2.5	3	2	1
Expert 2	4	2	2	3.5	4	3.5	0
Expert 3	3.5	3	3	3	4	2	0
Average	3.50	2.67	2.33	3.00	3.67	2.50	0.33
%age	0.70	0.53	0.47	0.60	0.73	0.50	0.07

has built-in tool to check ontology accuracy and consistency named as "Snap SPARQL Query"[4]. Fig. 6 shows the sample of SPARQL query and resultant table.

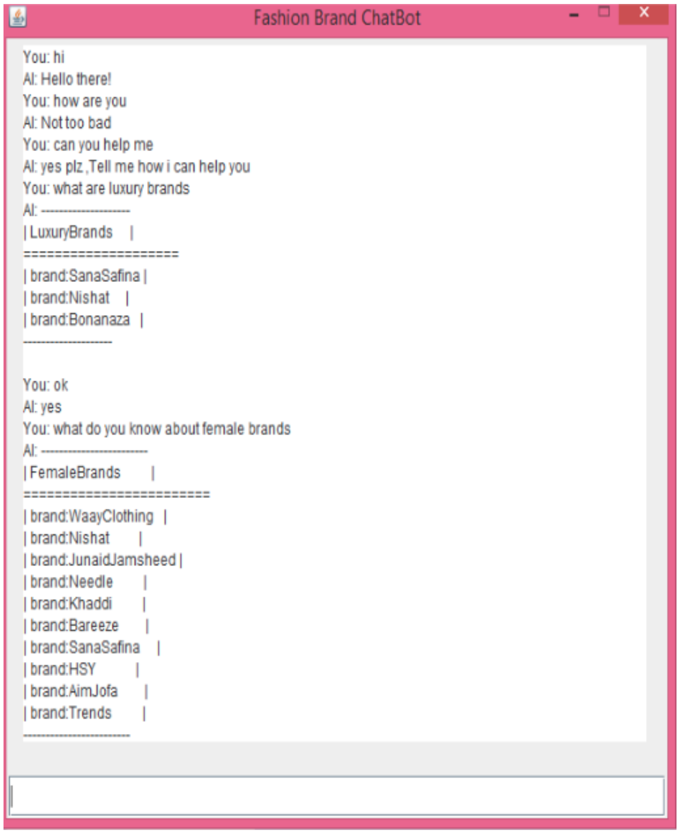
Further, in order to access the active ontology in Protégé and get results on the console, we have to set prefixes of OWL, RDF, and current ontology access path. Properties play a vital role while retrieving data from ontology. We access data based on object properties and data properties which are part of RDF Schema and retrieve results from RDF which are saved with ".owl" extension.

IV. RESULTS AND EVALUATION

A variety of techniques are available to evaluate system performance like BiLingual Evaluation Understudy (BLEU) [34], METEOR [35], and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [36]; but these metrics are often accounted as either weak or not up to the mark as compared with human judgment because there can be a lot of possible responses to any given turn [37], [38].

An insight to the conversation between chatbot and expert is shown in Fig. 7. However, our chatbot is thoroughly evaluated by using Turing's Loebner technique. According to this technique, we test chatbot by holding a conversation with it for 10 minutes. We assign performance grades based on how accurate, related, dynamic, non-sequitur, structured, history-oriented responses, retrieved by bot. The scale for judgment is from 0 to 4. For the experiment, we worked out the following criteria for the evaluation of chatbot constructed with the given methodology:

- *Accuracy*. How much accurate the results are.
- *Relatedness*. How much the responses are related to the query
- *Dynamic*. How much different results are produced if the customer asks the same question repeatedly.
- *Non-Sequitur*. How much responses are logical and reasonable.
- *Structured*. How much responses are grammatically structured and well-aligned according to sentence formation.
- *Context-Oriented*. How many results are history-based and give results in the same context.
- *Follow-up question*. Does bot ask in the response of customer, if some question is ambiguous or what a bot do to make conversation continue.



```
You: hi
AI: Hello there!
You: how are you
AI: Not too bad
You: can you help me
AI: yes plz ,Tell me how i can help you
You: what are luxury brands
AI: --------------------
| LuxuryBrands    |
====================
| brand:SanaSafina |
| brand:Nishat    |
| brand:Bonanaza  |
--------------------

You: ok
AI: yes
You: what do you know about female brands
AI: ------------------------
| FemaleBrands       |
========================
| brand:WaayClothing  |
| brand:Nishat        |
| brand:JunaidJamsheed |
| brand:Needle        |
| brand:Khaddi        |
| brand:Bareeze       |
| brand:SanaSafina    |
| brand:HSY           |
| brand:AimJofa       |
| brand:Trends        |
------------------------
```

Fig. 7. Chatbot replying the queries of customer utilizing ontology.

With the criteria discussed above, we have three experts to judge the system. The quantified results are compiled to get average and percentage-wise analyses.

We can see that the proposed methodology is useful as it produced meaningful results. For example, the performance of the system under the criteria of accuracy, non-sequitur, and structured answers show the better results by yielding .70, .60, and .73 % marks on average by the experts. However, the system shows a poor performance on the follow-ups; followed by failing to (behave dynamically) produce different answers, in more turns. It may be due to the lack of natural language generation within. The relatedness and context-orientation of the responses are a little above average. Although much considerable work can be done to improve this. The detailed results are presented in Table III.

V. CONCLUSION

In Pakistan, clothing brands lack instant AI assistants at their official websites and social web page, which is seen as a

[4] https://github.com/protegeproject/snap-sparql-query

core facility provided by international brands. Several tussles are required to make a well organized artificial bot to produce fast results. The proposed system resolves the problem for Pakistani fashion industry through developing clothing brand ontology, yielded through the handcrafted dataset of 5000 pairs of questions/answers, and integrating it with a conversation agent to facilitate online customers. In our work, we focus only on general-purpose information like brand facilities, services, garments, clothing stuff, and accessories based on information retrieved from Facebook pages and official websites.

VI. Limitations and Future Work

This research work is limited to only ten clothing brands and provides concern areas information to customers; thus, in the future, the scope of brand ontology can be increased by adding more national brands. We also intend to implement Semantic Web Rule Language (SWRL) and employing deep learning architectures.

References

[1] B. Gates, "Business@ the speed of thought," *Business Strategy Review*, vol. 10, no. 2, pp. 11–18, 1999.

[2] G. J. Avlonitis and D. A. Karayanni, "The impact of internet use on business-to-business marketing: examples from american and european companies," *Industrial Marketing Management*, vol. 29, no. 5, pp. 441–459, 2000.

[3] S. Aral, C. Dellarocas, and D. Godes, "Introduction to the special issue—social media and business transformation: a framework for research," *Information Systems Research*, vol. 24, no. 1, pp. 3–13, 2013.

[4] S. Gupta, D. Borkar, C. De Mello, and S. Patil, "An e-commerce website based chatbot," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 2, pp. 1483–1485, 2015.

[5] A. Turing, "Can machine think?"

[6] L. S. C. Rao, D. Kini, K. S, and K. K. N, "Chatbot-a java based intelligent conversational agent," *International Research Journal of Engineering and Technology*, vol. 4, no. 4, pp. 3575–3578, 2017.

[7] N. F. Noy, D. L. McGuinness *et al.*, "Ontology development 101: A guide to creating your first ontology," 2001.

[8] S. A. Abdul-Kader and J. Woods, "Survey on chatbot design techniques in speech conversation systems," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 7, 2015.

[9] T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[10] C. Bizer, "The emerging web of linked data," *IEEE intelligent systems*, vol. 24, no. 5, pp. 87–92, 2009.

[11] A. Bernstein, J. Hendler, and N. Noy, "A new look of the semantic web," 2016.

[12] K. Bollacker, N. Díaz-Rodríguez, and X. Li, "Beyond clothing ontologies: modeling fashion with subjective influence networks," in *KDD workshop on machine learning meets fashion*, 2016.

[13] D. Vogiatzis, D. Pierrakos, G. Paliouras, S. Jenkyn-Jones, and B. Possen, "Expert and community based style advice," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 647–10 655, 2012.

[14] S. Ajmani, H. Ghosh, A. Mallik, and S. Chaudhury, "An ontology based personalized garment recommendation system," in *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 03*. IEEE Computer Society, 2013, pp. 17–20.

[15] H. Al-Zubaide and A. A. Issa, "Ontbot: Ontology based chatbot," in *International Symposium on Innovations in Information and Communications Technology*. IEEE, 2011, pp. 7–12.

[16] G. S. Pathan, P. M. Bante, R. V. Dhole, and S. Kurzadkar, "An e-commerce web application based chatbot," in *International Journal for Research in Applied Science & Engineering Technology*, 2018, pp. 1263–1267.

[17] A. Augello, G. Pilato, A. Machi, and S. Gaglio, "An approach to enhance chatbot semantic power and maintainability: experiences within the frasi project," in *2012 IEEE Sixth International Conference on Semantic Computing*. IEEE, 2012, pp. 186–193.

[18] Y. Wu, W. Wu, C. Xing, M. Zhou, and Z. Li, "Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots," *arXiv preprint arXiv:1612.01627*, 2016.

[19] J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubézy, H. Eriksson, N. F. Noy, and S. W. Tu, "The evolution of protégé: an environment for knowledge-based systems development," *International Journal of Human-computer studies*, vol. 58, no. 1, pp. 89–123, 2003.

[20] G. Klyne and J. J. Carroll, "Resource description framework (rdf): Concepts and abstract syntax," 2006.

[21] G. Antoniou and F. Van Harmelen, "Web ontology language: Owl," in *Handbook on ontologies*. Springer, 2004, pp. 67–92.

[22] D. L. McGuinness, F. Van Harmelen *et al.*, "Owl web ontology language overview," *W3C recommendation*, vol. 10, no. 10, p. 2004, 2004.

[23] S. Lohmann, S. Negru, F. Haag, and T. Ertl, "Visualizing ontologies with vowl," *Semantic Web*, vol. 7, no. 4, pp. 399–419, 2016.

[24] S. Falconer, "Ontograf," *Protégé Wiki*, 2010.

[25] T. Segaran, C. Evans, and J. Taylor, *Programming the Semantic Web: Build Flexible Applications with Graph Data*. " O'Reilly Media, Inc.", 2009.

[26] B. McBride, "Jena: A semantic web toolkit," *IEEE Internet computing*, vol. 6, no. 6, pp. 55–59, 2002.

[27] J. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson, "Jena: implementing the semantic web recommendations," in *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. ACM, 2004, pp. 74–83.

[28] M. Horridge, S. Jupp, G. Moulton, A. Rector, R. Stevens, and C. Wroe, "A practical guide to building owl ontologies using protégé 4 and co-ode tools edition1. 2," *The university of Manchester*, vol. 107, 2009.

[29] K. Hadjar, "University ontology: A case study at ahlia university," in *Semantic Web*. Springer, 2016, pp. 173–183.

[30] T. B. of Trustees of the Leland Stanford Junior University. Protégé 5 documentation. [Online]. Available: http://protegeproject.github.io/protege/

[31] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "Owl 2 web ontology language primer," *W3C recommendation*, vol. 27, no. 1, p. 123, 2009.

[32] W. W. W. Consortium *et al.*, "Owl 2 web ontology language document overview," 2012.

[33] A. Grigoris and H. F. Van, "A semantic web primer a semantic web primer second edition," 2008.

[34] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[35] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[36] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[37] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *Acm Sigkdd Explorations Newsletter*, vol. 19, no. 2, pp. 25–35, 2017.

[38] J. H. Martin and D. Jurafsky, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall Upper Saddle River, 2009.

Multi-Sessions Mechanism for Decentralized Cash on Delivery System

Nghia Duong-Trung[1], Xuan Son Ha[2], Tan Tai Phan[3], Phuong Nam Trieu[4],
Quoc Nghiep Nguyen[5], Duy Pham[6], Thai Tam Huynh[7], Hai Trieu Le[8]
Can Tho University of Technology Can Tho city, Vietnam[1,2,4,5,8]
FPT University, Can Tho city, Vietnam[1,2]
National Chiao Tung University, Taiwan[3]
Hanoi University Science Technology, Ha Noi, Vietnam[6]
Transaction Technologies Pte. Ltd., Singapore[7]

***Abstract*—To date, cash on delivery (COD) is one of the most popular payment methods in developing countries thanks to the blossom of customer-to-customer e-commerce. With the widespread of a very small business model and the Internet, online shopping has become part of people's daily activity. People browse for desirable products at the comfort of their homes and ask the online vendor that a shipper can deliver the merchandise at their doorstep. Then, COD allows customers to pay in cash when the product is delivered to their desired location. Since customers receive goods before making a payment, COD is, therefore, considered as a payment system. However, the crucial issue that previous research has not yet addressed is that their models only support single delivering session at a time. More precisely, if the current buyer is not available to receive the goods, the shipper has to wastefully wait for the complete payment and he/she cannot start shipping another merchandise. The tracking system seems to poorly handle this issue. In particular, we propose a multi-session mechanism, which consists of blockchain technology, smart contracts and hyperledger fabric platform to achieve distributed and transparent across delivering sessions in the decentralized markets. Our proposed mechanism ensure the efficiency of delivering process. The authors release our sources codes for further reproducibility and development. We conclude that the integration of multi-session mechanism and blockchain technology will cause significant efficiency across several disciplines.**

***Keywords*—Blockchain; cash on delivery; multi-sessions; decentralized system**

I. Introduction

With the adoption of modern technology and the Internet, selling products online has become a very active market in developing countries. There is an immense need to have a delivery solution of any physical items ranging from catering, beverages, clothing and home facilities. Meeting the needs of today's customer-to-customer e-commerce, many third parties have launched delivery services which utilize blockchain technology. It seems counter-intuitive that, in developing countries like Vietnam, credit card and online payment are not widely used in the market. People still prefer to pay in cash because they want to make sure that the products must be in perfect condition.

Cash on Delivery (COD) allows customers to pay in cash when the products are delivered to their home or a location they choose. This is sometimes called a payment system because customers receive goods before making a payment. COD has become increasingly popular in recent years and been considered one of the main payment methods in many countries [1], [2], [3]. Among research articles, most investigated payment methods is in general, rather than focusing on COD in particular. Transfer agents are often used as postal services, but usually, consumer and business shipments will be sent to COD by courier companies, commercial truck forwarders or organizations own delivery services. COD sales usually involve a delivered fee charged by the shipping agents and is usually paid by the buyer. In retail and wholesale transactions, shipments rely on COD-based payment method when the buyer does not have a credit account and the seller does not choose a payment method in advance. COD postal services [4] were first introduced in Switzerland in 1849, India and Australia in 1877, the United States in 1913, Canada in 1922 and the United Kingdom in 1926. Particularly in Vietnam, COD is accepted by almost online vendors and customers.

However, the crucial issue that previous research has not yet addressed is that their models only support single delivering session at a time. During a working day, a shipper can take as many orders from customers across the local area. Then, the shipper delivery the products sequentially. If the current buyer is not available to receive the goods, the shipper has to wastefully wait for the complete payment and he/she cannot start shipping another merchandise. The tracking system seems to ineffectively track this issue. Consequently, a mechanism is missing in the buyer and seller's dilemma [5]. Addition to the current seven core components, e.g. product delivery, product payment, delivery trust, payment trust, escrow account, legal document and reputation system, the authors introduce a new part that can be integrated into the dilemma. To the best of our knowledge, this novel idea is firstly investigated and implemented by the authors.

The rest of the paper is organized as follows. The authors present related research in the field of COD and decentralized system in Section II. Then in Section III, the authors summarize the most important technical background for comprehending the proposed mechanism. The core contribution of the paper is presented in Section IV. Next, several real-world scenarios and remarks are demonstrated in Section V. Finally, the authors make conclusion in Section VII.

II. Related Work

Hanan and Salah have mentioned some limitations of proof of delivery (POD) process which relies on a trusted third party to implement the process [6]. Therefore, a new POD process using Blockchain technology from Ethereum has been introduced with the number of transportations through several intermediaries by their research. Besides, a dual deposit mortgage mechanism is used for parties to comply with the contract. The development of current e-commerce and the important role of Blockchain technology has been dedicated in Ha et al. research [7]. Besides, the limitation of the traditional CoD model is mentioned in detail such as trusted in the third party, order management, and the payment process between the parties in the system. So that, Blockchain technology using Hyperleder and smart contract is built to solve the issues of COD.

Camp et al. [8] had provided a digital purchasing method with a digital token on the network. They offer to issue invoices signed by sellers and digital goods which have been encrypted and sent from a seller to a customer. The participants such as sellers and buyers will be anonymous, making transactions through commitments by signing confirmation. There are no legal or property constraints. Le et al. [9] has mentioned the important roles of blockchain, especially, the decentralized users model, to builds the transportation process and offers mechanisms to promote and ensure the interests of participating parties. The benefits of the seller are enhanced and penalized the shippers who deliberately cheated. Therefore, the real data has applied to the system so that the delivery of multiple senders suitable for their Blockchain system is transparent. The process is built by all cash payments.

Altawy et al. [10] have compared the differences between buying handicrafts using cash payment and buying goods via the Internet. Online shipping needs more trust and the information of the parties who join the system to perform several actions such as making payment, delivery, and making sure the right items. Besides, the types of e-commerce used in trading which help the process faster. Anonymously purchase of a buyer is a big concern so that the Lelantos system has built based on Blockchain to detect and cancel the anonymous purchases which affect on the trading process. Discussing trading on a digital platform with the trust of the participants, Asgaonkar and Krishnamachari [11] have issued a deposit protocol for trading by the participants. They applied the trust of the Blockchain system to make payment by participants without a trusted third party. This protocol asked dual-deposit amount on the contract with payment of both sides and the price of the product is always fixed. However, the product for the parties conducting the transaction has not been verified.

Halaweh presented the rapid growth of the COD model as an important method in making payment and transport in e-commerce [12]. The author gave statistics on the study of the COD process to customers and conducted the prediction and testing of factors affecting the COD process by using a questionnaire methodology. Moreover, it also predicts the factors that affect the COD process such as safety and security of the system for the products and privacy of the participants. Barkhordari et al. have proposed a concern of the bank using the Internet to negotiate and solve customers' needs [13]. That is the trust of customers and the security of the system. Their article has deployed a survey regarding influencing factors to payment transactions. That surveys emphasize two factors, e.g. trust and security. Similar to the above survey, the payment and transportation in COD need the trust of the participants and the security of the system.

OpenBazaar platform [14] provides a procedure for making deposits when agreed by the buyer, seller, and trader which is known as multi-signature escrow. A third party will participate in the process of trading an item called a moderator. The moderator will resolve disputes when a problem occurs. Bitcoin currency will be used for payment in transactions. The process has not yet delivered the person, the role of the deliverer is not specified. Besides, the need for a third party to resolve the dispute will consume more assets and time of the parties involved. This has been solved by using the smart contract as a third party as well as solving the problems of the parties involved in the contract terms.

COD model using two smart contracts is introduced by Le et al. [15]. An outcome is a positive deployment process of decentralized applications, which enforces contracts with the exact terms. The price of the order is deposited by the participants. However, the management of orders becomes more difficult when the data and the number of orders scale up. According to the process, the second contract will be implemented immediately after the first contract is executed. The implementation is based on the memorization of the address of each contract. As a result, it is a major limitation in the deployment process if the process is applied to multiple orders since the system could not perform several contracts at the same time.

III. Materials and Technical Background

A. Cash On Delivery

COD is a service of collecting money collected in the group of services of buying goods by post (Cash On Delivery or Collect On Delivery). It is the association between postal service and money transfer service with several stages: First, the shipper proceeds to send the goods to the recipient via a courier company. Next, the courier company will send the goods to the recipient by delivery service. The branch or post office of the delivery company delivers the goods to the recipient and the receiver makes payment. After that, the branch or delivery office issued a COD check (similar to a money order) sent to the shipper. From this point on, COD will be similar to a money transfer service.

B. Blockchain Technology

Blockchain is a list of continuously written logs, called blocks, linked by encryption. Each block contains the previous block's cryptographic hash function, timestamp, and transaction data. Each block has a block header and a body containing data and hash values of the previous block. The hash value is the result of a hash function. The hash function transforms data of any length into a fixed-length string or numeric value, such as 256 bits (32 bytes) with SHA256. Blockchain is a technology that allows secure data transmission based on an extremely complex encryption system, similar to accounting books of a company where cash is closely monitored. In this case, the blockchain is an accounting ledger [16] that works in the digital field. A special feature of blockchain is that transactions are done at a high level of trust without disclosing information. All types of business and management can participate in the network and use the properties of the Blockchain system to ensure transparency of stakeholders.

C. Ethereum

Ethereum [17], [16], [18] is a distributed, public, and open-source computing platform based on blockchain technology. It features smart contracts (scenarios), facilitating online contract agreements. This platform includes a complete Turing virtual machine and Ethereum Virtual Machine [19], [20], which can execute scripts using an Ethereum computer network. Ethereum also provides a cryptocurrency called Ether, which can be transferred between accounts and used to pay peaches to help perform calculations. Gas is an internal transaction pricing mechanism, used to minimize spam and allocate resources on the network. When creating, each transaction is charged by a certain amount of gas, its purpose is to limit the amount of work needed to execute the transaction and pay for this execution at the same time.

D. Smart Contracts

A cryptocurrency is a decentralized platform that a distributed ledger is used to interact with virtual money. A contract is an instance of a computer program that executes on the Blockchain. Users transfer money by publishing transactions and interacting with contracts in the cryptocurrency network where information is propagated, data is stored among miners or network's nodes. An underlying cryptocurrency system supports the utilization of smart contracts. A smart contract contains program code, a stored file and an account balance. Any user can submit a transaction to an append-able-only log. When the contracted is created, its program code cannot be changed. An append-able-only log, called a blockchain, which imposes a partial or total arrangement on submitted transactions is the main interface provided by the cryptocurrency. The integration of smart contract in COD has been discussed in [15].

E. Decentralized Applications – dApps

DApps [21], [22], [23] are as similar as normal applications except that they are completely decentralized. It is also controlled by nodes running Ethereum networks. These dApps do not depend on any central server or third party for operating, and therefore, without the central point of failure. Thanks to the blockchain technology, the database is encrypted and stored in a decentralized fashion. By using a modern mean of communication protocols, participants can store and retrieve data without the risk of censor and intervention [24]. DApps are expected to resist attack and censorship while being able to operate in a fully autonomous model.

IV. Proposed Multi-Sessions COD Process

A. Abstract Model for a COD system

The authors start this session by presenting a general description of multi sections in COD transport process. The abstract model for COD system is illustrated in Fig. 1. First, the product information is uploaded to the sale contract by the seller where the buyer can verify through the app, and send the purchase request. The sale contract will trigger the purchase contract for activating the term that the buyer has to transfer the amount of mortgage money as same as the valuation of the order to the purchase contract. Thus, a delivery request is sent to the system by the shipper after seeing an available order. The sale contract will trigger to delivery contract for activating the terms that the shipper and seller have to transfer mortgage money and delivery fee, respectively. Finally, the money will be transferred to the seller from the purchase contract. The mortgage money and delivery fee will be transferred to the shipper by the delivery contract which used to store the delivery fee and order money from the seller and shipper respectively.

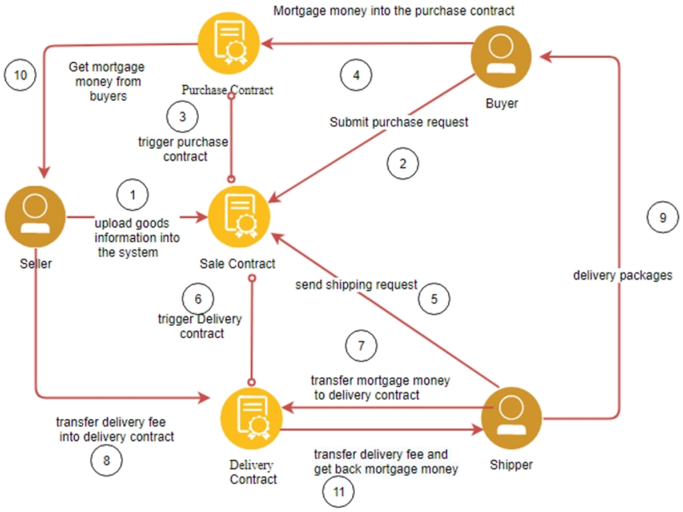

Mortgage money into the purchase contract
Purchase Contract
4
Buyer
10
Get mortgage money from buyers
3
trigger purchase contract
Submit purchase request
2
1
upload goods information into the system
Seller
Sale Contract
9
delivery packages
send shipping request
5
6
trigger Delivery contract
7
transfer mortgage money to delivery contract
transfer delivery fee into delivery contract
8
Delivery Contract
transfer delivery fee and get back mortgage money
11
Shipper

Fig. 1. General description of our proposed multi sections of COD transport process.

B. Detailed COD Scenarios

1) Shipper successfully delivers goods and the buyer successfully receives the goods: The sale contract triggers to purchase contract, seller contract and delivery contract for activating the money transferability function. The order money will be transferred to the seller by purchase contract when the buyer confirms successful delivery to the delivery contract. On the other hand, the delivery contract returns the mortgage money to the shipper which is already deposited before the receiving the order, and the delivery fee is also transferred to the shipper by seller contract immediately. This scenario is illustrated in Fig. 2.

2) Shipper unsuccessfully delivers goods: The sale contract triggers to purchase contract, seller contract and delivery contract for activating the money transferability function. The order mortgage money and delivery money will be transferred to the seller by seller contract and delivery contract due to the shipper failed delivery, then the purchase contract returns the mortgage money to the buyer. This scenario is illustrated in Fig. 3.

3) Buyer refuses to receive goods: The sale contract triggers to purchase contract, seller contract and delivery contract

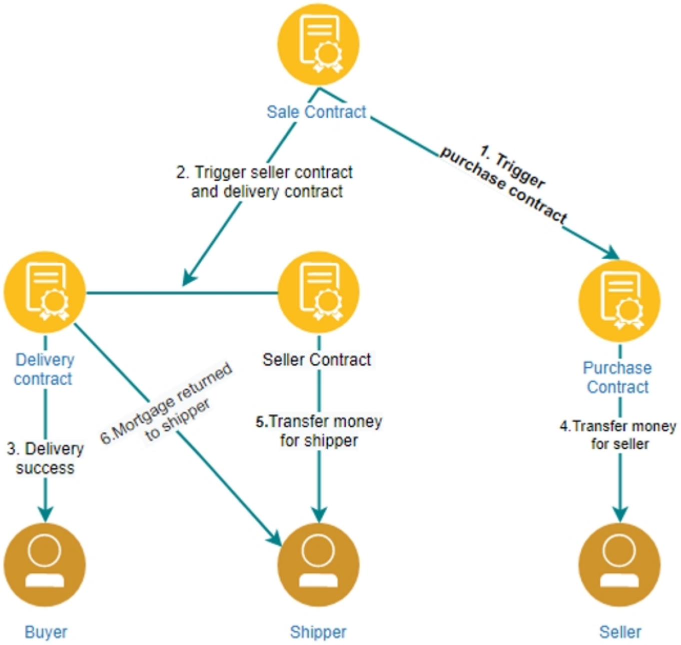

Sale Contract
2. Trigger seller contract and delivery contract
1. Trigger purchase contract
Delivery contract
Seller Contract
Purchase Contract
3. Delivery success
6.Mortgage returned to shipper
5.Transfer money for shipper
4.Transfer money for seller
Buyer
Shipper
Seller

Fig. 2. Case 1: Shipper successfully delivers goods and buyer successfully receives the goods.

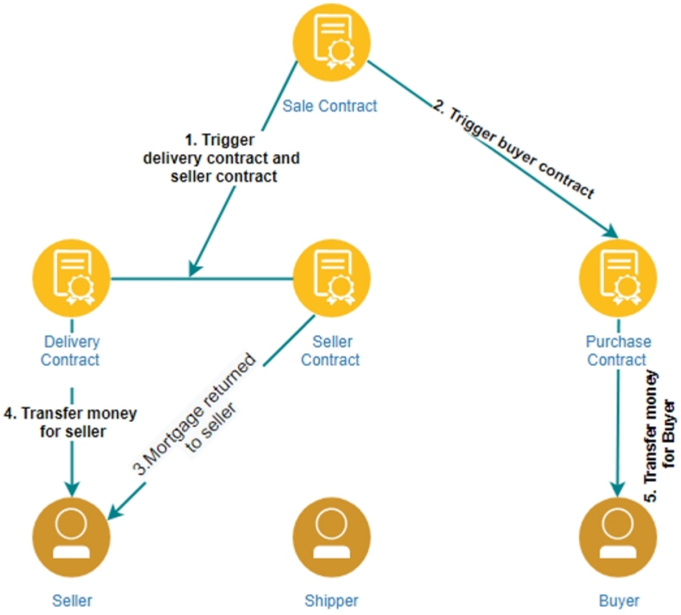

Sale Contract
1. Trigger delivery contract and seller contract
2. Trigger buyer contract
Delivery Contract
Seller Contract
Purchase Contract
4. Transfer money for seller
3.Mortgage returned to seller
5. Transfer money for Buyer
Seller
Shipper
Buyer

Fig. 3. Case 2: Shipper unsuccessfully delivers goods.

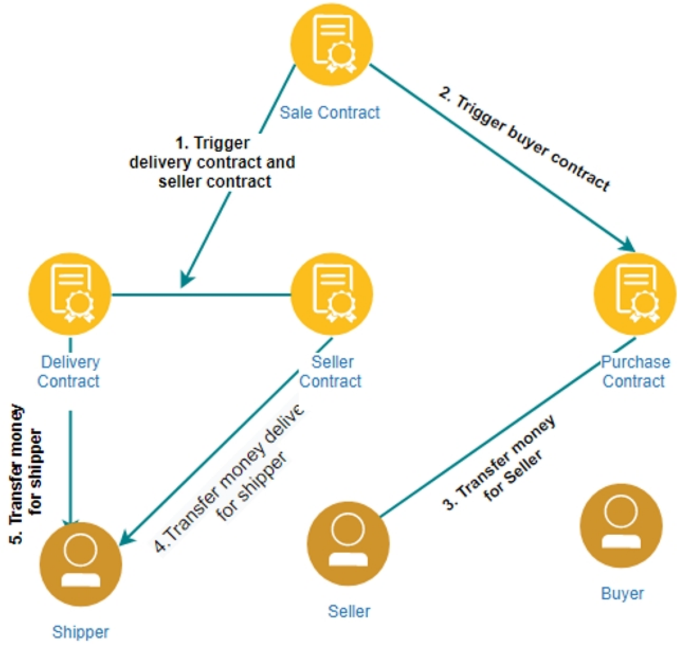

Sale Contract
1. Trigger delivery contract and seller contract
2. Trigger buyer contract
Delivery Contract
Seller Contract
Purchase Contract
5. Transfer money for shipper
4.Transfer money delive for shipper
3. Transfer money for Seller
Shipper
Seller
Buyer

Fig. 4. Case 3: Buyer refuses to receive goods.

for activating the money transferability function when the buyer does not receive the order. The purchase contract utilizes the mortgage money from the buyer to make a payment for the seller, this means that the buyer will be lost their deposit because of “Booming order”. On the other hand, the shipper will receive the mortgage money and delivery fee from the delivery contract and seller contract, respectively. This scenario is illustrated in Fig. 4.

4) Seller provides incorrect goods: The sale contract triggers to purchase contract, seller contract for activating the money transferability function when the seller is wrong order. The shipper checks order from sale contract and notifies the order is wrong, then mortgage money which stored in purchase contract will be returned to the buyer because of failure transaction. Therefore, the seller returns the mortgage money to the seller. This scenario is illustrated in Fig. 5.

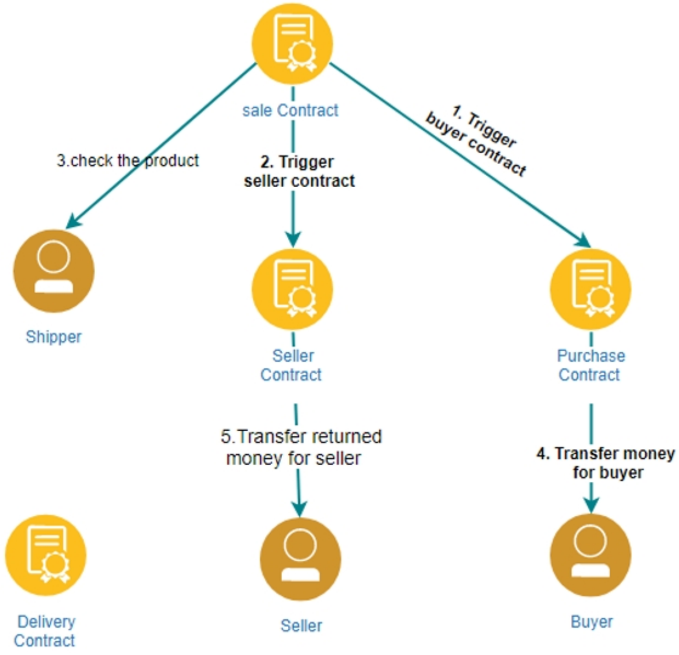

sale Contract
3.check the product
2. Trigger seller contract
1. Trigger buyer contract
Shipper
Seller Contract
Purchase Contract
5.Transfer returned money for seller
4. Transfer money for buyer
Delivery Contract
Seller
Buyer

Fig. 5. Case 4: Seller incorrectly provides goods.

C. Algorithms

The algorithm (1) is the money transferability algorithm. The temporary address is generated at line 1, 2 and 3. Line 4 is to trigger the seller contract for transferring the money to the shipper. The money is transferred to the seller by triggering purchase contract at line 5. Line 6 is to trigger the delivery contract to get back the mortgage money.

Algorithm 1 Money_transferability_algorithm

Input: Order code
Output: Trigger the contracts to execute money transferability and get back mortgage money

```
1:  Address seller_deposit_temp
2:  Address buyer_deposit_temp
3:  Address shipper_deposit_temp
4:  Trigger DepositSeller(seller_deposit_temp) contract and
    seller function to transfer money
5.  Trigger Depositbuyer(buyer_deposit_temp) contract and
    buyer function to transfer money
6.  Trigger DepositShiper(shipper_deposit_temp) contract
    and shipper function to get back mortgage money
```

The algorithm (2) is the refund algorithm. Line 1, 2, and 3 create temporary addresses. Triggering the seller contract to get a mortgage is executed at line 4. Trigger purchase contract to get back mortgage is done at line 5. Line 6 is to trigger delivery contract to transfer the mortgage of the shipper's products to the seller.

Algorithm 2 Refund_algorithm

Input: Order code
Output: Trigger the contracts to execute money transferability and get back mortgage money

```
1:  Address seller_deposit_temp
2:  Address buyer_deposit_temp
3:  Address shipper_deposit_temp
4:  Trigger DepositSeller(seller_deposit_temp) contract and
    function to get back mortgage money
5.  Trigger Depositbuyer(buyer_deposit_temp) contract and
    function to get back mortgage money
6.  Trigger DepositShiper(shipper_deposit_temp) contract
    and shipper function to transfer money
```

The algorithm (3) is called the seller failure algorithm Line 1, 2, and 3 create the temporary addresses. Shipper checks the order and returns the money if the order is not correct is done at line 4. Delivery mortgage money of the seller is a return to the seller is executed at line 5.

1) Case 1: The failure is caused by the shipper. The mortgage money as same as the valuation of the order is triggered to refund payment method as set at line 1 in the algorithm (4).

2) Case 2: The failure is caused by the buyer. When the situation happens, the shipper will receive the package and delivery fee. It is done at line 1 in the algorithm (5) to trigger the transfer money method.

Algorithm 3 Seller_failure_algorithm

Input: Order code
Output: Trigger the contracts to execute money transferability and get back mortgage money

```
1:  Address seller_deposit_temp
2:  Address buyer_deposit_temp
3:  Trigger DepositSeller(seller_deposit_temp) contract and
    function to get back mortgage money
4:  Trigger Depositbuyer(buyer_deposit_temp) contract and
    function to get back mortgage money
```

Algorithm 4 Case 1: Shipper_is_failed

Input: Order code
Output: Trigger refund money method

```
1:  Trigger refund money method
```

Algorithm 5 Case 2: Buyer_is_failed

Input: Order code
Output: Trigger refund money method

```
1:   Trigger transfer money function
```

3) Case 3: The shipment is done successfully. The buyer transfers money to the seller. The seller transfers money to the shipper. The shipper takes the money. Line 1 in the algorithm (6) triggers the transfer money method.

Algorithm 6 Case 3: The shipment is done successfully

Input: Order code
Output: Trigger transfer money method

```
1.  Trigger transfer money function
```

4) Case 4: The order is wrong because of the seller. Shipper checks the order at line 1 and 2 in the Algorithm (7). When the order is wrong, the ReUnfundSellerFail function is activated at line 3. The shipment stops unsuccessfully.

Algorithm 7 Case 4: Seller_is_failed

Input: Order code
Output: Trigger the reunfundSellerFail method

```
1:  Trigger package [ order code ] name = name
2:  If name !=_name
3:      Trigger ReUnfundSellerFail function
4:  EndIf
```

V. EXPERIMENTS

On a blockchain Ethereum model, all of the interaction with the blockchain such as contract reaction, command translation, execution of function has to pay a fee which called gas. Gas costs depend on the complexity and logic of that function. It is calculated based on how much computer resources will be required to perform the function. So that, code optimization is important in Ethereum to be able to save costs. The measurement in four experimental cases of COD has also performed. Case 1, the process takes place normally, the receiving and shipping take place successfully. Case 2, we will refer to the transaction error due to the problem on the shipper. Case 3 is a transaction error due to a buyer problem. Finally, the seller delivers the wrong product, e.g. Case 4. The details implementation of these cases is presented in this section. A complete codes solution is publicized on the authors' GitHub repository[1] (CC BY 4.0) to engage further reproducibility and improvement.

A. Case 1: Transport Successfully

TABLE I. CASE 1: STEP 1

From	0xca35b7d915458ef540ade6068dfe2f44e8fa733c
To	Seller.setPackage(string,uint256,string) 0x692a70d2e424a56d2c6c27aa97d1a86395877b3a
Transaction cost	106490
Execution cost	83618

1) Step 1: Seller creates a package. See Table I:

TABLE II. CASE 1: STEP 2

From	0x14723a09acff6d2a60dcdf7aa4aff308fddc160c
To	DepositBuyer.(constructor)
Transaction cost	270620
Execution cost	166432

2) Step 2: Buyer deposits an amount of money. See Table II:

TABLE III. CASE 1: STEP 3

From	0x14723a09acff6d2a60dcdf7aa4aff308fddc160c
To	Seller.buyItem(uint256,string,address) 0x692a70d2e424a56d2c6c27aa97d1a86395877b3a
Transaction cost	89622
Execution cost	66110

3) Step 3: Buyer buys goods. See Table III:

TABLE IV. CASE 1: STEP 4

From	0xca35b7d915458ef540ade6068dfe2f44e8fa733c
To	DepositSeller.(constructor)
Transaction cost	239581
Execution cost	143213

4) Step 4: Seller places a mortgage. See Table IV:

5) Step 5: Shipper places a mortgage and agrees to deliver goods. See Tables (V, VI, VII, and VIII):

TABLE V. CASE 1: STEP 5A

From	0x4b0897b0513fdc7c541b6d9d7e929c4e5364d2db
To	DepositShipper.(constructor)
Transaction cost	239169
Execution cost	142813

TABLE VI. CASE 1: STEP 5B

From	0x4b0897b0513fdc7c541b6d9d7e929c4e5364d2db
To	Seller.setShipperDepositAddress(uint256,address) 0x692a70d2e424a56d2c6c27aa97d1a86395877b3a
Transaction cost	43443
Execution cost	20635

TABLE VII. CASE 1: STEP 5C

From	0x4b0897b0513fdc7c541b6d9d7e929c4e5364d2db
To	Seller.setSellerDepositAddress(uint256,address) 0x692a70d2e424a56d2c6c27aa97d1a86395877b3a
Transaction cost	43465
Execution cost	20657

TABLE VIII. CASE 1: STEP 5D

From	0x4b0897b0513fdc7c541b6d9d7e929c4e5364d2db
To	Seller.setFlagBuyerAndShiper(uint256) 0x692a70d2e424a56d2c6c27aa97d1a86395877b3a
Transaction cost	56198
Execution cost	34798

B. Case 2: Transport Failure Caused by Buyer.

In this case, a flag is set to indicate that the buyer is the one who causes the transport cancellation. See Table IX.

TABLE IX. CASE 2

From	0x4b0897b0513fdc7c541b6d9d7e929c4e5364d2db
To	Seller.setFlagBuyerFail(uint256) 0x692a70d2e424a56d2c6c27aa97d1a86395877b3a
Transaction cost	56156
Execution cost	34820

C. Case 3: Seller Provides Incorrect Goods.

In this case, a flag is set to indicate that the seller is the one who causes the transport cancellation. See Table X.

TABLE X. CASE 3

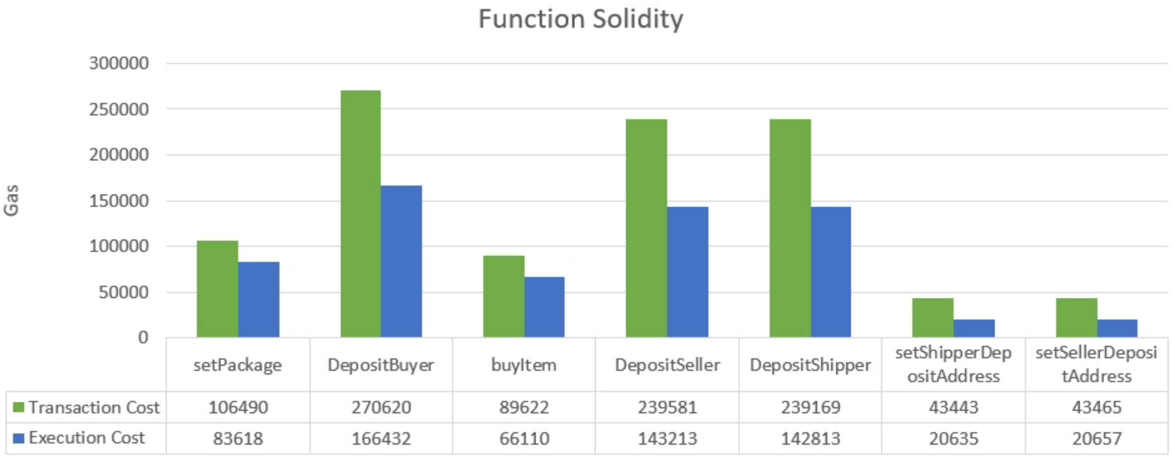
From	0xca35b7d915458ef540ade6068dfe2f44e8fa733c
To	Seller.setFagSellerFail(uint256,string) 0x692a70d2e424a56d2c6c27aa97d1a86395877b3a
Transaction cost	34487
Execution cost	27319

D. Case 4: Shipper Fails to Deliver Goods. See Table XI.

TABLE XI. CASE 4

From	0x4b0897b0513fdc7c541b6d9d7e929c4e5364d2db
To	Seller.setFlagShipperFail(uint256) 0x692a70d2e424a56d2c6c27aa97d1a86395877b3a
Transaction cost	54439
Execution cost	33039

[1] https://github.com/TrieuNam/Smart-Contract-Cash-on-delivery-4.0



	setPackage	DepositBuyer	buyItem	DepositSeller	DepositShipper	setShipperDepositAddress	setSellerDepositAddress
🟩 Transaction Cost	106490	270620	89622	239581	239169	43443	43465
🟦 Execution Cost	83618	166432	66110	143213	142813	20635	20657

Fig. 6. Function solidity

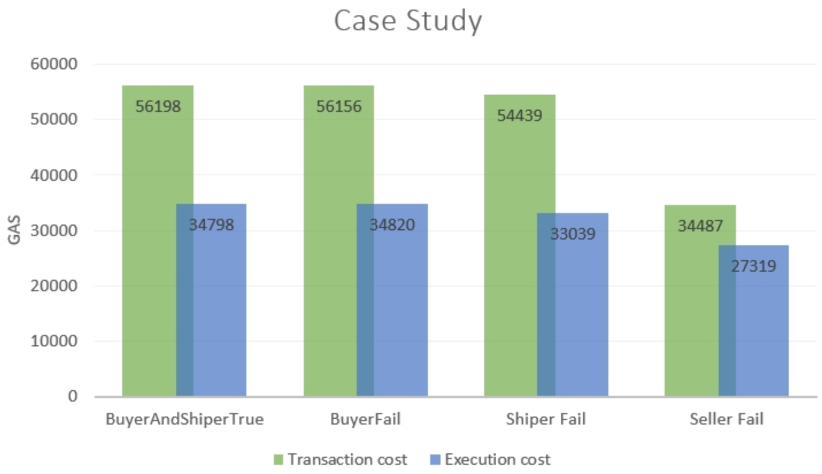


Fig. 7. Gas consumption in every case.

VI. Remarks

In the function solidity diagram, the gas consumption (Fig. 6), we see that the amount of gas in the modality does not pass the transaction and the execution are 300000 and 200000, respectively. The amount of gas increases due to the access to smart contracts as well as the complexity of the methods caused. In the Case Study diagram, e.g. Fig. 7, the amount of Gas in successful trading scenario shows that stable transactions and execution do not exceed 60000 and 40000 respectively. It is important to note that the amount of gas for contract transactions of the participants is very small. In case of transaction of the failure scenarios, the amount of the Gas will be lower than that of in the successful trade. The Gas will be decided by the smart contract when the execution stops, ensuring the amount of loss during the transaction process is insignificant.

VII. Conclusion

As we have demonstrated, the integration of multi-session mechanism in any cash on delivery systems is very effective. Our proposed idea is given to not only enhance the effectiveness of the shipper but also improve the overall performance of decentralized systems. The mechanism works transparently across participants. Several real-world scenarios have been discussed the feasibility of the proposed multi-sessions in boosting the performance and robustness of the COD systems. The crucial delivering issue that previous research has not yet addressed is sufficiently solved. Our proposed mechanism ensure the overall efficiency of delivering process. We are pleased to announce that a new core component of the buyer and seller's dilemma. The authors release our sources codes for further reproducibility and development. We believe that the integration of multi-session mechanism, blockchain technology and smart contracts will cause significant efficiency across several disciplines.

References

[1] M. Halaweh, "Cash on delivery (cod) as an alternative payment method for e-commerce transactions: Analysis and implications," *International Journal of Sociotechnology and Knowledge Development (IJSKD)*, vol. 10, no. 4, pp. 1–12, 2018.

[2] ——, "Intention to adopt the cash on delivery (cod) payment model for e-commerce transactions: An empirical study," in *IFIP International*

Conference on Computer Information Systems and Industrial Management. Springer, 2017, pp. 628–637.

[3] U. Tandon and R. Kiran, "Study on drivers of online shopping and significance of cash-on-delivery mode of payment on behavioural intention," *International Journal of Electronic Business*, vol. 14, no. 3, pp. 212–237, 2018.

[4] J. D. Alie and P. E. Vliek, "International cash-on-delivery system and method," Jul. 24 2007, uS Patent 7,249,069.

[5] A. Asgaonkar and B. Krishnamachari, "Solving the buyer and seller's dilemma: A dual-deposit escrow smart contract for provably cheat-proof delivery and payment for a digital good without a trusted mediator," *arXiv preprint arXiv:1806.08379*, 2018.

[6] H. R. Hasan and K. Salah, "Blockchain-based solution for proof of delivery of physical assets," in *International Conference on Blockchain*. Springer, 2018, pp. 139–152.

[7] H. X. Son, M. H. Nguyen, N. N. Phien, H. T. Le, Q. N. Nguyen, V. D. Dinh, P. T. Tru, and P. Nguyen, "Towards a mechanism for protecting seller's interest of cash on delivery by using smart contract in hyperledger," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 4, 2019. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100405

[8] L. J. Camp, J. D. Tygar, and M. R. Harkavy, "Anonymous certified delivery," Jun. 13 2000, uS Patent 6,076,078.

[9] N. T. T. Le, Q. N. Nguyen, N. N. Phien, N. Duong-Trung, T. T. Huynh, T. P. Nguyen, and H. X. Son, "Assuring non-fraudulent transactions in cash on delivery by introducing double smart contracts," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, 2019. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100584

[10] R. AlTawy, M. ElSheikh, A. M. Youssef, and G. Gong, "Lelantos: A blockchain-based anonymous physical delivery system," in *2017 15th Annual Conference on Privacy, Security and Trust (PST)*. IEEE, 2017, pp. 15–1509.

[11] A. Asgaonkar and B. Krishnamachari, "Solving the buyer and seller's dilemma: A dual-deposit escrow smart contract for provably cheat-proof delivery and payment for a digital good without a trusted mediator," *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pp. 262–267, 2018.

[12] M. Halaweh, "Intention to Adopt the Cash on Delivery (COD) Payment Model for E-commerce Transactions: An Empirical Study," in *16th IFIP International Conference on Computer Information Systems and Industrial Management (CISIM)*, ser. Computer Information Systems and Industrial Management, K. Saeed, W. Homenda, and R. Chaki, Eds., vol. LNCS-10244. Bialystok, Poland: Springer International Publishing, Jun. 2017, pp. 628–637, part 7: Various Aspects of Computer Security. [Online]. Available: https://hal.inria.fr/hal-01656252

[13] M. Barkhordari, Z. Nourollah, H. Mashayekhi, Y. Mashayekhi, and M. S. Ahangar, "Factors influencing adoption of e-payment systems: an empirical study on iranian customers," *Information systems and e-business management*, vol. 15, no. 1, pp. 89–116, 2017.

[14] "Sites like ebay or etsy but decentralized - our features." [Online]. Available: https://openbazaar.org/features/

[15] N. T. T. Le, Q. N. Nguyen, N. N. Phien, N. Duong-Trung, T. T. Huynh, T. P. Nguyen, and H. X. Son, "Assuring non-fraudulent transactions in cash on delivery by introducing double smart contracts," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, 2019. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2019.0100584

[16] G. Wood *et al.*, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum project yellow paper*, vol. 151, pp. 1–32, 2014.

[17] "Ethereum." [Online]. Available: https://www.ethereum.org/

[18] C. Dannen, *Introducing Ethereum and Solidity*. Springer, 2017.

[19] Y. Hirai, "Defining the ethereum virtual machine for interactive theorem provers," in *International Conference on Financial Cryptography and Data Security*. Springer, 2017, pp. 520–535.

[20] M. Wohrer and U. Zdun, "Smart contracts: security patterns in the ethereum ecosystem and solidity," in *2018 International Workshop on Blockchain Oriented Software Engineering (IWBOSE)*. IEEE, 2018, pp. 2–8.

[21] "What is a dapp? decentralized application on the blockchain." [Online]. Available: https://blockchainhub.net/decentralized-applications-dapps/

[22] S. Raval, *Decentralized applications: harnessing Bitcoin's blockchain technology*. " O'Reilly Media, Inc.", 2016.

[23] W. Mougayar, *The business blockchain: promise, practice, and application of the next Internet technology*. John Wiley & Sons, 2016.

[24] A. Wright and P. De Filippi, "Decentralized blockchain technology and the rise of lex cryptographia," *Available at SSRN 2580664*, 2015.

Prediction of Academic Performance Applying NNs: A Focus on Statistical Feature-Shedding and Lifestyle

Shithi Maitra[1], Sakib Eshrak[2], Md. Ahsanul Bari[3],
Abdullah Al-Sakin[4], Rubana Hossain Munia[5], Nasrin Akter[6], Zabir Haque[7]
Dept. of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh[1,2,3,4,5,7]
Dept. of Electronics and Telecommunication Engineering, Daffodil International University, Dhaka, Bangladesh[6]

Abstract—**Automation has made it possible to garner and preserve students' data and the modern advent in data science enthusiastically mines this data to predict performance, to the interest of both tutors and tutees. Academic excellence is a phenomenon resulting from a complex set of criteria originating in psychology, habits and according to this study, lifestyle and preferences–justifying machine learning to be ideal in classifying academic soundness. In this paper, computer science majors' data have been gleaned consensually by surveying at *Ahsanullah University*, situated in Bangladesh. Visually aided exploratory analysis revealed interesting propensities as features, whose significance was further substantiated by statistically inferential Chi-squared (χ^2) independence tests and independent samples *t*-tests for categorical and continuous variables respectively, on median/mode-imputed data. The initially relaxed *p-value* retained all exploratorily analyzed features, but gradual rigidification exposed the most powerful features by fitting neural networks of decreasing complexity i.e., having 24, 20 and finally 12 hidden neurons. Statistical inference uniquely helped shed off weak features prior to training, thus optimizing time and generally large computational power to train expensive predictive models. The *k*-fold cross-validated, hyper-parametrically tuned, robust models performed with average accuracies wavering between 90% to 96% and an average 89.21% F1-score on the optimal model, with the incremental improvement in models proven by statistical *ANOVA*.**

Keywords—Educational Data Mining (EDM); Exploratory Data Analysis (EDA); median and mode imputation; inferential statistics; t-test; Chi-squared independence test; ANOVA-test

I. Introduction

The research field of Educational Data Mining (EDM) applies statistics and machine learning to information stemming from educational environments and is thus contributing to educational psychology. EDM leverages precise, fine-grained data to discover types of learners, examine effectiveness/suggest improvements of instructional learning environments, predict students' learning behavior and advance learning sciences. Baker, Yacef [1] critically identified learners, educators, researchers and administrators to be the four stakeholders of EDM.

The bulk of the academic literature, while addressing problems from the domain of EDM, has taken past academic credentials into account. Fewer academicians resorted to mental health and personality traits. However, the application of features related to students' lifestyle and preferences, as done in this study to predict academic excellence, is a novel approach to the field. In this study, we choose ten such features and apply an evidential function—mapping them to students' expertise in the respective field. The study shows that attributes apart from academic track-records alone can predict academic success which can help institutions to foresee the aptitude of the graduates they are producing, admitting, strategizing for hiring or educating.

Systematic collection of educational data and ML methodologies enable researchers to explore the similarities and dissimilarities among academically sound and unsound students. Recent such researches in the EDM arena have gained momentum using Neural Networks (NNs). NNs are surpassing traditional learning models such as Logistic Regression, Support Vector Machines in performance—characteristically having multiple hidden layers with different activation functions. NNs are versed in fitting complex functions spread through many dimensions featuring multiple independent variables. Backpropagation allows refinement of its initial parameters through numerous epochs, with derivatives showing the direction and learning rate indicating the magnitude of refinement. The weights represent a hierarchical mapping from lower (learns comparatively simpler features) layers to the higher (learns sophisticated features) layers.

The research work addresses a binary classification problem in categorizing final-year Computer Science (CS) students from *Ahsanullah University, Bangladesh* as of their academic performance, following the four EDM phases [2]:

- It is generally held that if a CS student is able to maintain a $\mathbf{CGPA} \geq \mathbf{3.40}$ until the final semester, he/she is faring academically well. First, we exploratorily choose unconventional, unique features by finding their consistent relations with CGPA.
- Then the best use of available data is made by imputing both categorical and continuous variables.
- Third, NN models are proposed to predict academic status.
- The models and features are statistically cross-validated and finer conclusions are drawn.

The sequencing of this paper renders the second section as a review of existing literature, the third section as descriptions of methods followed, the fourth section as a depiction of experimental results and the final section as concluding notes.

II. Related Works

Artificial intelligence-based and statistically analytical methods (Fig. 1) applied in classifying academic performance can be discussed in light of three prototypical dimensions as below.

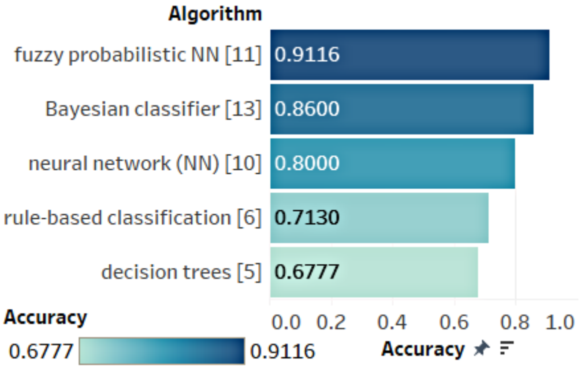


Fig. 1. Comparison among related researches.

A. Conventional Statistics and Decision Trees

Wilkinson, Zhang et al. [3] conducted a study on 706 undergraduate medical students in three consecutive years at the University of Queensland with their objective of modestly determining how precisely each of prior academics, admission tests and interviews accounted for students' performance at post-graduation. These altogether served as the selection criteria which accounted for 21.9% variation in overall scores. They explored GPA to correlate most strongly with performance (*p-value* < 0.001), followed by interviews (*p-value* $= 0.004$) and admission tests (*p-value* $= 0.08$), respectively.

Chamorro-Premuzic et al. [4] established through two longitudinal studies (sample size, $n = 70$, 75 respectively) that personality-measures could testify for students' academic ability. The setting examined students over three academic years at two British universities along academic behavior and personality traits. Sample-1 proved that neuroticism negatively and conscientiousness positively impacted students' academics, accounting for 10% variance. Sample-2 used EPQ-R showing three personality factors were instrumental in predicting academic performance and accounted for 17% variance.

Yadav et al. [5] explored C4.5, ID3 and CART decision trees on engineering students' data to predict final exam's scores. They obtained a true positive rate (TPR) of 0.786 on the 'fail' class using ID3 and using C4.5 decision trees, the highest accuracy of 67.77%. Ahmad et al. [6] proved the impact of demographic information of students spanning eight educational years in predicting academic success. They found rule-based classification techniques to fit the data best with 71.3% accuracy.

B. Unsupervised Clustering Approaches

Oyelade et al. [7] analyzed students' data at a private Nigerian institution using k-means clustering. The cluster analysis was combined with standard statistical methods and a deterministic model was $k = 3$-fold cross-validated using different cluster sizes. The study clustered students labeling them in 5 categories depending on marks' thresholding. However, the study utilized typical academic indicators. Shovon et al. [8] utilized k-means clustering to analyze learning behavior in terms of quizzes, mids and finals in three classes.

C. Supervised, Parametric Learning Approaches

Bhardwaj et al. [9] applied a Naive Bayes classifier on the data of 300 students by preprocessing and transforming the features of raw data. They selected features with probabilities > 0.5. They classified among four classes: first, second, third and fail. The study succeeded in finding interesting features such as living location, mother's qualifications etc. Naser et al. [10] devised an NN based on multilayer perceptron topology and trained it using sophomores' data of five consecutive engineering intakes. They considered high school scores, scores at math and circuitry-based courses during freshman-year, gender among the predictors—gaining 80% accuracy on test-set.

Arora et al. [11] proposed a fuzzy probabilistic NN model for generating personalized prediction which outperformed traditional ML models. The personalized results showed cross-stream generalization capabilities and produced 90%, 96% and 87.5% accuracies on three ranks upon training over 570 instances. The model converged to an error of 0.0265 and included interest, belief, family etc. among eighteen features. Taylan et al. [12] designed an adaptive neuro-fuzzy inference system (ANFIS), a combination of NN and fuzzy systems, to enhance speed and adaptability. The new trend in soft computing produced predictions of students' academics with crisp numerics. Mueen et al. [13] took into account academic participation and scores of two courses and modeled them to Naive Bayes, NN and decision tree—finding the Bayesian classifier to provide the highest accuracy of 86%.

III. Implemented Methodology

Ethical collection of students' data, followed by exploratory analysis, preprocessing, predictive modeling and methodical estimation of metrics led to interesting findings (Fig. 2).

A. Preparation of AUST CS Students' Data

1) Collection of Final Semester's Data:

- **Questionnaire:** Students' responses were gathered via a survey containing questions of multifarious forms including numerical entries, multiple choices and sentential expressions.
- **Environmental setting:** The subjects were surveyed using *Google forms* and the responses were recorded as structured data. There were multiple phases of data-collection either in the labs of AUST or within the comfort of home. No time-constraint allowed subjects to amply think before responding.

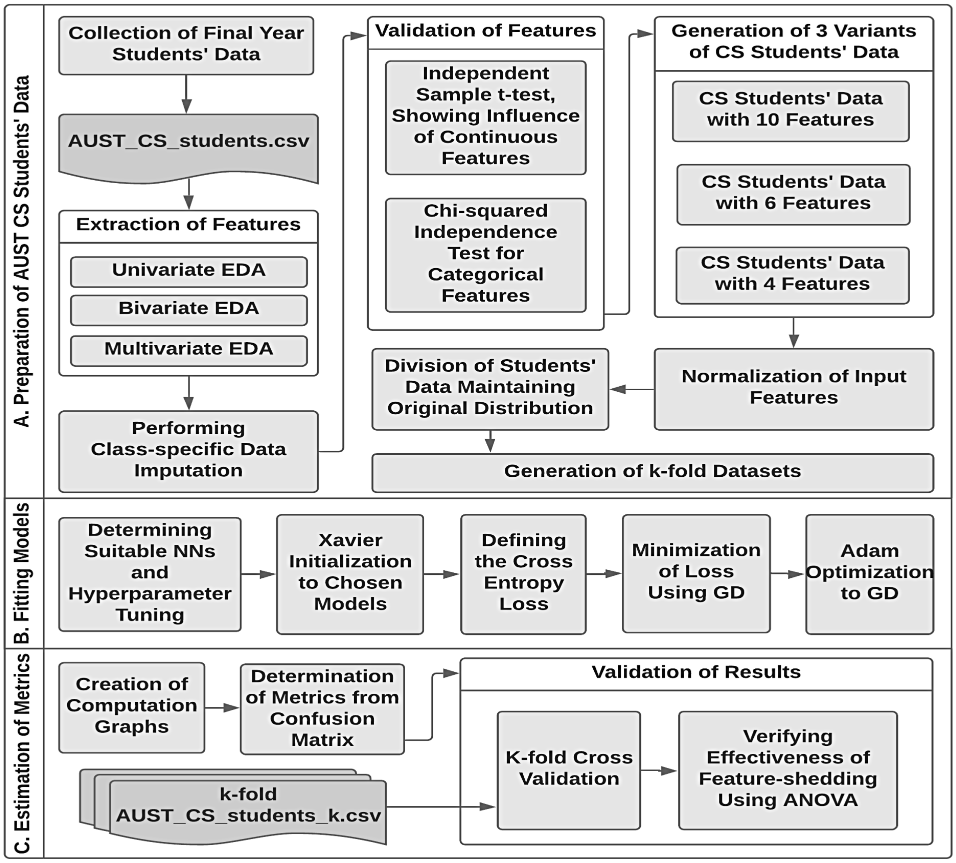

A. Preparation of AUST CS Students' Data
Collection of Final Year Students' Data
AUST_CS_students.csv
Extraction of Features
Univariate EDA
Bivariate EDA
Multivariate EDA
Performing Class-specific Data Imputation
Validation of Features
Independent Sample t-test, Showing Influence of Continuous Features
Chi-squared Independence Test for Categorical Features
Generation of 3 Variants of CS Students' Data
CS Students' Data with 10 Features
CS Students' Data with 6 Features
CS Students' Data with 4 Features
Normalization of Input Features
Division of Students' Data Maintaining Original Distribution
Generation of k-fold Datasets
B. Fitting Models
Determining Suitable NNs and Hyperparameter Tuning
Xavier Initialization to Chosen Models
Defining the Cross Entropy Loss
Minimization of Loss Using GD
Adam Optimization to GD
C. Estimation of Metrics
Creation of Computation Graphs
Determination of Metrics from Confusion Matrix
Validation of Results
k-fold AUST_CS_students_k.csv
K-fold Cross Validation
Verifying Effectiveness of Feature-shedding Using ANOVA

Fig. 2. Workflow of the proposed prediction of academic performance

- **Representativeness:** The sample size (also the population) of 103 subjects represent the whole CS-batch and thus the findings may be generalized among educated youth.

- **Consensual usage:** A pattern recognition lab-project was afoot and students contributed with conscious knowledge and consent to any research thereof.

2) Extraction of Features using Exploratory Data Analysis (EDA): EDA is the statistical process of summarizing tendencies within different attributes of a dataset, assisted by visualizations. The outcome of data-collection, *AUST_CS_students.csv*, had above 30 features and EDA extracted insights beyond predictive analysis to hypothesize features underpinned by data.

A bivariate exploratory visualization (Fig. 3) exposes that pupils with a high attendance rate are the top-scorers (CGPA: 3.3069) and this gradually falls along low and medium attendance. A multivariate observation shows that learners with the strongest passion for both sessional and theory are the highest achievers (CGPA: 3.4398) in terms of academia.

A univariate box-and-whisker exploration (Fig. 4) shows that learners have a median CGPA of 3.25 and programmers investing five or more hours daily in coding are rare. Interestingly, seniors with lower-than-threshold CGPA tend to spend more time (2.261 hours) on social media than their counterparts. The lighter shades of violet tell that either family

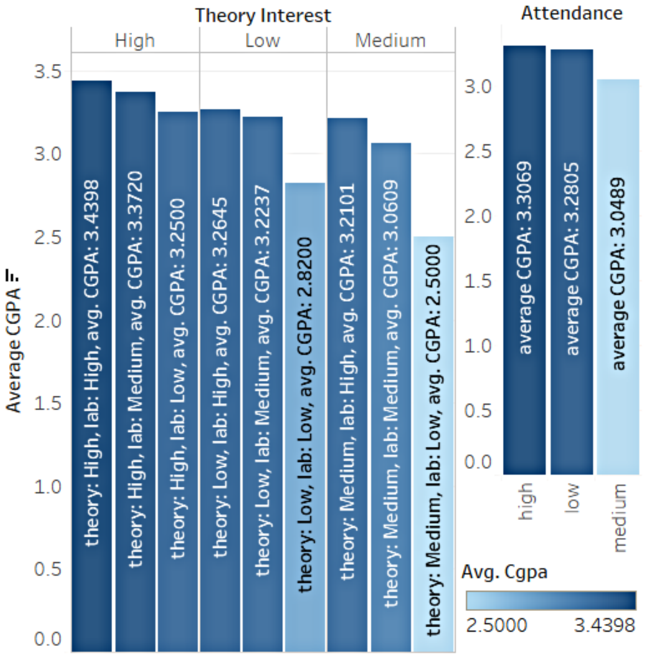


Fig. 3. Relationship of interest in theoretical/sessional CS and attendance (both categorical) with CGPA (continuous)

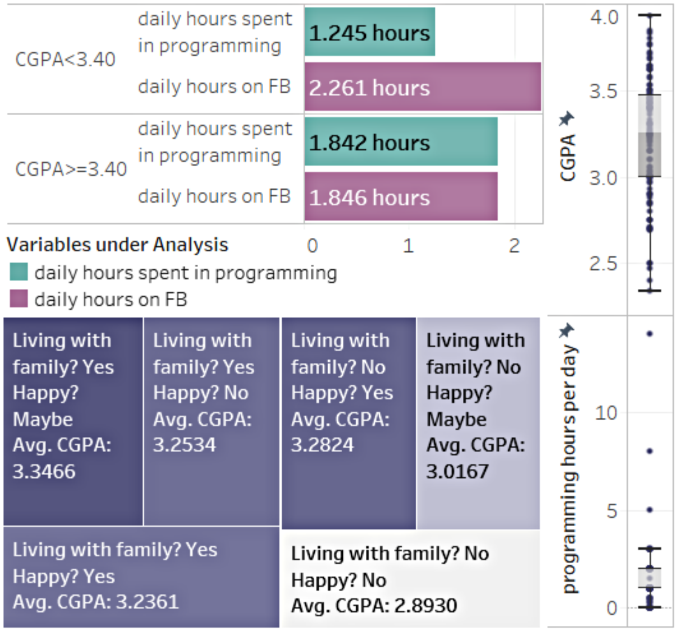


Fig. 4. Univariate and multivariate analyses of lifestyle-factors with CGPA

or happiness should probably be present for a brighter CGPA.

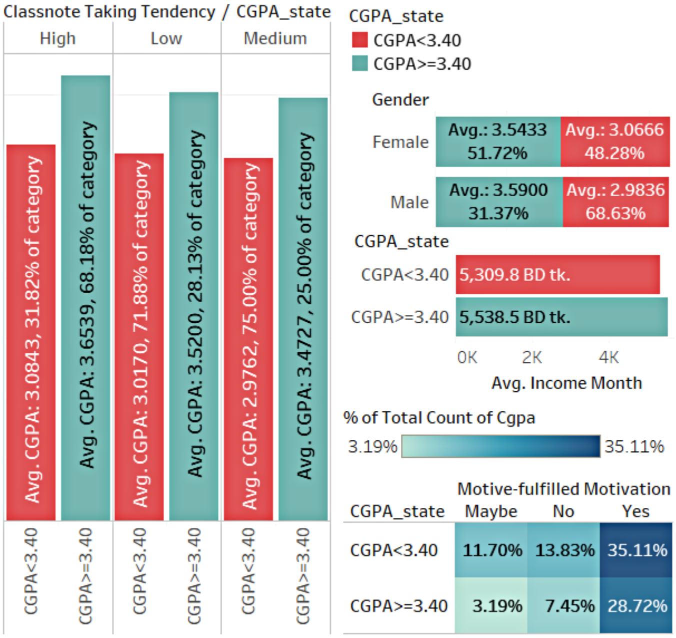


Fig. 5. Thresholded CGPA with respect to preferences (class note, motivation), facts (gender), figures (income)

In another discovery (Fig. 5), class-note taking shows promise in not only that this being high holds the highest CGPA-holders but also in that even the lower-than-threshold students are the highest scorers in their respective category (CGPA $<$ 3.40). More than half (51.72%) of the females hold high CGPA, contrary to their male counterparts.

It is a tendency among students to engage in tutoring and other part-time jobs for self-sufficiency. We find that academically high-achievers tend to earn more than their peers (Fig. 5). Another unintuitive but intriguing cross-tabular finding is that lower-threshold students assert to remain more loyal to their passion (35.11%) even if motive (money, parents' satisfaction, social status) is fulfilled in some other way.

The analyzed attributes clearly show correlations with academic performance and are thus initially justified as features. Data has been visualized according to the best practices, admitting that statistical findings may not always map absolute reality.

3) Performing Class-specific Data Imputation: The statistical process of assigning inferred values to absent fields in accordance with existing fields and summary of the dataset is known as imputation. The *AUST_CS_students.csv* file had numerous blank entries both at categorical and continuous fields, which were eventually filled with class-specific modes and medians respectively (Fig. 6).

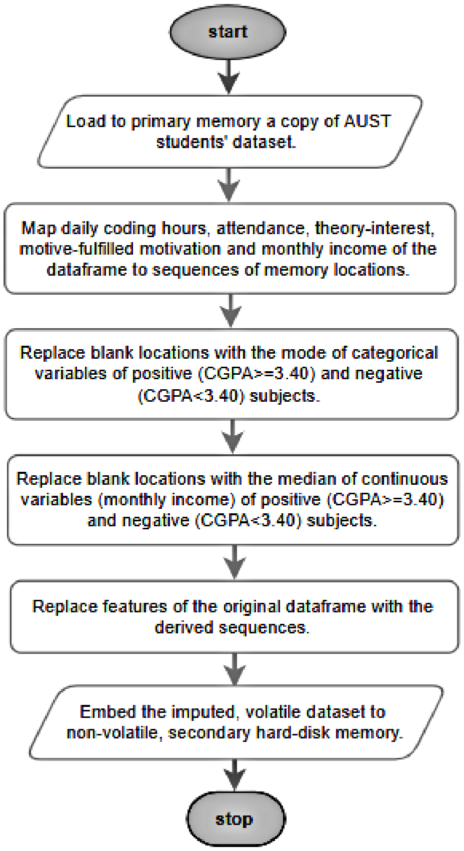

start
Load to primary memory a copy of AUST students' dataset.
Map daily coding hours, attendance, theory-interest, motive-fulfilled motivation and monthly income of the dataframe to sequences of memory locations.
Replace blank locations with the mode of categorical variables of positive (CGPA>=3.40) and negative (CGPA<3.40) subjects.
Replace blank locations with the median of continuous variables (monthly income) of positive (CGPA>=3.40) and negative (CGPA<3.40) subjects.
Replace features of the original dataframe with the derived sequences.
Embed the imputed, volatile dataset to non-volatile, secondary hard-disk memory.
stop

Fig. 6. Class-specialized mode/median imputation algorithm

4) Feature-validation and Generation of Three Variants of Dataset: Inferential statistics is generating statistical models to test hypotheses about a population by producing additional data and eventually deducing propositions using the said model. Most statistical inferences signify *p-value* < 0.05 (a 95% probability of the alternative hypothesis being true), we, however, relax this condition initially and gradually solidify.

To determine if the association between two qualita-

TABLE I. INFERRED STATISTICAL SIGNIFICANCE OF FEATURES

Pearson's χ^2-test			
discrete features	χ^2	**degrees of freedom**	**p-value**
daily hours on FB, state of CGPA	45.254	1	1.73E-11
classnote-taking tendency, state of CGPA	18.553	2	9.36E-05
interest in theory, state of CGPA	4.956	2	8.39E-02
living with family, state of CGPA	2.7991	1	9.43E-02
interest in sessional, state of CGPA	2.7272	2	2.56E-01
attendance in class, state of CGPA	1.978	2	3.72E-01
gender, state of CGPA	0.2086	1	6.48E-01
motive fulfilled motivation, state of CGPA	0.59718	2	7.42E-01
Welch Two Sample t-test			
continious feature	*t*-score	**degrees of freedom**	**p-value**
daily programming hours, state of CGPA	0.21972	36.864	8.27E-01
monthly income, state of CGPA	-0.63789	24.137	5.30E-01

tive variables is statistically significant, we conduct the χ^2-independence test. Firstly, we define the null hypothesis, $H_{\circ}$: *no significant association exists between daily hours spent on social media and CGPA*. Conversely, the alternative hypothesis is H_a. To find evidence against $H_{\circ}$, we compare the observed counts with the expected counts using,

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} \tag{1}$$

Looking up 45.254 in the χ^2-table for 1 degree of freedom, we find the *p-value*: 1.731E-11, highly statistically significant. Other features are analyzed the same way (Table I). The independent samples *t*-test is a test to determine whether the difference between two groups' (CGPA above or below 3.40) means are significant. If so, an attribute can constitute a feature, where the *t*-statistic is calculated as:

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} \tag{2}$$

where,
$\bar{x_1}$ = mean of sample-1
$\bar{x_2}$ = mean of sample-2
n_1 = number of subjects in sample-1
n_2 = number of subjects in sample-2
${s_1}^2$ = variance of sample-1 $= \frac{\sum(x_1 - \bar{x_1})^2}{n_1}$
${s_2}^2$ = variance of sample-2 $= \frac{\sum(x_2 - \bar{x_2})^2}{n_2}$

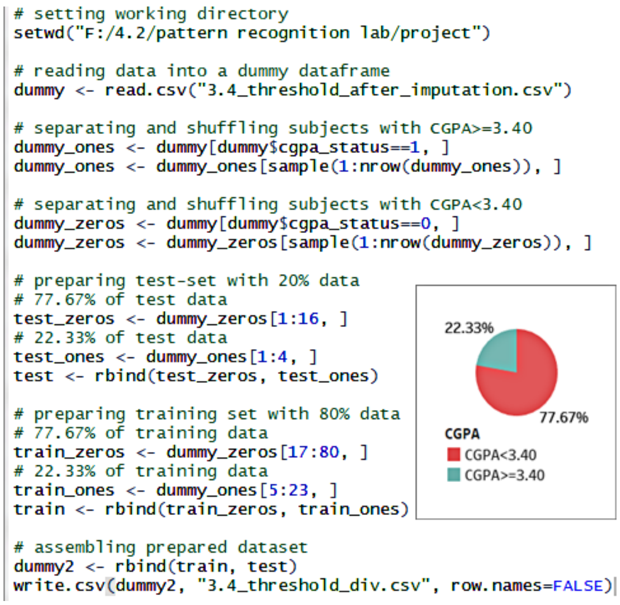
Not all exploratorily extracted features show a strong rejection of the null hypothesis. We start out by retaining all features and gradually drill down to the more significant ones (e.g., *p-value* < 0.4 and *p-value* < 0.1), thus generating three variants.

5) Normalization of Input Features: Preprocessing mandates inputs and parameters to belong to the same range and scale for fair comparison and for the gradient descent to converge following an aligned orientation.

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3}$$

The above formula rescaled all numerics (both categorical: gender, attendance, interest, etc. and continuous: income, daily hours) within the range $[0, 1]$.

6) Maintained Division of Data and k-fold Datasets: Standard ML practices have been followed by assigning a larger set of 80% (83 examples) of total examples for training and the rest 20% (20 examples) for cross-validation. The original distribution of data, i.e., 22.33% positive and 77.67% negative examples, have been maintained throughout training and test data, in order to eliminate any bias during training or cross-validation (Fig. 7).



```
# setting working directory
setwd("F:/4.2/pattern recognition lab/project")

# reading data into a dummy dataframe
dummy <- read.csv("3.4_threshold_after_imputation.csv")

# separating and shuffling subjects with CGPA>=3.40
dummy_ones <- dummy[dummy$cgpa_status==1, ]
dummy_ones <- dummy_ones[sample(1:nrow(dummy_ones)), ]

# separating and shuffling subjects with CGPA<3.40
dummy_zeros <- dummy[dummy$cgpa_status==0, ]
dummy_zeros <- dummy_zeros[sample(1:nrow(dummy_zeros)), ]

# preparing test-set with 20% data
# 77.67% of test data
test_zeros <- dummy_zeros[1:16, ]
# 22.33% of test data
test_ones <- dummy_ones[1:4, ]
test <- rbind(test_zeros, test_ones)

# preparing training set with 80% data
# 77.67% of training data
train_zeros <- dummy_zeros[17:80, ]
# 22.33% of training data
train_ones <- dummy_ones[5:23, ]
train <- rbind(train_zeros, train_ones)

# assembling prepared dataset
dummy2 <- rbind(train, test)
write.csv(dummy2, "3.4_threshold_div.csv", row.names=FALSE)
```

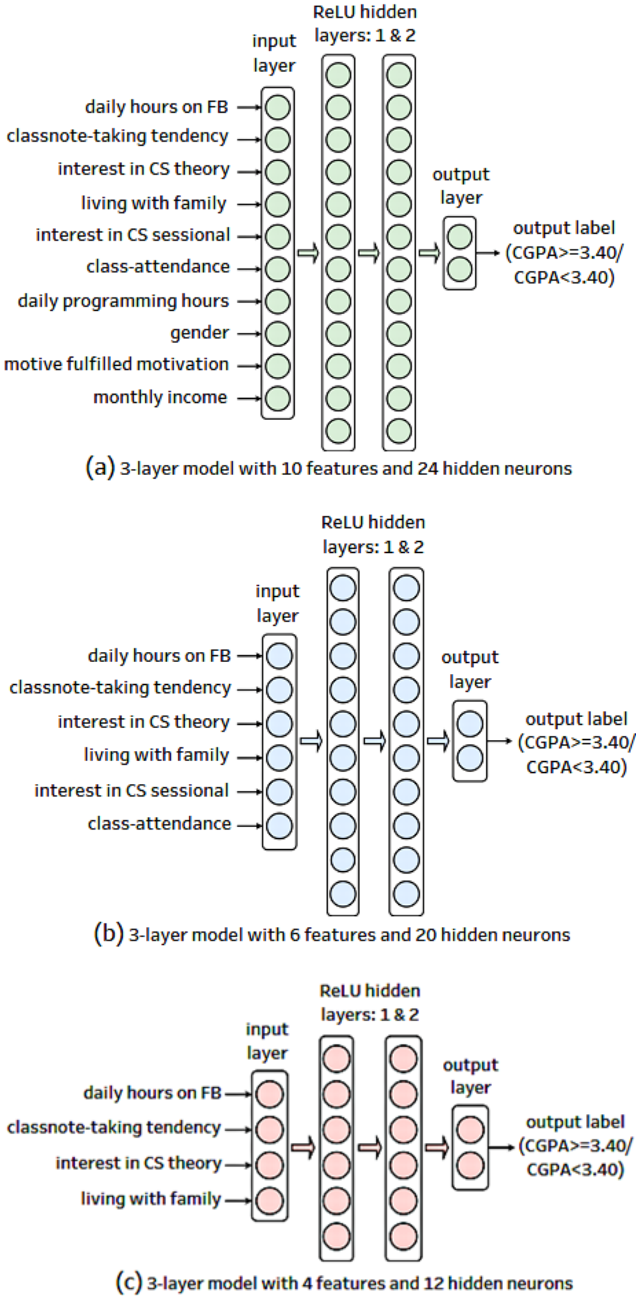
Fig. 7. R script to divide data into an 80%-20% ratio, with the original distribution as inset

K-fold cross-validation is an independent analysis of a model's consistent performance on *k* different training and validation sets. Running the *R* script *k* times provided *k* differently permuted datasets due to shuffle before each binding, thereby allowing the generation of *k*-fold data.

B. Fitting the Models

1) Determining Suitable NNs and Hyperparameter Tuning: Continuous and categorical features' numeric representations were fed to the input layer, with weighted inputs eventually propagating through two *ReLU*-activated hidden layers to the probabilistic *SoftMax* output layer (Fig. 8).

Hyperparameters, upon which the most favorable outcome of a learning model depends besides learnable weights, have been tuned to the following values.



(a) 3-layer model with 10 features and 24 hidden neurons

(b) 3-layer model with 6 features and 20 hidden neurons

(c) 3-layer model with 4 features and 12 hidden neurons

Fig. 8. Proposed three-layer neural network models

- **Number of layers, neurons:** A scarce 83 training examples demanded a simple two hidden-layered network to avoid overfitting. An identical number of hidden neurons were chosen to preclude underfitting. Narrowing the scope to features of greater significance, the complexity reduces; e.g. from 24 (Fig. 8(a)) to 20 (Fig. 8(b)), 12 (Fig. 8(c)).

- **Number of epochs:** 150 for models Fig 8(a, b) and a larger 550 for Fig. 8(c), to converge to an optimum set of parameters.

- **Learning rate:** Depending on epochs, 0.02 for models Fig. 8(a, b) and as small as 0.001 for Fig. 8(c), in order to avoid overshooting across minima.

- **Size of minibatch:** Given the availability of 3.78 GB physical memory, batch gradient descent has been used.

2) Xavier Initialization of Chosen Models: Xavier initialization was used for delicate initialization of weights in order to keep them reasonably ranged across multiple layers as:

$$Var(W) = \frac{1}{n_{in}} \tag{4}$$

Where W is the initialization distribution with zero mean for the neuron in question and n_{in} is the number of neurons feeding in. The distribution is typically *Gaussian* or uniform.

3) Defining the Cross-Entropy Loss Function: The cross-entropy loss has been optimized for the classification problem with a view to obtaining most optimally refined parameters. Here we represent the precise cross-entropy [14], summed over all training examples:

$$\begin{aligned} -logL(\{y^{(n)}\}, \{\hat{y}^{(n)}\}) &= \sum_n [-\sum_i y_i \, log \, \hat{y}_i^{(n)}] \\ &= \sum_n H(y^{(n)}, \hat{y}^{(n)}) \end{aligned} \tag{5}$$

where n denotes the number of training examples, $y^{(n)}$ indicates the ground-truth for a separate example, $\hat{y}^{(n)}$ is prediction generated by the model and i renders the sequence of activation within a layer.

4) Minimization of Loss using Gradient Descent: A set of parameters θ was to be selected in order to minimize loss $J(\theta)$. Gradient descent algorithm [14] initialized θ, then repeatedly performed the following update.

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \tag{6}$$

This update was parallelly performed for all features, i.e., $j = 0, 1, ..., n$ with α being the learning rate. This is a quite natural algorithm that iteratively took steps towards the steepest decrease of $J(\theta)$. Its implementation required the partial derivative term to be computed. Considering only one training example (x, y), we have:

$$\begin{aligned} \frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (h_\theta(x) - y)^2 \\ &= 2 . \frac{1}{2} (h_\theta(x) - y) . \frac{\partial}{\partial \theta_j} (h_\theta(x) - y) \\ &= (h_\theta(x) - y) . \frac{\partial}{\partial \theta_j} (\sum_{i=0}^{n} \theta_i x_i - y) \\ &= (h_\theta(x) - y) x_j \end{aligned}$$

$$Therefore, \quad \theta_j := \theta_j + \alpha (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)} \tag{7}$$

To modify the above for a set of more than one examples, the statement should be replaced by the algorithm below:

Repeat until convergence {

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} \;\; (for\ every\ j) \quad (8)$$

}

5) Adam Optimization to Gradient Descent: Adam is a first-order gradient-based optimization algorithm for stochastic objective functions, using adaptive estimates of lower-order moments. The parameters used for *Adam* in this study are as follows:

- **α :** The learning rate or step size, whose decay is permissible for *Adam*, but has not been used.
- **β_1 :** The exponential decay for first-order moment estimates (e.g. 0.9).
- **β_2 :** The exponential decay for second-order moment estimates (e.g. 0.999).
- **ϵ :** An infinitesimal number to prevent division by 0 in the implementation (e.g. 10E-8).

C. Estimation of Metrics

1) Creation of Computation Graphs: A computation graph is a collective mathematical function represented using the frameworks of graph theory. The round nodes indicate operations while the rectangular ones denote operands, with the directed edges delineating the sequence of mathematical operations performed.

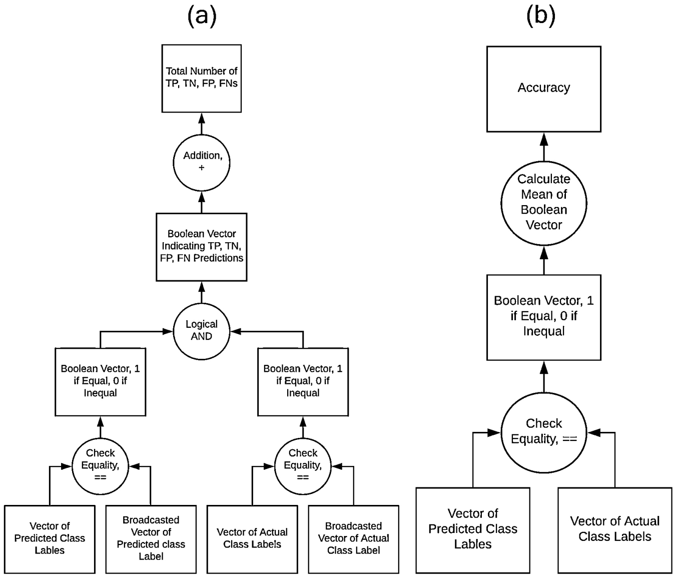
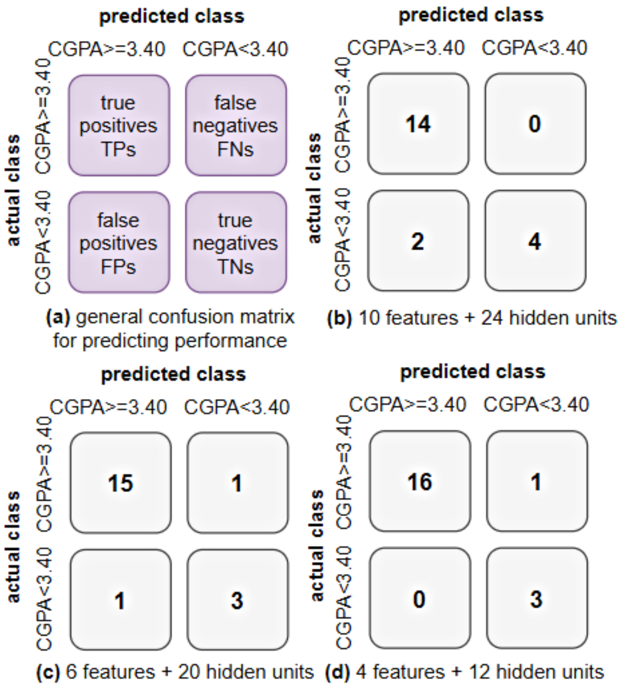

(a)
Total Number of TP, TN, FP, FNs
Addition, +
Boolean Vector Indicating TP, TN, FP, FN Predictions
Logical AND
Boolean Vector, 1 if Equal, 0 if Inequal
Boolean Vector, 1 if Equal, 0 if Inequal
Check Equality, ==
Check Equality, ==
Vector of Predicted Class Lables
Broadcasted Vector of Predicted class Label
Vector of Actual Class Labels
Broadcasted Vector of Actual Class Label
(b)
Accuracy
Calculate Mean of Boolean Vector
Boolean Vector, 1 if Equal, 0 if Inequal
Check Equality, ==
Vector of Predicted Class Lables
Vector of Actual Class Labels

Fig. 9. (a) Generalized computational graph to determine entries associated with confusion matrix; (b) Computation graph portraying computation of accuracy.

TensorFlow's NN framework requires a computation graph to be devised before running a session to refine numerics. The one-hot Boolean representation of class labels has been used to concoct two bottom-up graphs in order to determine entries associated (Fig. 9(a)) with confusion matrices and accuracy on cross-validation set.

After equality-checking, the boolean vector of outputs gave 'high's against the examples identified correctly and 'low's against the converse as to having a CGPA above the threshold. The mean of this data structure rendered the fraction of correct identification (Fig. 9(b)).

2) Determination of Metrics from Confusion Matrix: In the domain of statistical classification, a confusion matrix (Fig. 10(a)) is a special type of contingency table with identical sets of classes in both dimensions—used to account for the performance of a classification model on cross-validation data for which the actual labels are available.



actual class		predicted class	
		CGPA>=3.40	CGPA<3.40
	CGPA>=3.40	true positives TPs	false negatives FNs
	CGPA<3.40	false positives FPs	true negatives TNs

(a) general confusion matrix for predicting performance

actual class		predicted class	
		CGPA>=3.40	CGPA<3.40
	CGPA>=3.40	14	0
	CGPA<3.40	2	4

(b) 10 features + 24 hidden units

actual class		predicted class	
		CGPA>=3.40	CGPA<3.40
	CGPA>=3.40	15	1
	CGPA<3.40	1	3

(c) 6 features + 20 hidden units

actual class		predicted class	
		CGPA>=3.40	CGPA<3.40
	CGPA>=3.40	16	1
	CGPA<3.40	0	3

(d) 4 features + 12 hidden units

Fig. 10. Confusion matrices of our models for some random *k*-th cross-validation

Rows of the tabular layout (Fig. 10) represent instances in an actual class and columns represent predicted labels. The name originates from its making viable to verify if the system is confusing the classes. For our binary classification, we select the popular accuracy, precision, recall and F1-score as evaluative metrics.

- **Accuracy:** proportion of actually correct predictions (both upper and lower-threshold),

$$accuracy = (TP + TN)/(P + N)$$

- **Precision:** proportion of actually correct CGPA>=3.40 identifications,

$$precision = TP/(TP + FP)$$

- **Recall:** proportion of actual CGPA>=3.40 was identified correctly,

$$recall = TP/(TP + FN)$$

- **F1-score:** a trade-off between accuracy and precision, their harmonic mean,

$$F1\text{-}score = (2 * TP)/(2 * TP + FP + FN)$$

3) K-fold, ANOVA-tested Validation of Improvement in Models: Hypothesis-testing technique *ANOVA* (Analysis of Variance) tested the incremental improvement of proposed models' mean accuracies by examining their variances (each having $k = 5$ instances). The samples are random and independent, to the fulfillment of *ANOVA*'s assumptions.

Equality of all sample means is the null hypothesis of *ANOVA*. Hence, $H_\circ$: $\mu_1 = \mu_2 = \mu_3$. Thus, the alternative hypothesis is given as, H_a: *The mean accuracies are reliably unequal*. It essentially calculates the ratio:

F = variance between groups / variance within groups

The greater the ratio, the more the likelihood of rejection of $H_\circ$. The results of *ANOVA* is written in the format $F(b, w)$ where b and w are degrees of freedoms between and within groups, respectively.

Here,
b = number of groups – 1
w = total number of observations – number of groups

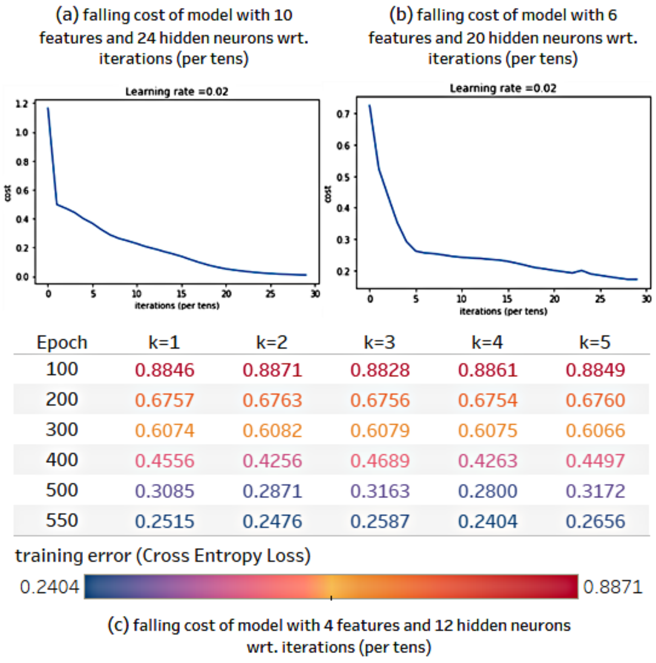
IV. Experimental Results and Discussion

The results originating from fitting three models using the CS students' data are transcending in that even the lowest achieved accuracy surpassed orthodox learning algorithms reviewed in the literature. Application of class-specific median and mode imputation ensured no shrinkage of the already small dataset of 103 tuples, leading to the best use of already existing and further inferred data. Features have been cut down and models' complexity has been gradually reduced, all statistically validated.

For some random *k*, the cross-entropy loss fell with each epoch during training the first two models through 150 epochs with a learning rate of 0.02 (Fig. 11(a, b)). The training was stopped when the error plateaued to a reasonably small value. The third model was trained for 550 epochs with a 0.001 learning rate, whose *k*-fold (*k* = 5) cooling down of error from warmer-shaded greater errors are shown in (Fig. 11(c)).

Firstly, we present the 5-fold consistent results fitting the 10-feature model (Fig. 8(a)) on different cross-validation sets (Fig. 12). The $k = 5$ cases of a consistent 90% test-accuracy can be differentiated by optimized training errors. The model seems to fit training data impressively and is already surpassing traditional models in accuracy (Fig. 14). All cross-validations are consistently giving promising F1-scores (greater or equal to 0.75).

Secondly, we fit another model (Fig. 8(b)) with the same hyperparameters except that now we extract out 6 most significant features as per Table I instead of retainment of all exploratorily discovered features. This scaled down the model's complexity from 24 hidden units to 20. The 5-fold cross-validations resemble training and testing accuracies closely, leading to perfect fitting with test-accuracies as impressive as the former model (Fig. 12).


(a) falling cost of model with 10 features and 24 hidden neurons wrt. iterations (per tens)

(b) falling cost of model with 6 features and 20 hidden neurons wrt. iterations (per tens)

Epoch	k=1	k=2	k=3	k=4	k=5
100	0.8846	0.8871	0.8828	0.8861	0.8849
200	0.6757	0.6763	0.6756	0.6754	0.6760
300	0.6074	0.6082	0.6079	0.6075	0.6066
400	0.4556	0.4256	0.4689	0.4263	0.4497
500	0.3085	0.2871	0.3163	0.2800	0.3172
550	0.2515	0.2476	0.2587	0.2404	0.2656

(c) falling cost of model with 4 features and 12 hidden neurons wrt. iterations (per tens)

Fig. 11. 10, 6-Feature models' learning curves and 4-feature model's lessening of error with epochs

TABLE II. 5-Fold cross-validated results upon training the 4-featured 3-layer final model

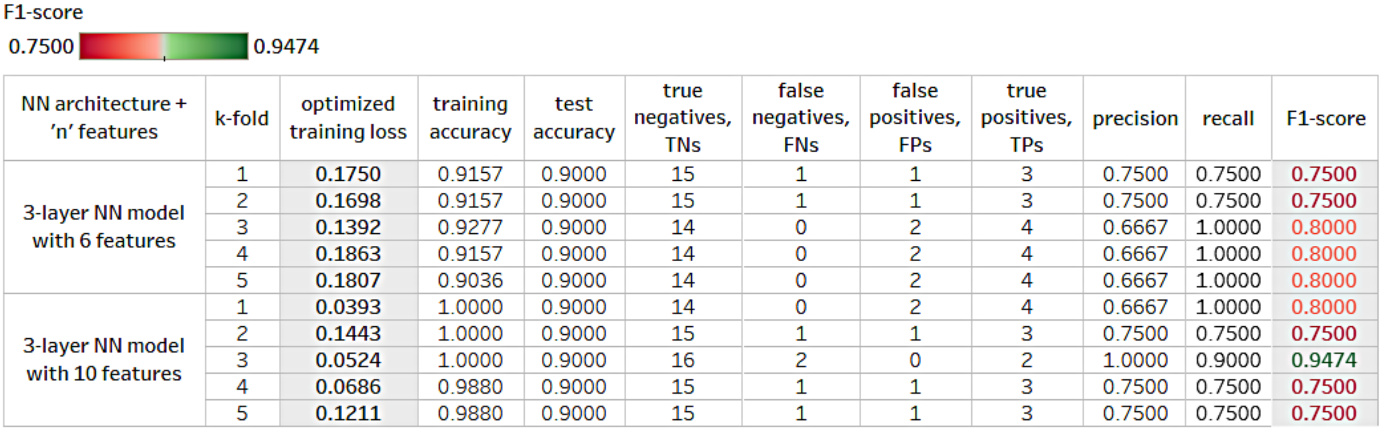
k-fold	optimized training loss	test accuracy	precision	recall	F1-score
1	0.251543	0.95	1	0.75	0.857143
2	0.247627	1	1	1	1
3	0.258734	0.95	1	0.75	0.857143
4	0.24039	0.95	0.8	1	0.888889
5	0.265609	0.95	1	0.75	0.857143

Finally, we become more selective by cherrypicking features with more stringent *p-values* < 0.1 (90% chance of the alternative hypothesis to be true). The network (Fig. 8(c)) thus deprecated its complexity to just 12 hidden neurons, yielding comparatively the most promising (Fig. 13) and consistent (Fig. 12) metrics.

TABLE III. ANOVA-test results verifying the incremental improvement of models

ANOVA (Analysis of Variance) test metrics	Values
degrees of freedom for numerator (ind)	2
degrees of freedom for denominator (residuals)	12
sum of squares of numerators (ind)	0.012
sum of squares of denominators (residuals)	0.002
mean of squares of numerators (ind)	0.006
mean of squares of denominators (residuals)	0.000167
analysed value	36
p-value, Pr(>F)	8.50E-06

Applying *ANOVA* on test-accuracy data from Fig. 12 and Table II, we attempt to test whether the mean accuracies of


F1-score
0.7500
0.9474

NN architecture + 'n' features	k-fold	optimized training loss	training accuracy	test accuracy	true negatives, TNs	false negatives, FNs	false positives, FPs	true positives, TPs	precision	recall	F1-score
3-layer NN model with 6 features	1	0.1750	0.9157	0.9000	15	1	1	3	0.7500	0.7500	0.7500
	2	0.1698	0.9157	0.9000	15	1	1	3	0.7500	0.7500	0.7500
	3	0.1392	0.9277	0.9000	14	0	2	4	0.6667	1.0000	0.8000
	4	0.1863	0.9157	0.9000	14	0	2	4	0.6667	1.0000	0.8000
	5	0.1807	0.9036	0.9000	14	0	2	4	0.6667	1.0000	0.8000
3-layer NN model with 10 features	1	0.0393	1.0000	0.9000	14	0	2	4	0.6667	1.0000	0.8000
	2	0.1443	1.0000	0.9000	15	1	1	3	0.7500	0.7500	0.7500
	3	0.0524	1.0000	0.9000	16	2	0	2	1.0000	0.9000	0.9474
	4	0.0686	0.9880	0.9000	15	1	1	3	0.7500	0.7500	0.7500
	5	0.1211	0.9880	0.9000	15	1	1	3	0.7500	0.7500	0.7500

Fig. 12. 5-Fold cross-validated results upon training 10, 6-featured 3-layer models (6-featured 3-layer model better fitting the data by overcoming overfitting)

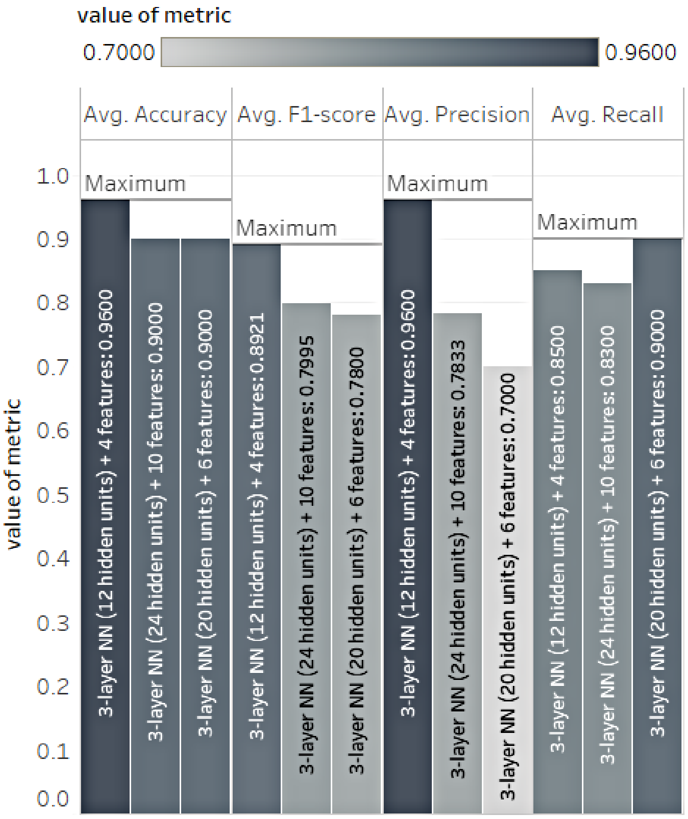


Fig. 13. Comparison among proposed models' average performance measures

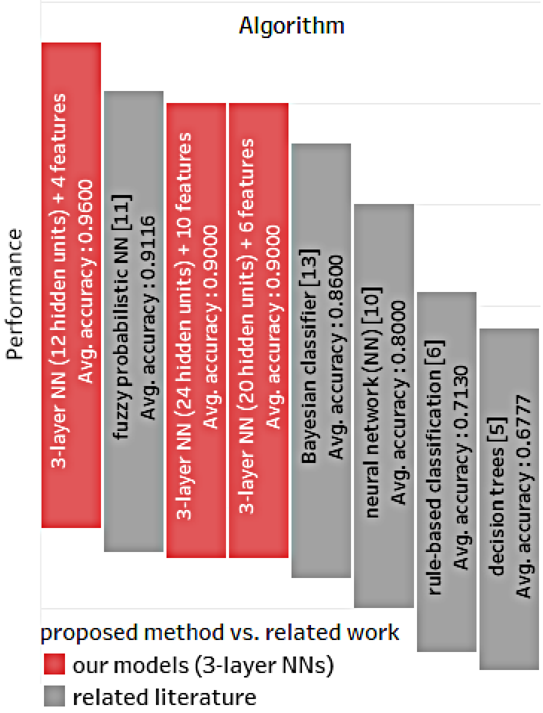


Fig. 14. Comparison between our methodology and reviewed literature

the architectures are systematically different or are just due to sampling errors. The *ANOVA* results (Table III) show:

$$F(2, 12) = 36,\ p\text{-}value = 8.50\text{E-}06 < 0.05,$$

leading us to safely conclude, the models have a systematic effect on the accuracy and similar results can be expected if further data-points are added.

A comparative analysis (Fig. 13) reveals that the most optimized model does brilliantly in accuracy, precision and F1-score. The 6-feature model performs best in terms of average recall. Deployment of the suitable model should be done carefully as different models excel differently. Another comparative study (Fig. 14) manifests that the 3-layer NNs proposed in this paper outsmart many existing methods utilized to solve similar problems.

V. Conclusion

The curious problem of predicting students' performance has, till date, been addressed using direct predictive modeling—this paper proves the effectiveness of visually exploratory and statistical analysis prior to that objective, leading to the following landmarks.

- The study avoids random, carefree, holistic selection of features by first examining their relevance through hypothesis testing, thus establishing the importance of statistical preprocessing.
- The research endorses data-engineered median and mode imputation in handling missing values, introducing no outside noise to training data.
- The paper testifies robustness of the incrementally developed proposed models through *k*-fold cross-validated, *ANOVA*-tested, significant results.

It is recognized that setting the threshold to a CGPA of 3.40 may not epitomize aptitude, which depends on factors external to the scope of this endeavor. However, this study approves and incentivizes further researches to consider lifestyle and personal preferences as useful features towards that end.

REFERENCES

[1] Ryan SJD Baker and Kalina Yacef. The state of educational data mining in 2009: A review and future visions. *JEDM— Journal of Educational Data Mining*, 1(1):3–17, 2009.

[2] 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.

[3] David Wilkinson, Jianzhen Zhang, Gerard J Byrne, Haida Luke, Ieva Z Ozolins, Malcolm H Parker, and Raymond F Peterson. Medical school selection criteria and the prediction of academic performance. *Medical journal of australia*, 188(6):349–354, 2008.

[4] Tomas Chamorro-Premuzic and Adrian Furnham. Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of research in personality*, 37(4):319–338, 2003.

[5] Surjeet Kumar Yadav and Saurabh Pal. Data mining: A prediction for performance improvement of engineering students using classification. *arXiv preprint arXiv:1203.3832*, 2012.

[6] Fadhilah Ahmad, Nur Hafieza Ismail, and Azwa Abdul Aziz. The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences*, 9(129):6415–6426, 2015.

[7] OJ Oyelade, OO Oladipupo, and IC Obagbuwa. Application of k means clustering algorithm for prediction of students academic performance. *arXiv preprint arXiv:1002.2425*, 2010.

[8] Md Hedayetul Islam Shovon and Mahfuza Haque. Prediction of student academic performance by an application of k-means clustering algorithm. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(7), 2012.

[9] Brijesh Kumar Bhardwaj and Saurabh Pal. Data mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*, 2012.

[10] S Abu Naser, Ihab Zaqout, Mahmoud Abu Ghosh, Rasha Atallah, and Eman Alajrami. Predicting student performance using artificial neural network: In the faculty of engineering and information technology. *International Journal of Hybrid Information Technology*, 8(2):221–228, 2015.

[11] Nidhi Arora and JR Saini. A fuzzy probabilistic neural network for student's academic performance prediction. *International Journal of Innovative Research in Science, Engineering and Technology*, 2(9):4425–4432, 2013.

[12] Osman Taylan and Bahattin Karagözoğlu. An adaptive neuro-fuzzy model for prediction of student's academic performance. *Computers & Industrial Engineering*, 57(3):732–741, 2009.

[13] Ahmed Mueen, Bassam Zafar, and Umar Manzoor. Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science*, 8(11):36, 2016.

[14] Ng, A., 2000. CS229 Lecture notes. CS229 Lecture notes, 1(1), pp.1-3.

Extending Conditional Preference Networks to Handle Changes

Eisa Alanazi
Department of Computer Science
Umm Al-Qura University
Makkah, Saudi Arabia

***Abstract*—Conditional Preference Networks (CP-nets) are a compact and natural model to represent conditional qualitative preferences. In CP-nets, the set of variables is fixed in advance. That is, the set of alternatives available during the decision process are always the same no matter how long the process is. In many configuration and interactive problems, it is expected that some variables are subject to be included or excluded during the configuration process due to users showing interest or boredom on some aspects of the problem. Representing and reasoning with such changes is important to the success of the application and therefore, it is important to have a model capable of dynamically including or excluding variables. In this work, we introduce active CP-nets (aCP-nets) as an extension of CP-nets where variable participation is governed by a set of activation requirements. In particular, we introduce an activation status to the CP-net variables and analyze two possible semantics of aCP-nets along with their consistency requirements.**

Keywords*—*AI; changes; CP-nets; preferences; decision making; product configuration

I. Introduction

Preferences, as a notion for desires and wishes, plays an important role in any decision making process [1]. Hence, representing and reasoning about preferences is an important task to develop successful applications. Classical decision theory approaches usually assume the existence of a utility function which maps an alternative u to a numerical value $\mu(u)$ that represent the desirability of the decision maker in having u among other alternatives. Another important type of preference representations is the qualitative models where preference statements are expressed in a comparative way i.e., u is more preferred than u' where u and u' are two distinct alternatives. The latter is known to be less demanding in terms of the cognitive effort required from the user [2]. One of the well-studied models for qualitative preferences is the of Conditional Preference Networks (CP-nets) [3]. A CP-net is a fixed model to represent and reason about qualitative conditional preferences. In CP-net, the preference of one attribute may depend on the value of other attributes. Roughly speaking, the CP-net is a directed graph where vertices represent attributes and the edges show the preferential dependencies. Given a set of attributes or variables V, one usually define a CP-net that is fixed and static through out all the process of decision making. In other words, the set of solutions available to the decision maker (DM) are the same through out the process.

Two important questions arise when dealing with CP-nets: i) What is the best alternative given the current preference information? and ii) Given two alternatives which one is better according to the underlying CP-net? The latter is also known as dominance testing, i.e., deciding which alternative dominates the other. Clearly, the best alternative is the one that is not dominated by any other alternative or solution. Solving or answering dominance questions require searching the space of solutions. Needless to say, for a fixed structure like CP-nets, the answer to the above questions is the same through out the decision process. However, one may expect the answer to differ from time to time due to some changes happening in the network.

Also, while having the same answer to both questions is acceptable on some static domains, it is not the case in interactive and configuration problems. In the latter, users are usually interested in different subsets of the variables satisfying certain requirements and hence, the answers need to take into account the changes. It is intuitive to assume the user interest in having one attribute to be part of the solution space is conditioned upon the existence of other attributes.

Consider, for example, a PC configuration website where customized PCs can be manufactured per customer's requirements and preferences. Assume the user is interested in the type of screen only if high performance graphic card was chosen as part of the customized configuration. In all other cases, she is not interested in knowing the screen type as all the market screens are indifferent to her. In this case, it is clear that there is no need to include the screen type preference for all configurations. But only to those where high performance graphic card is included. This is evident in situations where users build their own products or mass customization products. The users are interested in having a mechanism to include or exclude some variables based on some predefined criteria. This differs from situations where the system (or the user) has a vague idea on the preferences. The latter has been tackled by the literature in a probabilistic manner for uncertain preferences in CP-nets [4].

However, to the best of our knowledge, there exists no attempt to augment scenarios of the form *"If a variable X is included (resp. excluded) in the network then include (resp. exclude) variable Y from the search"* or *"If variable X has the value x then include (resp. exclude) variable Y"*. This has the potential to make the preference representation applicable to wide problems where the set of variables change during the decision making process. In this work, we propose an extension to CP-net called active CP-net (aCP-nets) to tackle variables' inclusion and exclusion in the problem. The main idea is to associate every variable with a status (active or inactive) where the status change is governed by a set of inclusion and exclusion constraints. And, at any given time of the decision making process, only those active variables are included in the search.

This paper is organized as follows: Background information is provided in the next section. Section 3 presents related attempts in the literature. In Section 4, we introduce aCP-nets and the participation constraints. Section 5 discusses two different possible semantics of aCP-nets and show how to solve dominance testing in each semantic. Lastly, conclusion and foreseeable work is discussed in Section 6.

II. Conditional Preference Networks (CP-nets)

A CP-net [3] is a graphical model to represent qualitative preference statements including conditional preferences of the form "*I prefer x to x'* " or "*I prefer x to x' when y holds*". A CP-net works by exploiting the notion of preferential dependency based on the *ceteris paribus* (with all other things being without change) assumption. The CP-net is a set of ceteris paribus preference statements which assumed to be valid only when two alternatives differ in exactly one variable value. Graphically, a CP-net can be represented by a directed graph where vertices represent features (or variables) $V = \{V_1, V_2, \ldots, V_n\}$ and arcs represent preference dependencies among features. Every variable $X \in V$ is associated with a set of possible values (its domain) D_X. An edge from X to Y means the preference of Y depends on the values of X. In such case we say X is a parent of Y and use Pa(Y) to denote the set of parents for Y. Every variable X is associated with a ceteris paribus table (denoted as $CPT(X)$) expressing the order ranking of different values of X given the values of the parents $Pa(X)$. An outcome for a CP-net is an assignment for each variable from its domain. A variable X is an ancestor of another variable Y if X resides in a path from any root node of the graph to Y and X is descendant of Y if Y is an ancestor of X.

Given a CP-net, the users usually have some queries about the preference statements in the network. One of the main queries is to find the best outcome given the set of preferences. We say an outcome o is better than another outcome o' if there exists a sequence of worsening flips going from o to o' [3]. A worsening flip is a change in the variable value to a less preferred value according to the variable's CPT. The relation between different outcomes for a CP-net can be captured through an induced graph. The graph is constructed as follows: Each node in the induced graph represents an outcome of the network. An edge going from o' to o exists if there is an improving flip according to the CPT of one of the variables in o' all else being equal.

Consider the simple CP-net and its induced graph shown in Fig. 1. The CP-net has three variables A, B and C where A and B are unconditionally prefer a and b to $\bar{a}$ and $\bar{b}$ respectively. However, the preference function for C depends on different values of A and B. For instance when $A = a$ and $B = \bar{b}$, the decision maker prefers $\bar{c}$ to c as value of the variable C. The induced graph represents all the information we need to answer different dominance relations between outcomes. An outcome o dominates another outcome o' if there is a path from o' to o in the induced graph otherwise they are incomparable (denoted as $o \bowtie o'$). An outcome is said to be optimal if there exists no other outcome that dominates it. It is known that for acyclic CP-nets there exists a single optimal outcome that dominates all other solution [3]. For example, the optimal outcome for CP-net in Fig. 1 is abc. Apparently, the size of the induced graph is exponential in the number of attributes. Due to the dependency nature of CP-nets, one needs to consult all the values of $Pa(X)$ before deciding which value is preferred for X. This is mainly because missing a value of any $Y \in Pa(X)$ may lead to inconsistent conclusions, i.e., x being preferred to x' and vice versa at the same time. This turns to be the main property we need to guarantee when including or excluding variables and dependencies.

III. Related Work

Since its inception by Boutilier et al. [3], CP-nets have received a considerable attention from the artificial intelligence (AI) community. Many attempts have been proposed tackling different aspects of CP-nets including their semantics [5]–[8], learning [9], [10], and representation [11], [12]. In particular, several works have been made toward extending the semantics and the expressive power of the CP-nets. For instance, the work in [7] extended the CP-net to include preference languages beyond ceteris paribus and thus allow statements to differ in more than more attribute. Another extension of CP-net is the weighted CP-net [8] where the user is able to associate weights to variables. The work in [12] has introduced (conditional) importance over variables. Also, [13] extended the model to augment the notion of comfort when choosing one alternative over another. In [14], [15], the preference-based optimization problem was investigated where hard constraints are assumed to co-exist with a CP-net and the goal is to find a most preferred and feasible solution.

As for extending the semantics of CP-nets to dynamic situations, Bigot et al. [4], [9] studied the case where preferences are uncertain and a probability distribution is associated with a statement. The same problem has been also tackled by [16] where dependencies are associated with probability of existence and every variable X is associated with a distribution over the set of total orders for X's values. The work in [11] considered situations where a webpage content is governed by a CP-net in an adaptive way. Based on the user clicks, the most preferred content is rendered on the page.

However, none of previous attempts has discussed the dynamic aspect of the CP-nets in handling changes that is deterministic and of incremental nature. In particular, the variables' inclusion and exclusion during the search. To this end, there are dynamic models to represent configuration problems similar to the problems tackled in this work. However, they target different knowledge information. One notable representation in this class is the conditional constraint satisfaction problem [17] where constraints and variables are included or excluded during the search. The conditional CSP formalism is limited to constraints and cannot directly applied to qualitative preferences as it is the case in this work.

A closely related area is the preference-based product configuration systems [18]–[22] where a configurator is responsible for customizing the product based on the user preferences. Such configurators allow for a greater flexibility in meeting the users needs and desires. However, we are not aware of

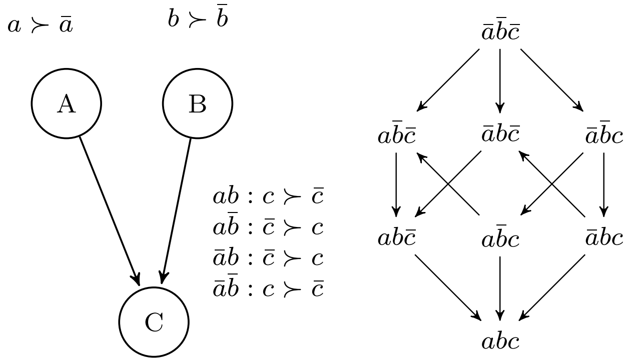


Fig. 1: An acyclic CP-net (left) and its induced preference graph (right).

any previous work targeting conditional qualitative preferences where attributes are included or excluded during the process of configuration.

IV. Active CP-nets (aCP-nets)

In this section we introduce aCP-net extension of CP-nets in the spirit of Conditional CSPs framework [17], [23]. In particular, we add participation conditions upon the CP-net variables that allow variables to be included or excluded during the search. However, unlike conditional CSPs, CP-nets have rich semantics that we need to take into account when reasoning about variables' participation. Specifically, the set of participant variables must be consistent with the preferential dependecies shown in the CP-net structure. Typical CP-net can be viewed as a pair $\langle V, \phi \rangle$ where V is the set of all possible variables in the domain and ϕ is the set of CPTs. Now we define our aCP-net framework.

Definition 1 (aCP-net)**.** *aCP-net is a tuple $\langle V, V_I, \hat{C}, \phi \rangle$ where V is the set of all variables s.t. each variable is associated with an activation status. V_I is the set of initial variables $V_I \subseteq V$ and $\hat{C}$ is the set of activity constraints and ϕ is the set of CPTs.*

Given a variable $X \in V$, we denote the status of X as STATUS(X) . Each variable can be in exactly one state either ACTIVE or INACTIVE at any given time. The set of participation requirements $\hat{C}$ describes different possible changes over the structure while V_I is the set of always active variables that cannot be removed from the domain. V_I can be viewed as the core variables that minimally describe any configuration and thus it is the default preference network for the system. When reasoning about aCP-net, only active variables are taken into account. Given aCP-net instance α, the two main operations over α are including and excluding variables. We describe variables inclusion in terms of Require Variable (RV) and Always Required Variable (ARV) conditions. Similarly, excluding variables from α is done through the conditions of Require Not (RN) and Always Require Not (ARN).

A. Participation Requirements

In this section, we discuss different possible participation conditions over the aCP-net structure. These conditions are Required Variable (RV) and Always Required Variable (ARV) to include variables and Require Not (RN) and Always Require Not (ARN) to exclude variables. Such participation requirements have been proven to be useful in many configuration problems and in particular in the framework of conditional CSP [17]. Generally speaking, all conditions have the following form:

```
condition::result
```

where `condition`$\subset V$ represents the condition variables and `result`$\in V - V_I$ represents a variable and $:: \in \{\xrightarrow{incl}, \xrightarrow{excl}\}$ represents the type of condition (exclusion or inclusion). Whenever `condition` becomes true, `result` is executed and a variable is either included or excluded from the problem domain.

Note that aCP-nets is a generalisation of CP-nets in the sense that the set of conditional dependencies can be preserved along with CPTs in aCP-nets under the special case where $V_I = V$. In such case, the aCP-net will have only one possible instance.

Example 1. *Consider CP-net in Fig. 2. It is easy to see that we can represent the original CP-net as aCP-net instance where $V = \{A, B, C, D, E, F\}$, $V_I = V$ and $\hat{C} = \emptyset$.*

1) Including Variables: One of the simplest yet effective participation requirements to include variables is the required variable (RV) condition. RV has the form $A_1 = a_1 \wedge A_2 = a_2 \wedge .... \wedge A_n = a_n \xrightarrow{incl} X$ where $A_i \in V$ and $X \in (V - V_I)$. This simply says X will be activated (i.e., included) into the problem if every $A_i = a_i$ becomes true.

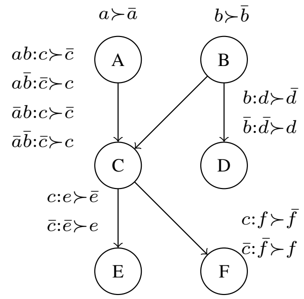

$a{\succ}\bar{a}$
$b{\succ}\bar{b}$
$ab{:}c{\succ}\bar{c}$
$a\bar{b}{:}\bar{c}{\succ}c$
$\bar{a}b{:}c{\succ}\bar{c}$
$\bar{a}\bar{b}{:}\bar{c}{\succ}c$
A
B
$b{:}d{\succ}\bar{d}$
$\bar{b}{:}\bar{d}{\succ}d$
C
D
$c{:}e{\succ}\bar{e}$
$\bar{c}{:}\bar{e}{\succ}e$
$c{:}f{\succ}\bar{f}$
$\bar{c}{:}\bar{f}{\succ}f$
E
F

Fig. 2: Simple CP-net Structure

Similarly, Always Required Variable (ARV) is used to include a variable X when subset of variables $A_1, ..., A_n \in V$ are active regardless of their values and has the form of $A_1 \wedge ... \wedge A_n \xrightarrow{incl} X$.

2) Excluding Variables: Intuitively, Required Not (RN) requirement asserts the exclusion of variable based on other variables values. RN has the form $A_1 = a_1 \wedge A_2 = a_2 \wedge ... \wedge A_n = a_n \xrightarrow{excl} X$ where $A_i \in V$ and $X \in (V - V_I)$. Similar to RN, ARN has the form of $A_1 \wedge ... \wedge A_n \xrightarrow{excl} X$.

V. Semantics of aCP-nets

So far we have described the conditions under which variables may be included or excluded from the network domain without relating them to the underlying aCP-net structure and semantics. Arbitrary changes might lead to violating the semantics of CP-nets. For instance, assume removing one of the parents of a variable X. How CPT(X) should be updated for such changes? or consider including a variable X where one of its parents is not active, how the aCP-net should behave in such circumstances? Therefore, in this work, we study different possible changes and define conservative and open rules for applying different changes into the aCP-net structure. The goal of conservative rules is to represent a valid CP-net (defined below) at anytime of the aCP-net process. On the other hand, the open semantics aim to represent most general case where the resulted instance of aCP-net are not necessarily a semantically correct CP-net.

A. Conservative Semantics

The core concept here is that the changes must result in a valid CP-net at any given time of the solution process.

Definition 2 (Valid CP-net). *Given a set of variables R and their corresponding CPTs ψ, $\langle R, \psi \rangle$ represents a valid CP-net iff for any variable $X \in R$, $Pa(X)$ also exists in R.*

Example 2. *Consider $R = \{A, C, E, F\}$ and ψ is their CPTs for the CP-net in Fig. 2, here R does not represents valid CP-net since the variable $B \in Pa(C)$ is not in R.*

In the conservative semantics of aCP-nets the changes will always result in a valid CP-net. To reflect the conditional dependencies in the structure of CP-net, we assert the activation of a variable based on its parents activation. That is for any variable X with set of parents $Pa(X) \subset V$, for any parent variable $I \in Pa(X)$, either `STATUS(`I`)==ACTIVE` or there exists $c \in \hat{C}$ where I will be activated.

Definition 3 (Consistent Inclusion). *An aCP-net has the consistent inclusion property if for any $c \in \hat{C}$ whenever a variable $X \in V$ to be included, $Pa(X)$ is also included in the domain.*

In the context of aCP-net, we need to be careful in excluding variables. The excluded variable X may be either a leaf node (thus there are no other variables depend on it) or not a leaf node in the aCP-net structure. In the first case, we can safely remove X since we are guaranteed that there will be no other variable S where $X \in Pa(S)$ that might be activated later. In the second case, we can use a procedure to look for whether any of X's descendants will be activated in $\hat{C}$.

Definition 4 (Consistent Exclusion). *An aCP-net has the consistent exclusion property if for any $c \in \hat{C}$ whenever a variable $X \in V$ to be excluded, X has no descendants or for any variable $Y \in V$ of X's descendants, there is no $c \in \hat{C}$such that Y will be activated.*

Definition 5 (Consistent aCP-net). *aCP-net is consistent if it satisfies the consistent inclusion and exclusion properties.*

The goal of conservative rules is to reflect precisely different valid CP-nets from the original CP-net $\langle V, \phi \rangle$ without violating its semantics and dependencies. This is formally proved by the following lemma:

Lemma 1. *If $\mathcal{A}$ is a consistent aCP-net then the set of variables available at any given time of the search represents a valid CP-net.*

Proof: The proof is by contradiction. Assume $\mathcal{A}$ to be a consistent aCP-net but the set of variables available at time t form a CP-net that is not valid. By definition, this means there exists at least one variable X where $Y \in Pa(X)$ was not included at time t but X was included. First assume $X \in V_I$, this is impossible as V_I is available at any given time, and for any X, the set of parents must be part of the initial variables as well. Second, assume $X \notin V_I$, then there must exist at least one participation constraint $c \in \hat{C}$ that result in including X. Given that $Y \in Pa(X)$ was not included at time t, then the inclusion was not consistent and thus the aCP-net does not have the consistent inclusion property which contradicts with our assumption of $\mathcal{A}$ being a consistent aCP-net. ■

Algorithm 1: Consistency Test for aCP-nets

```
input : ⟨$V, V_I, \hat{C}, \phi$⟩: aCP-net Structure
output: True or False
foreach $c \in \hat{C}$ do
    Let $X$ = result in $c$
    if $c$ is inclusion condition then
        foreach $Y \in Pa(X)$ do
            if STATUS($Y$)==INACTIVE then
                Return False
            end
        end
    end
    else
        foreach $P \in$ Descendants($X$) do
            if STATUS($P$)==ACTIVE then
                Return False
            end
        end
    end
end
Return True
```

Although this might seem too restrictive conditions, it may apply in different domains where the changes are known a priori and the dependencies between variables cannot be changed. We describe a procedure in Algorithm 1 to check whether a given aCP-net is consistent or not. The complexity of the algorithm is $O(n|\hat{C}|)$ where $n = |V|$ is the number of variables and $|\hat{C}|$ is the number of participation requirements assuming the parent size for any variable X is bounded by a constant.

1) Reasoning with aCP-net in Conservative Semantics: In the previous section, we have listed some conservative rules for aCP-nets to follow in order to be consistent with the information provided in V and ϕ. These consistency conditions are based on the valid CP-net definition and the set of activation requirements $\hat{C}$. It is clear that the order of conditions listed in $\hat{C}$ plays an important rule in verifying the consistency of a given aCP-net structure. Thus, we order $\hat{C}$ in the following way. First all inclusion conditions are sorted before exclusion ones. For the set of inclusion conditions, we order them from top to bottom according to V. For example for two conditions $c_1, c_2 \in C$ where c_1 and c_2 assert including A and B respectively such that $B \in Pa(A)$ then c_2 is ordered before c_1. The exclusion conditions are ordered in the opposite way from bottom to top.

Example 3. *Consider aCP-net structure θ= $\langle V, V_I, \hat{C}, \phi\rangle$ where V and ϕ represent the set of variables and their CPTs in Fig. 2, respectively. Let $V_I = \{A, B, D\}$ and $\hat{C}$ contains the following set of activity conditions: $a_1 : A = a \xrightarrow{incl} C$, $a_2 : B = \bar{b} \xrightarrow{incl} E$, $a_3 : C \xrightarrow{incl} F$. Fig. 3 shows the CP-net instances resulted from θ when executing a_1, a_2 and a_3 respectively. Observe that θ is a consistent aCP-net. Now, assume we add the activity constraint $a_4 : A = \bar{a} \xrightarrow{excl} C$, θ will not be consistent anymore as E is already activated. This will not be the case if for example there was an exclusion constraint for E.*

a) Answering aCP-net Queries:: Given the activation requirements, we are interested in answering different queries related to the underlying aCP-net structure. The key observation here is that any instance of aCP-net is a valid CP-net representation. In other words, it is a sub-network of the original CP-net. Therefore, for any aCP-net, the semantics of CP-net are directly inherited and utilized. For instance, finding the best outcome for a particular aCP-net instance is done through the sweep-forwarding procedure presented in [3]. Given two outcomes α and β, deciding which one is better can be classified into two cases: (i) α and β are derived from the same network. (ii) They are derived from different networks. In the former, it is clear that we can adopt the CP-net semantics to answer such query. In particular, assume α and β are derived from the network π , we can answer the dominance task by finding a sequence of flips from one outcome to another in π.

However, aCP-net structures usually contain outcomes with different domain spaces. Here, we need to have a mechanism under which we can conclude whether a given outcome is better than another. In other words, we need to find out a new *dominance* relation over aCP-net structure in case outcomes α and β were derived from different networks.

Example 4. *Consider the aCP-net structure θ= $\langle V, V_I, \hat{C}, \phi\rangle$ where V and ϕ represent the set of variables and their CPTs in Fig. 2, respectively. Let $V_I = \{A, B, D\}$ and $\hat{C}$ contains the following set of activity conditions: $a_1 : A = a \xrightarrow{incl} C$, $a_2 : B = \bar{b} \xrightarrow{incl} E$, $a_3 : C \xrightarrow{incl} F$ and $a_4 : D = \bar{d} \xrightarrow{excl} C$. Fig. 4 shows the complete search space for this example with classic baktracking search algorithm. Here, θ has different solutions with different domain spaces. For instance, solutions $ab\bar{d}$ and $abdcf$ are derived from different networks. The crossed out paths represent inconsistent aCP-net instances. For example, assignment $a\bar{b}\bar{d}$ is inconsistent since executing a_4 will lead to removing the variable C while E is included.*

2) Dominance Testing: In this section we provide a method to answer dominance queries when outcomes have different domain spaces. We do so by utilizing the original network structure Ω where $\Omega = \langle V, \phi\rangle$ for a given aCP-net structure θ=$\langle V, V_I, \hat{C}, \phi\rangle$. Let α be any outcome in θ. We know that α might be delivered from any valid CP-net π contained in Ω. This is due to the fact that consistent aCP-net always result in valid CP-nets. Thus, we have $\pi \subseteq \Omega$ for any outcome defined over network π in θ. Our method works as follow. Given two outcomes α and β over θ, if $D_\alpha \neq D_\beta$, we extent them to outcomes $\bar{\alpha}$ and $\bar{\beta}$ over Ω and then do dominance test over the new extended outcomes. Assume $\varphi = V_\Omega - V_\alpha$ to be the set of variables not listed in α. We go top to bottom assigning each variable $X \in \varphi$ to $X = x_i$ such that x_i is the least preferred value given $Pa(X)$.

Example 5. *Consider the aCP-net structure in Example 4 with the following two solutions $\alpha = ab\bar{d}$ and $\beta = abdcf$, it is clear that both Ω_α and Ω_β are subset of the original network Ω, thus we extend α and β to solutions over Ω and then do the dominance testing task. In particular, we have $\bar{\alpha} = ab\bar{d}\bar{c}fe$ and $\bar{\beta} = abdcf\bar{e}$ and $\bar{\beta} \succ_\Omega \bar{\alpha}$.*

B. Open Semantics

So far we have discussed a restrictive but useful semantics for changes over CP-net. The result of the conservative semantics is the guarantee of valid CP-nets during the decision process. This means that whenever we include a variable X, the set of parents are already included (i.e. active). Analogy, when removing a variable Y, we assert that Y has no other variables depend on it in the given aCP-net structure or its participation constraints. Particularly, we assumed a large and original CP-net exist in advance. From the imposed consistency conditions, we implicitly handled the inclusion and exclusion of variables.

However, it could be the case where the user is interested in changing the structure and semantics of the original network. In such cases, we need to distinguish between including/excluding variables through the participation requirements and adding/removing variables and dependencies to the domain. The latter will result in changing original network $\langle V, \phi\rangle$ of the aCP-net structure. In such cases we can relax the consistency conditions posed by the conservative semantics to a more flexible ones. In particular, we consider changes correspond to adding and removing variables and dependencies.

1) Adding Variables or Dependencies: Adding a variable X to the aCP-net structure can be done through two steps. First, we add X to V. Second, we pose an ARV condition in the form $\emptyset \xrightarrow{incl} X$. This will result in X included in the problem. This way, it is possible to add a completely new

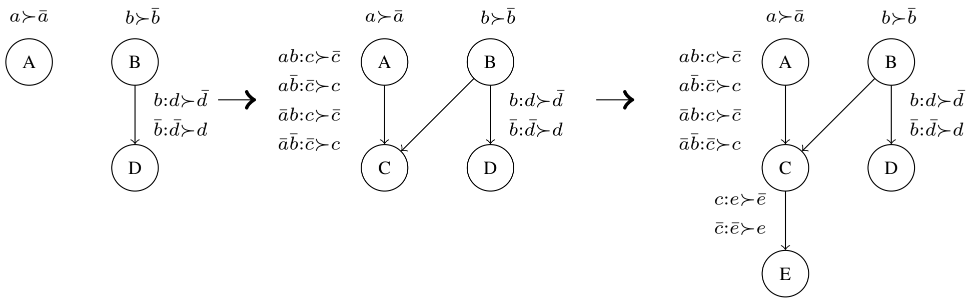

$a{\succ}\bar{a}$
$b{\succ}\bar{b}$
A
B
$b{:}d{\succ}\bar{d}$
$\bar{b}{:}\bar{d}{\succ}d$
D
$ab{:}c{\succ}\bar{c}$
$a\bar{b}{:}\bar{c}{\succ}c$
$\bar{a}b{:}c{\succ}\bar{c}$
$\bar{a}\bar{b}{:}\bar{c}{\succ}c$
C
$c{:}e{\succ}\bar{e}$
$\bar{c}{:}\bar{e}{\succ}e$
E

Fig. 3: Different CP-net instances resulted from the aCP-net in Example 3.

variable to the domain and then reason about its most preferred value.

To add dependency between X and Y where $X, Y \in V$. The only condition here is that the new dependency will not lead to cycles in the aCP-net structure. We refer to $I(X,Y)$ as a dependency of Y on the values of X (i.e., for different $x \in X$ we have an order over D_Y). Adding dependency $I(X,Y)$ will result in updating the CPT(Y)$\in \phi$ in a way that $Pa(Y) = Pa(Y) \cup X$ and for each unique assignments of the parents, we have an order over D_Y.

2) Removing Dependencies and Variables: Before removing a dependency or variable, we first introduce the process of marginalization of a variable $X \in Pa(Y)$ in CPT(Y).

Definition 6 (Marginalization)**.** *Given a CPT ℓ for a variable Y , marginalising $X \in Pa(Y)$ over ℓ (denoted as $\ell^{\downarrow X}$) is a new CPT λ where D_λ=$D_{\ell-X}$ and for any value $x \in D_X$, x has been removed from λ.*

After marginalising X over CPT(Y) ($CPT(Y)^{\downarrow X}$) it might be the case where we have the same assignment of the parents with different orders over D_Y. Thus, we next provide a definition for valid CPTs in the aCP-net structure.

Definition 7 (Valid CPT)**.** *CPT(X) is a valid CPT iff for each assignment γ of the parents, we have the same ordering over D_X.*

For instance, consider the CP-net in Fig. 1, assume we are interested in removing the dependency between B and C ($I(B,C)$). First we marginalise B over $CPT(C)$. The resulted CPT is not valid since for $A = a$ we have two different orders.

Lastly, in order to remove a variable X from the domain, we first need to remove the set of dependencies hold between X and its immediate descendants (i.e., children) and then we can safely remove X.

3) Posing Queries: Consider removing $I(B,C)$ from the CP-net structure in Fig. 1, CPT(C) will have the following statements: $a : c \succ \bar{c}$, $a : \bar{c} \succ c$, $\bar{a} : c \succ \bar{c}$ and $\bar{a} : \bar{c} \succ c$. Obviously, these statements contradict with each other and breaks the intuitive meaning of CP-net of having exactly one order for the same assignment of the parents. How the CPT(C) should be updated in such cases? We can overcome such contradictions by revising the order of the variable. One way to do so is to engage the user by asking different questions. In this particular example, we can ask the user whether she prefers abc to $ab\bar{c}$ in order to know the order when $A = a$. In particular, if $abc \succ ab\bar{c}$ then the CPT(C) is updated to $c \succ \bar{c}$ for $A = a$. The same goes for $A = \bar{a}$ with the query whether $\bar{a}bc$ is preferred to $\bar{a}b\bar{c}$. Such queries hold the promise of revising invalid CPTs and make them valid during the process of decision making.

VI. Conclusion and Future Work

This paper presented *aCP-net*, an extension for CP-nets to include and exclude variables during the search. We listed some consistency conditions under which the resulted changes always form valid CP-nets and, thus will preserve the semantics of CP-nets. We have also analyzed the situation of changes leading to inconsistencies in the preference information and suggested possible techniques to overcome the inconsistency and answer the dominance testing.

Foreseeable work include defining relaxed conditions to allow arbitrary changes over variables and dependencies. Another important future work is to learn the participation requirements from historical interactions with the system. This holds the promise of lowering the burden of specifying participating requirements by the end users.

References

[1] T. Walsh, "Representing and reasoning with preferences," *AI Magazine*, vol. 28, no. 4, pp. 59–70, 2007.

[2] C. Domshlak, R. I. Brafman, and S. E. Shimony, "Preference-based configuration of web page content," in *IJCAI*, 2001, pp. 1451–1456.

[3] C. Boutilier, R. I. Brafman, C. Domshlak, H. H. Hoos, and D. Poole, "Cp-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements," *J. Artif. Intell. Res. (JAIR)*, vol. 21, pp. 135–191, 2004.

[4] D. Bigot, H. Fargier, J. Mengin, and B. Zanuttini, "Probabilistic conditional preference networks," in *Proc. 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, 2013.

[5] R. I. Brafman and Y. Dimopoulos, "A new look at the semantics and optimization methods of cp-networks," in *IJCAI*, 2003, pp. 1033–1038.

[6] C. Boutilier, F. Bacchus, and R. I. Brafman, "Ucp-networks: A directed graphical representation of conditional utilities," in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2001, pp. 56–64.

[7] N. Wilson, "Extending cp-nets with stronger conditional preference statements," in *AAAI*, vol. 4, 2004, pp. 735–741.

[8] H. Wang, J. Zhang, W. Sun, H. Song, G. Guo, and X. Zhou, "Wcp-nets: a weighted extension to cp-nets for web service selection," in *International Conference on Service-Oriented Computing*. Springer, 2012, pp. 298–312.

[9] D. Bigot, J. Mengin, and B. Zanuttini, "Learning probabilistic cp-nets from observations of optimal items." in *STAIRS*, 2014, pp. 81–90.

[10] F. Koriche and B. Zanuttini, "Learning conditional preference networks," *Artificial Intelligence*, vol. 174, no. 11, pp. 685–703, 2010.

[11] R. I. Brafman, "Adaptive rich media presentations via preference-based constrained optimization," in *Proceedings of the IJCAI-05 Workshop on Advances in Preference Handling*, 2005.

[12] R. I. Brafman, C. Domshlak, and S. E. Shimony, "On graphical modeling of preference and importance," *J. Artif. Intell. Res. (JAIR)*, vol. 25, pp. 389–424, 2006.

[13] S. Ahmed and M. Mouhoub, "Extending conditional preference network with user's genuine decisions," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2018, pp. 4216–4223.

[14] C. Boutilier, R. I. Brafman, C. Domshlak, H. H. Hoos, and D. Poole, "Preference-based constrained optimization with cp-nets," *Computational Intelligence*, vol. 20, no. 2, pp. 137–157, 2004.

[15] E. Alanazi and M. Mouhoub, "Variable ordering and constraint propagation for constrained cp-nets," *Applied Intelligence*, vol. 44, no. 2, pp. 437–448, 2016.

[16] C. Cornelio, J. Goldsmith, N. Mattei, F. Rossi, and K. B. Venable, "Updates and uncertainty in cp-nets," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2013, pp. 301–312.

[17] S. Mittal and B. Falkenhainer, "Dynamic constraint satisfaction problems," in *AAAI*, 1990, pp. 25–32.

[18] Y. Wang, D. Y. Mo, and M. M. Tseng, "Relative preference-based product configurator design," *Procedia CIRP*, vol. 83, pp. 575–578, 2019.

[19] H. L. Jakubovski Filho, T. N. Ferreira, and S. R. Vergilio, "Preference based multi-objective algorithms applied to the variability testing of software product lines," *Journal of Systems and Software*, vol. 151, pp. 194–209, 2019.

[20] P. Zheng, Z. Sang, R. Y. Zhong, Y. Liu, C. Liu, K. Mubarok, S. Yu, X. Xu *et al.*, "Smart manufacturing systems for industry 4.0: Conceptual framework, scenarios, and future perspectives," *Frontiers of Mechanical Engineering*, vol. 13, no. 2, pp. 137–150, 2018.

[21] S. Shafiee, A. Felfernig, L. Hvam, P. Piroozfar, and C. Forza, "Cost benefit analysis in product configuration systems," in *Configuration Workshop 2018 (ConfWS 2018)*. CEUR-WS, 2018.

[22] E. Gençay, P. Schüller, and E. Erdem, "Applications of non-monotonic reasoning to automotive product configuration using answer set programming," *Journal of Intelligent Manufacturing*, vol. 30, no. 3, pp. 1407–1422, 2019.

[23] M. Sabin, E. C. Freuder, and R. J. Wallace, "Greater efficiency for conditional constraint satisfaction," in *CP*, 2003, pp. 649–663.

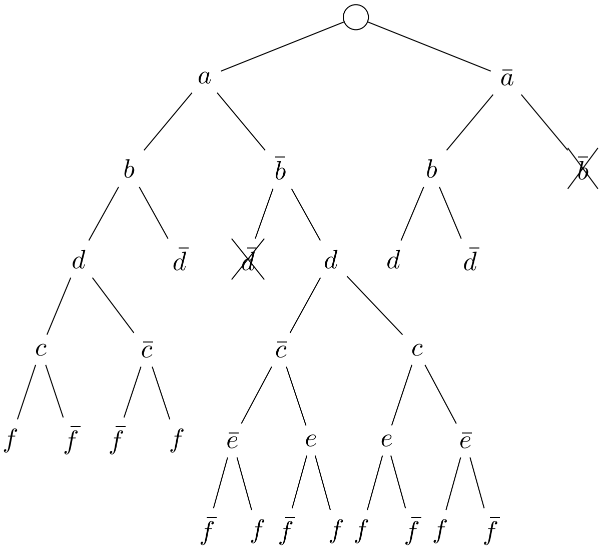


Fig. 4: The complete search space for Example 4

Thai Agriculture Products Traceability System using Blockchain and Internet of Things

Thattapon Surasak[1]*, Nungnit Wattanavichean[2], Chakkrit Preuksakarn[3], Scott C.-H. Huang[4]
Institute of Communications Engineering, National Tsing Hua University, Hsinchu City, Taiwan[1,4]
Department of Applied Chemistry, National Chiao Tung University, Hsinchu City, Taiwan[2]
Department of Computer Engineering, Kasetsart University, Nakhon Pathom, Thailand[3]

Abstract—**In this paper, we successfully designed and developed Thai agriculture products traceability system using blockchain and Internet of Things. Blockchain, which is the distributed database, is used for our proposed traceability system to enhance the transparency and data integrity. OurSQL is added on another layer to easier query process of blockchain database, therefore the proposed system is a user-friendly system, which cannot be found in ordinary blockchain database. The website and android application have been developed to show the tracking information of the product. The blockchain database coupling with Internet of Things give a number of benefits for our traceability system because all of the collecting information is in real-time and kept in a very secured database. Our system could have a huge impact on food traceability and supply chain management become more reliable as well as rebuild public awareness in Thailand on food safety and quality control.**

Keywords*—*Blockchain; internet of things; supply chain management; product traceability; distributed database; data integrity; ourSQL

I. Introduction

The Kingdom of Thailand is situated in Southeast Asia. Thailand is an agricultural country. According to The Trading Economics website reported that the Gross Domestic Product (GDP) from agriculture is the major section that raise the overall GDP in 2018. Although there is the rapid expansion of industrial sectors, agriculture is still a majority [1].

However, there are a number of problems and concerns regarding the quality of Thai agricultural products. Not only due to the plant diseases and chemical contamination, which is the main problems, but also the uncontrollable factor such as weather conditions or disasters. These kind of problems cause the reduction of agricultural products quality. Eventually, this problem leads to the lower income of farmer as well as consumer confidence for their receiving products. [1], [2].

The traceability system is an appropriate solution to control, prevent and exterminate various problems and concerns in supply chain, especially in food and agricultural sector [3]. In this process the details of the product since it was farming until delivered to the consumers has been controlled to meet required food safety and quality standards (GMP, HACCP, and HALAL). European Union (EU) has launched a law and policy for the meat products registration and identification to guarantee that the meats products sold in European countries are verifiable and traceable [4]. The traceability system is reliable when it consists of secured database and trusted monitoring system. Both components can be provided by technology [5]. In this paper, blockchain database is implemented with IoT devices.

Blockchain is a new technology that catch many attentions in various fields of researches at this moment. The structure of blockchain illustrates in Fig. 1. It is a chain of blocks which each block stores all information of network activities after the block was added to the chain [6]. This feature makes blockchain become traceable database. Blockchain allows every user to add data as a transaction. Moreover, all data in the blockchain can be reviewed by every user, but no one able to change it [7]. The security and validity of the information in blockchain controlled by one process called ′mining process′ [8]. This process aims to verify the information before adding any block to the chain. A person who verifies the new transactions and create the block known as miners. Miners utilize a consensus algorithm to add the new block. This algorithm is similar to a voting system. Once the transaction is confirmed by enough number of nodes (voter), it will be validated and permanently be a part of the database. After the block was added to the database, it cannot be changed. Therefore, this method ensures the transparency, trust, and traceability in a system. As this reason, blockchain is very famous among financial applications. The most well-known is Bitcoin, which is a peer-to-peer electronic cash system [9]. In addition, blockchain has one interesting feature called smart contract, which is the digital agreement. This feature is suitable for supply chain management because it can create a deal between farmers and consumers [7].

Internet of things (IoT) is enabling the connection between machine to machine (M2M) over the Internet [10]. With IoT, sensors, actuators and electronic devices can collect and exchange the data with each others [11]. In food supply chain traceability system, the IoT devices is calibrated to make sure that its measurement value is accurate. The data collected from the IoT device is reported in real-time [12]. One advantage of using IoT device to collect the data is because it reading value cannot manipulate by human [13]. With the implementation to blockchain database, all true value is kept directly in the database without changing, which leads to high reliable of the system [14], [15].

In this paper, we propose a promising solution in which Internet of Things, blockchain distributed database have been integrated. Section II reviews the works related with tracability system by applying the SQL database, blockchain, and IoT technology. Section III explains our proposed blockchain traceability system which consists of the blockchain database technology (OurSQL) and system architecture. Section IV presents

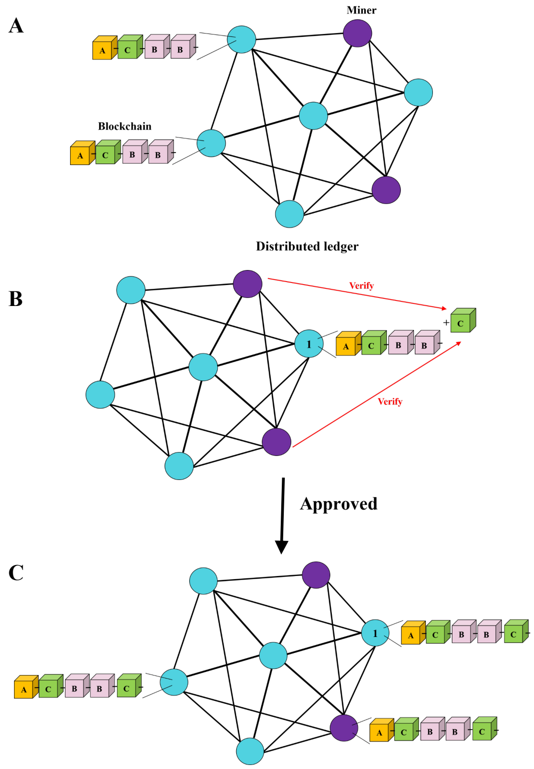


Fig. 1. Blockchain and the block creation process: (A) blockchain and distributed ledger. (B) The block creation process starts when one node (user) requests to make a transaction over the whole blockchain network. The particular transaction needs to be verified by the others. After transaction verified, the miner can create the block and connect it to the chain. (C) After the block creation process is approved, every node in the system is also updated with a new block created from part (B).

the case study to verify the feasibility of proposed system. The results are shown on website and android application. The evaluation and benchmark of our proposed system design have been discussed. Finally, Section V describes the conclusion and future work.

II. Related Work

In traceability system for food and agricultural supply chain, users are able to track and trace the products from the origin [16]. In the past, only database technology was required to create a traceability system and the user could directly put all information into the database. The most popular database to store the product information is SQL database [17]. Although this solution can track the product information, the human error is still another big issue in this data collection process [13].

The use of Internet of Things (IoT) technology introduces many benefits for both business and human aspects [18]. The supply chain management has various advantages by integrating the IoT [19], [20]. For example, the user will be able to use the real-time monitoring system, which is the most important system in the food supply chain, by implementing the IoT traceability system with the main food supply chain logistics [21]. This provides a number of benefits regarding the food quality and security. RFID and QR-code also be implemented in order to allow the users to easily retrieve the information [6], [20], [22]. For the previous work, ZigBee mesh network has applied to work on the real time GPS tracking and monitoring. This increases the number of benefits in the logistic management for agricultural products. However, there have some problems regarding security issues due to the server misusing [23]. In order to control the temperature and humidity, the wireless sensor networks have been applied. The application can automatically determine whether it going to water the tomato greenhouse. With the system, the tomato greenhouse is in controlled [24]. Furthermore, one of smart farm research in Thailand also used the wireless sensor networks to detect the environmental data of the farm using several sensors. Air temperature, humidity, water temperature and pH have been chosen as the parameters that represent the quality of planting process. After the data collected, the data were analyzed, summarized, and store in the database, respectively. This information includes name of plants, origins, and farm certification. To monitor the smart farm conditions, an Android application was created. The QR code also be generated to allow the customers to access additional information after the products have been shipped to the stores. The main advantage of a smart farm and traceability system using IoT, which proposed in this research, let the customer to have more confidence in the agricultural product [25]. Anyway, the more benefits provided, the more emerging security issues about the IoT server and devices [26]. In previous year, an authentication mechanism has been created to protect the system [10]. However, an authentication itself can avoid the misuse problem only because the data could be changed after the users authenticated themselves to the system [27]. Therefore, data integrity is another important part that developer need to carefully designed along with creating the IoT system [11].

The implementation of blockchain with the IoT will be the most strongest system that can applied for the food supply chain [28]–[31]. Recently, the researchers have impressively adapt the blockchain technology with IoT for the supply chain traceability system [31]. One of these applied the blockchain and IoT with soybean traceability [32]. This work creates the smart contract which is one of the blockchain feature to use as a real agreement between seed company and the farmer. The contract is used when the farmer wants to ship their product to distributors or retailers. With this work, the researchers can guarantee that the farmer gets the most benefit they deserve. Blockchain technology has also be adapted with the RFID to use as a food supply chain tracking system in China [6]. The blockchain technology provides the secure database for the food supply chain and the RFID is applied as an interactive database that can be used by scanning the RFID and getting interested information. A research applied the Ethereum blockchain, which is well-known blockchain technology, for traceability system [33]. In order to show the data from Ethereum blockchain, web3 API is used [29]. However, the data stored in ledger can be seen by any user in the Ethereum network which is not suitable for sensitive data. Moreover, using a large-scale network dramatically consumes a lot of time and cost for block creation process. Another shortage of using blockchain as the database is because it required a specific command based on each blockchain technology to query the information [34], [35]. In other word, the common blockchain database is not easy to query the information we need. This problem can be solved by implementation of the

SQL database with the blockchain technology [7], [36].

III. Blockchain-based Traceability System

As mentioned before, each blockchain technology requires a specific command to retrieve the information. This problem can be solved by implementation of the SQL database with the blockchain technology. In this research, the IoT sensors and OurSQL (Blockchain replication database technology) have been combined together to immediately store the agricultural product information into the blockchain database. In this study, OurSQL is used as a core of the system in order to collect the data from the IoT sensors. OurSQL platform provides a unique ability to share selected information including location tracking, temperature, humidity and ownership transfers. The first implementation of our system provides for the beef products in Ratchaburi Province of Thailand. There are three important information to collect for the beef which are temperature, humidity and location. Therefore, the temperature, humidity and GPS sensors were used to collect the data on our system. In our design, the product information is collected across the whole supply chain. Therefore, the data collection has been divided into three parts. The first part is the information from the cow farm. The second part collected during the beef production. Then, products shipping is the last part of our collection. These three parts will collect the data in the blockchain database by sensors. From the basic concept of the system, the implementation process was designed based on the supply chain system of the beef production as can see in Fig. 2. The last step of our system design is to create website to allow the customers or restaurant owners who bought the beef to track the beef information. With our website, we show the location tracking information with the temperature and humidity of the beef during the shipping process. Moreover, the information of the particular cow is presented to confirm the quality of the whole beef production.

A. *OurSQL*

OurSQL is a server software running between a mysql server and software for the Database client. It has been created by Roman Gelembjuk. OurSQL enables the use of basic SQL client tools and libraries to perform distributed ledger as a SQL database. This tool has two core components, including the blockchain management server and database proxy server. Each node operates with a single database of mysql. Blockchain data are kept together with data tables in the same database. After creating a decentralized database and starting blockchain or joining an existing decentralised database on initial, OurSQL node server can be created. A node server listens on two ports: (1) A local mysql client database proxy server, (2) A port to interact with other existing database cluster OurSQL nodes. mysql client connects on a known port to a proxy. This method is the same as connecting directly on "localhost" to mysql server, just specific another port number. With this decentralized database, mysql client can be a SQL application or a DApp. Moreover, the user can be able to perform any sort of SQL query using SQL command with this decentralized database as well. OurSQL system architecture can be seen in Fig. 3.

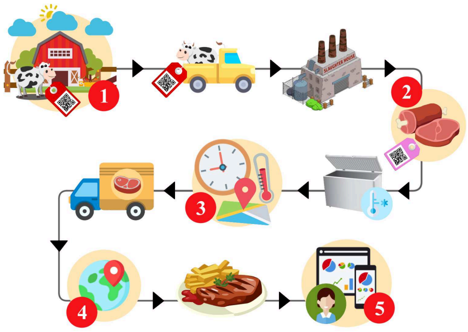


Fig. 2. Data collection methods: (1) The assigned government officer will create a tag (tracking ID and QR code) for a cow and will also track the feeding information during the cow production. When the cow is ready to be slaughtered, the IoT sensors will be used in order to track the cow information during the shipping and transport. (2) After the beef production is finished, the new tag will be allocated with the relation to the previous tag from step (1). (3) and (4) The information will continue track from while the beef has been shipped or delivered from the warehouse. (5) The customers will be able to see the tracking data using the QR code and website.

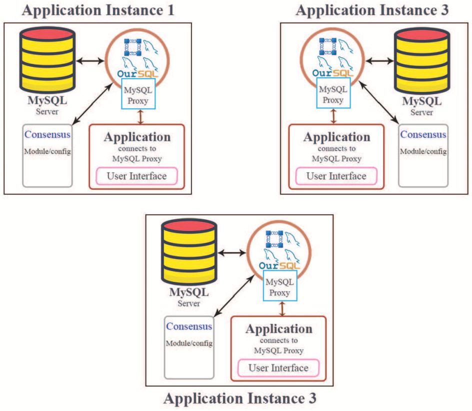


Fig. 3. OurSQL system architecture: It is a no GUI standalone server working on top of mysql server. The key function of OurSQL is to immediately share the updated information when there is an update in the database using Proof of Work (PoW) consensus protocol.

B. *System Architecture*

In the aspect of hardware-oriented design, there are three main sections including IoT devices, server, and smart phone. they are connected together via the Internet. IoT devices sends the location with humidity and temperature to the Raspberry Pi. Then the Raspberry Pi forwards the data to the server. If any problem occurs, the actuator system will notify the user using LED and buzzer. Server install the blockchain database (OurSQL) and automatically receive the data from Raspberry Pi. Mobile application is used to control and monitoring the system. Our system design can be seen in Fig. 4(A). For the service-oriented design, this paper divides the system in to three main parts, including client service, system management service, and cloud service can be seen in Fig. 4(B).

1) Client service: The client service can be divides in to two parts: farmer (seller) and general user. Farmer is an user who can add the product lists and information. Moreover, farmer can query the list of the product and also track the

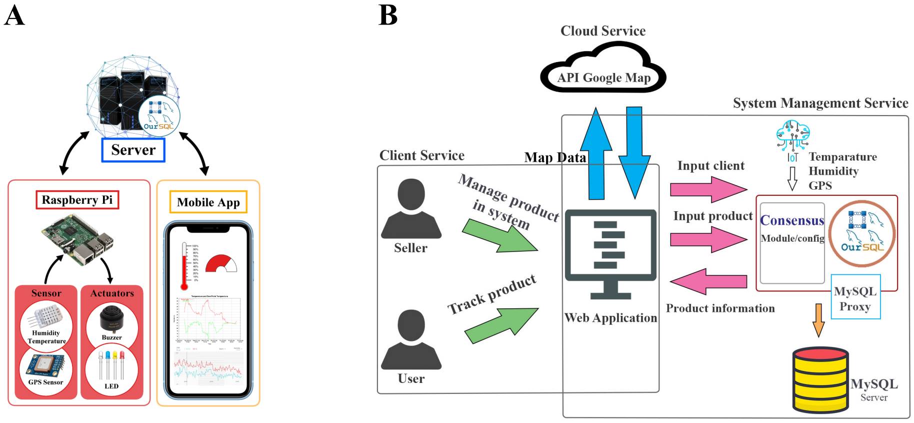


Fig. 4. System design and Overview: In the hardware aspect, this system consists of three main parts: Server, IoT device and Smartphone Application (A). For the service aspect, client service, system management service, and cloud service are three main services need to be separated from each others (B).

product during the shipping process. On the other hand, general user can only see the public information about the product.

2) System management service: It is the control part to manage the website and send/receive the data between the blockchain database and IoT devices.

3) Cloud service: In order to show the real-time tracking system, the Google map API has been used to indicate the coordinate in the map. With this API feature, we are able to track the shipping route in real-time as well as know the specific location of the truck.

IV. Results and Discussion

A. Website and Android Application

After tracking processes finished, our design system automatically generates the QR code for the customer to let the users see the product information. For example, the customer is able to know where the beef come from and when it was first cultivated. With this information, the seller can confirm the quality of the product. We are now successfully created the website and developed the android application in order to show the real-time product information, including location, temperature and humidity.

The website shows the product information from the blockchain database. The location has automatically pinned on the Google map by using the API. Humidity and temperature can be shown during the whole distribution process. The website can be seen in Fig. 5. An android application shows the dash board including virtual temperature and humidity gauges. The status of the system is shown as an alarm on/off. In an ordinary case, the status is shown as an alarm off. However, when a problem is detected, the system status is changed to alarm on. The alarm status and dashboard on the android application can be seen in Fig. 6.

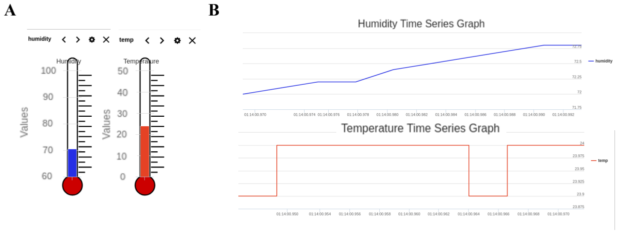
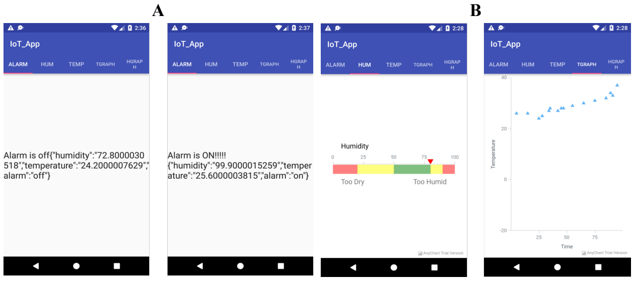


Fig. 5. Website illustrates the humidity and temperature and showed as virtual gauge (A) and time series graph (B).



Alarm is off{"humidity":"72.8000030518","temperature":"24.2000007629","alarm":"off"}

Alarm is ON!!!!! {"humidity":"99.9000015259","temperature":"25.6000003815","alarm":"on"}

Fig. 6. Android application has two main features: alarm status (A) and time series graph and virtual gauge (B).

B. Blockchain Enabled Traceability System

1) Security of the Database: As mentioned in Section II, data collection methods by the IoT have been applied for a number of studies in order to solve the human errors. However, the current traceability system without the blockchain integration cannot guarantee the data integrity because hackers still can change the data in the database [26]. There are many attempts which provide authentication procedure to increase data integrity, but these procedures might not be an appropriate solution to these issues and concerns [10], [26]. In the ordinary authentication procedure, it works to avoid the authentication created by the hackers. Unfortunately, the data can also be changed by the users who able to authenticate themselves to the system [10].

So, the traceability systems with the blockchain integration have recently proposed by many researchers. The blockchain technology will increase the data integrity collected by the IoT devices [28]–[31]. The most popular blockchain technology is the Ethereum, which is a public blockchain. Nevertheless, it automatically allows any user in its network to see the data stored in ledger. As a result, the sensitive data cannot be stored in this kind of the database. Additionally, using a large-scale network dramatically consumes a lot of time and cost for block creation process.

The implementation of OurSQL, which is a blockchain database, confirm the information security of our system by using the proof of work algorithm in order to create the transactions and the blocks. Therefore, our proposed methods can work with the sensitive data with reasonable time consuming and no block creation fees is needed. This is because OurSQL is a controllable database technology that will allow only selected users to use the system and users do not need to pay for the block creation fees.

2) Query layer for the data stored in the blockchain: In order to store the blockchain to hard disk, the data is required to be compressed. This is meant that these valuable data sets are extremely hard to be reused or shown to the users [36]. Moreover, the normal blockchain database is not easy to query the information because some specific commands based on each blockchain technology are required to query the information [34], [35]. This problem can be solved by implementation of the SQL database with the blockchain technology [7], [36]. In this work, we integrated OurSQL, an efficiency blockchain database, with the IoT data collection method in order to implement the traceability system and show the real-time information via both website and android application. With our proposed database, the website and application development can be finished with the use of SQL commands while using a very secured blockchain replication database as mentioned in subsection III-A.

V. Conclusions and Future Work

In this paper, we have integrated the OurSQL with the IoT real-time data collection. To compare with the traditional system (without the blockchain integration), the use of OurSQL blockchain provides unchangeable data when the data stored into the database. In the aspect of blockchain traceability system, our proposed system can utilize the use of SQL commands in the website and application development part. Moreover, OurSQL, controllable blockchain database can work with the sensitive data with a faster query time and no block creation fees is needed. These are our significant improvement points against the traditional methods both with and without the blockchain integration. Our system can check the temperature and humidity of the product in real-time by using website or android application. Moreover, the users can get the notifications when our system found some problems related to the temperature and humidity values. In the future, we plan to add more sensors in order to get more information to ensure the quality of the products. To use all blockchain features, future research can further integrating the traceability system with Hyperledger blockchain technology, which is a permissioned blockchain. This implementation should continue using the SQL database in order to allow the user to use SQL commands with permissioned blockchain. In this permissioned blockchain, the control layer runs on top of the blockchain can differentiate the actions that performed by each user. Therefore, permissioned blockchain has a better transaction performance because we can set the block size limitation and the validate information by adjusting the Chaincode during implementation process. Within this type of blockchain, we can assure the security level of the system because every user has the different priority status. In addition, the smart contract is also recommended in order to avoid the middlemen problem. The deploying of smart contract can be used over many production steps. For example, we can use the smart contract to set up the price and the product quality for both farmers and consumers. Therefore, the farmers are forced to produce their products as good as they can to reach the standard and get the price as mentioned in the contract.

Acknowledgment

We would like to thank Roman Gelembjuk for the patient guidance and advice regarding OurSQL integration he has provided throughout this project development.

References

[1] N. Poapongsakorn, M. Ruhs, and S. Tangjitwisuth, “Problems and outlook of agriculture in thailand,” *Thailand Development Research Institute Quarterly Review*, vol. 13, 01 1998.

[2] N. Chomchalow, *Agricultural development in Thailand*. Dordrecht: Springer Netherlands, 1993, pp. 427–443.

[3] L. U. Opara, “Traceability in agriculture and food supply chain: A review of basic concepts, technological implications, and future prospects,” in *Food Agricultural and Environment*, 2003, pp. 101–106.

[4] S. Ammendrup and L. Barcos, “The implementation of traceability systems,” *Revue scientifique et technique (International Office of Epizootics)*, vol. 25, pp. 763–73, 09 2006.

[5] N. V. Vafiadis and T. T. Taefi, “Differentiating blockchain technology to optimize the processes quality in industry 4.0,” in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, April 2019, pp. 864–869.

[6] F. X. Tian, “An agri-food supply chain traceability system for china based on rfid and blockchain technology,” *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, pp. 1–6, 2016.

[7] M. Muzammal, Q. Qu, and B. Nasrulin, “Renovating blockchain with distributed databases: An open source system,” *Future Generation Computer Systems*, vol. 90, pp. 105–117, 2019.

[8] Feng, Tian, “A supply chain traceability system for food safety based on haccp, blockchain and internet of things,” in *2017 International Conference on Service Systems and Service Management*, June 2017, pp. 1–6.

[9] S. A. Swamy and N. Jayapandian, “Secure bitcoin transaction and iot device usage in decentralized application,” in *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, Oct 2018, pp. 271–274.

[10] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, “A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications,” *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1125–1142, Oct 2017.

[11] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, “A survey on security and privacy issues in internet-of-things,” *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1250–1258, Oct 2017.

[12] W. Hong, Y. Cai, Z. Yu, and X. Yu, “An agri-product traceability system based on iot and blockchain technology,” in *2018 1st IEEE International Conference on Hot Information-Centric Networking (HotICN)*, Aug 2018, pp. 254–255.

[13] B. Bordel Sánchez, R. Alcarria, D. Martín, and T. Robles, “Tf4sm: A framework for developing traceability solutions in small manufacturing companies,” *Sensors*, vol. 15, pp. 29 478–29 510, 11 2015.

[14] S. Rahmadika, B. J. Kweka, C. N. Z. Latt, and K. Rhee, “A preliminary approach of blockchain technology in supply chain system,” in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, Nov 2018, pp. 156–160.

[15] S. Aich, S. Chakraborty, M. Sain, H. Lee, and H. Kim, “A review on benefits of iot integrated blockchain based supply chain management implementations across different sectors with case study,” in *2019 21st International Conference on Advanced Communication Technology (ICACT)*, Feb 2019, pp. 138–141.

[16] T. M. Bhatt and J. Zhang, “Food product tracing technology capabilities and interoperability.” *Journal of food science*, vol. 78 Suppl 2, pp. B28–33, 2013.

[17] B. Adam, R. B. Holcomb, M. Buserc, B. Mayfieldd, J. Thomase, C. A. O’Bryanf, P. Crandallg, D. K. R. Knipei, and S. C. Ricke, “Enhancing food safety , product quality , and value-added in food supply chains using whole-chain traceability,” in *International Food and Agribusiness Management Review*, 2016.

[18] 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)*, Jan 2018, pp. 1–8.

[19] L. B. Campos and C. E. Cugnasca, “Towards an iot-based architecture for wine traceability,” in *2015 International Conference on Distributed Computing in Sensor Systems*, June 2015, pp. 212–213.

[20] W. Liang, J. Cao, Y. Fan, K. Zhu, and Q. Dai, “Modeling and implementation of cattle/beef supply chain traceability using a distributed rfid-based framework in china,” *PloS one*, vol. 10, p. e0139558, 10 2015.

[21] K. Wongpatikaseree, P. Kanka, and A. Ratikan, “Developing smart farm and traceability system for agricultural products using iot technology,” in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, June 2018, pp. 180–184.

[22] W. Cao, L. Zheng, H. Zhu, and P. Wu, “General framework for animal food safety traceability using gs1 and rfid,” in *CCTA*, 2009.

[23] G. Angel and A. Brindha, “Real-time monitoring of gps-tracking multifunctional vehicle path control and data acquisition based on zigbee multi-hop mesh network,” in *2011 International Conference on Recent Advancements in Electrical, Electronics and Control Engineering*, 2011, pp. 398–400.

[24] M. U. H. A. Rasyid, E. M. Kusumaningtyas, and F. Setiawan, “Application to determine water volume for agriculture based on temperature amp; humidity using wireless sensor network,” in *2016 International Conference on Knowledge Creation and Intelligent Computing (KCIC)*, 2016, pp. 105–112.

[25] K. Wongpatikaseree, P. Kanka, and A. Ratikan, “Developing smart farm and traceability system for agricultural products using iot technology,” in *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, June 2018, pp. 180–184.

[26] A. H. Ngu, M. Gutierrez, V. Metsis, S. Nepal, and Q. Z. Sheng, “Iot middleware: A survey on issues and enabling technologies,” *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 1–20, Feb 2017.

[27] S. Thattapon and H. Scott C.-H., “Enhancing voip security and efficiency using vpn,” in *2019 International Conference on Computing, Networking and Communications (ICNC)*, Feb 2019, pp. 180–184.

[28] M. Samaniego and R. Deters, “Blockchain as a service for iot,” in *2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 2016, pp. 433–436.

[29] M. Kim, B. Hilton, Z. Burks, and J. Reyes, “Integrating blockchain, smart contract-tokens, and iot to design a food traceability solution,” in *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Nov 2018, pp. 335–340.

[30] M. Singh, A. Singh, and S. Kim, “Blockchain: A game changer for securing iot data,” in *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, 2018, pp. 51–55.

[31] A. Reyna, C. Martín, J. Chen, E. Soler, and M. Díaz, “On blockchain and its integration with iot. challenges and opportunities,” *Future Generation Computer Systems*, vol. 88, pp. 173–190, 2018.

[32] K. Salah, N. Nizamuddin, R. Jayaraman, and M. Omar, “Blockchain-based soybean traceability in agricultural supply chain,” *IEEE Access*, vol. 7, pp. 73 295–73 305, 2019.

[33] M. P. Caro, M. S. Ali, M. Vecchio, and R. Giaffreda, “Blockchain-based traceability in agri-food supply chain management: A practical implementation,” in *2018 IoT Vertical and Topical Summit on Agriculture - Tuscany (IOT Tuscany)*, May 2018, pp. 1–4.

[34] R. Adams, G. Parry, P. Godsiff, and P. Ward, “The future of money and further applications of the blockchain,” *Strategic Change*, vol. 26, no. 5, pp. 417–422, 2017.

[35] E. Karafiloski and A. Mishev, “Blockchain solutions for big data challenges: A literature review,” in *IEEE EUROCON 2017 -17th International Conference on Smart Technologies*, July 2017, pp. 763–768.

[36] Y. Li, K. Zheng, Y. Yan, Q. Liu, and X. Zhou, “Etherql: A query layer for blockchain system,” in *Database Systems for Advanced Applications*, S. Candan, L. Chen, T. B. Pedersen, L. Chang, and W. Hua, Eds. Cham: Springer International Publishing, 2017, pp. 556–567.

Mobile Agent Platform based Wallet for Preventing Double Spending in Offline e-Cash

Irwan*[1], Armein Z. R. Langi[2], Emir Husni[3]
School of Electrical Engineering and Informatics
Institut Teknologi Bandung,
Bandung, Indonesia

***Abstract*—Electronic cash (or e-cash) research has been going on for more than three decades since it was first proposed. Various schemes and methods are proposed to improve privacy and security in e-cash, but there is one security issue that less discussed mainly in offline e-cash, namely, double-spending. Generally, the mechanism to deal with double-spending in offline e-cash is performing double-spending identification when depositing the coin. Even though the mechanism is successful in identifying double-spender, but it cannot prevent double-spending. This paper proposes the Mobile Agent Platform based Wallet (MAPW) to overcome the double-spending issue in offline e-cash. MAPW uses the autonomy and cooperation of agents to give protection against malicious agent, counterfeit coin and duplicate coin. This model has been verified using Colored Petri Nets (CPN) and has proven to be successful in preventing double-spending, and overcoming malicious agent, and counterfeit coins.**

Keywords*—*e-Cash; double-spending; MAPW; CPN

I. Introduction

Nowadays, electronic payment has proliferated along with the use of the Internet. The electronic payments can be classified into four categories: online credit card, electronic check, smart cards based electronic payment, and e-cash [1]. E-cash is the only electronic payment that provides not only security but also the privacy of its users. E-cash generally consists of three types of entities: bank, user, and merchant. The user withdraws coins from the bank and spends it on the merchant who then deposits the coins to the bank. The security aspect of e-cash should cover some main properties (1) *unforgeability*: no user can create coin other than the authorized party; (2) *no framing*: no one except the owner of a coin can spend it; and (3) *double-spending prevention*: coin can only be spent once. The privacy aspect covers the *anonymity of users*, which mean no one can discover the true identity of a user correlated with withdrawal or spending transaction.

Following the first e-cash scheme introduced by Chaum [2], [3], many e-cash schemes [4], [5], [6], [7] have been proposed, and most of them focus on improving privacy and security (unforgeability and no framing). These proposed e-cash schemes only provide double-spending identification to overcome double-spending. Double-spending is a security issue in which the same coin can be spent more than once since an e-cash's coin is a set of digital data that can be duplicated easily. Double-spending identification is a mechanism to identify whether a coin is a duplicated or not. If the coin is identified as duplicated coin, e-cash system revoke anonymity and discover the identity of the duplicated coin owner. The double-spending identification successfully discovers the identity of double-spender, but it cannot prevent double-spending in advance, especially for offline e-cash scheme.

Several existing proposed methods to fulfill the property of double-spending prevention such as blockchain [8], smartcard [9], and mobile agent [10]. Nakamoto proposed the use of blockchain in a peer-to-peer e-cash system to prevent double-spending [8]. Blockchain is a global ledger where all transaction recorded. Every transaction must be broadcasted to all nodes and added to the ledger. Because all nodes keep this ledger, it is impossible to perform double-spending. The blockchain-based method is comparatively slow in transaction speed and confirmation. As reported by Statista, average confirmation time of bitcoin, which is one of e-cash scheme that implements the blockchain method, is 9.47 minutes[1].

In order to manage the double-spending problem in offline e-cash scheme, Liu proposed a method that uses a smartcard that records a pair of all withdrawn coins [9]. When a customer spends a coin, a merchant requests the pair of the spent coin to the smartcard. Smartcard searches the pair to prove that the coin has not been spent yet. If the pair exists in the smartcard, merchant accepts the coin while smartcard deletes the pair of the coin. The customer cannot spend the same coin because a pair of the coin no longer exist in the smartcard. However, this method does not provide any mechanism to prevent a pair of the spent coin rewritten in the smartcard.

Furthermore, Salama proposed a more advanced method in against the double-spending problem by using Optical Memory Card (OMC) and mobile agent [10]. OMC is a write-only card that used for recording the serial number of the spent coin. The mobile agent is used as a coin that can identify the spent coin which its serial number has been recorded in OMC. Hence, customer cannot spend the same coin if the serial number of the coin recorded in OMC. This method can prevent double-spending in advance but has limitation in OMC memory. When OMC has no memory space, the double-spending prevention capability cannot be performed.

From the above analysis, the existing methods have not been able to prevent double-spending optimally. These methods still have open issues in the slow confirmation time, the inability to counter data of spent coin to be copied on the smartcard and limited data storage space on OMC. Double-spending causes financial losses, so this issue is paramount to

[1] Average confirmation time of Bitcoin transactions from June 2017 to June 2018 according to https://www.statista.com/statistics/793539/bitcoin-transaction-confirmation-time / accessed 31 July 2019

resolve. Thus, we need to construct a method that meets the double-spending prevention property and prove the security of the method. This paper proposes the Mobile Agent Platform based Wallet (MAPW) model that is not only able to deter counterfeit coin and prevent double-spending but also protect the mobile agent platform from malicious agents.

This paper is organized as follows. In Section 2, we describe the preliminaries on offline e-cash and agent's technology in e-cash. Overview of the proposed MAPW model is presented in Section 3. In Section 4, the Colored Petri Nets (CPN) model of MAPW is described, and the analysis of security is presented in Section 5. Finally, Section 6 concludes the proposed model and Section 7 gives pointers to future work.

II. Preliminaries

This section introduces some concepts related to offline e-cash system and the use of agent's technology in the e-cash.

A. Overview of Offline e-Cash

The prevalent model of the e-cash schemes involved three different parties, namely a bank, customers, and merchants. The life cycle of e-cash coin involves these three parties as given in Fig. 1. This life cycle begins when a customer withdraws the coin from the bank (withdrawal protocol). Then, the customer spends the coin by sending it to a merchant in trading for some goods or services (payment protocol). Finally, the merchant ends the cycle by depositing coin (deposit protocol). There are two types of e-cash schemes, namely, online and offline.

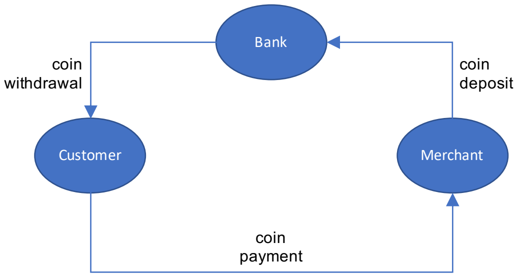
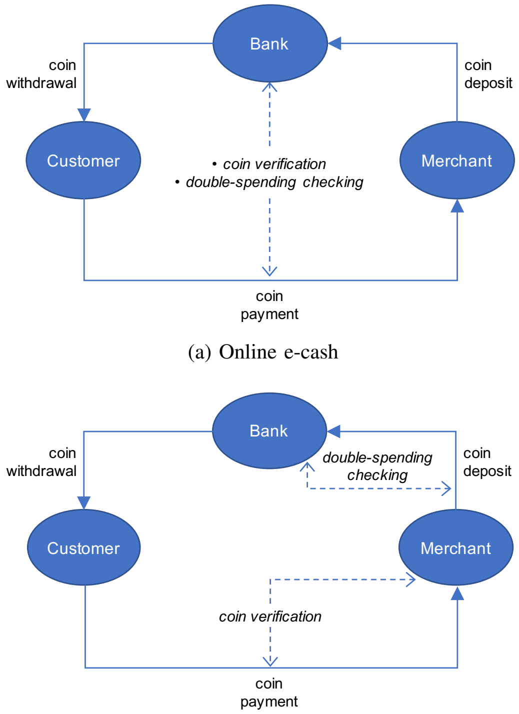


Fig. 1. The circulation of e-cash coin.

Both online and offline e-cash schemes perform two major verifications, namely coin's validity verification and identification of the coin's double-spending. In an online e-cash scheme [11], as illustrated in Fig. 2a, the merchant has to be online with the bank when performing both verifications on the payment protocol. The coin from the payment is accepted if the coin is valid and never used before. While in offline e-cash scheme [12], [13], [14], as illustrated in Fig. 2b, the merchant accepts a coin on payment protocol and subsequently verifies the validity of the coin without involving the bank. The bank identifies double-spending of a coin on deposit protocol.

The potential for dishonest customers to double-spend coins is higher in offline e-cash since the coins are not verified at the coin payment protocol. Therefore, an offline e-cash scheme must have a mechanism to prevent double-spending. Besides, e-cash also must be resistant to counterfeit coin and adversary users.



(a) Online e-cash

(b) Offline e-cash

Fig. 2. Coin verification in traditional e-cash

B. Agent's Technology in e-Cash

An agent is a computer program that is able to take independent action on behalf of its user or owner. Generally, the agent is categorised into two types: static and mobile agent. The static agent always stays in one place and performs the operations according to its intended purposes. The mobile agent is the one with mobility capability that allows it to migrate from one host to another to perform the operations for its owner.

Within the last decade, the paradigm of the agent system was discussed broadly. Many suggestions for future fields of application of the agent system have been made in a distributed system such as distributed database [15] and digital library [16]. In the e-cash system, the paradigm of the agent system address the challenges of communication bottleneck and resource ability of a user. There are two main research fields about e-cash and agent. The first field concern of agent is an e-commerce framework, including the notion of payment [17]. Furthermore, the possibility of an agent to carry and spend e-cash is the second field [10], [18].

The agent technology adoption in many fields led to sophisticated design and security threats. Since mobile agent migrates from one host to another, mobile agent more vulnerable than the static agent. Thus, the mobile agent platform protection is vital because it is an environment where a mobile agent gets executed. Mobile agent platform suffers security threats from a foreign agent that performs denial of service attack and unauthorized access. Protection of mobile agent platform from a malicious agent can adopt various techniques. The first is sandbox technique which isolates untrusted agent so cannot alter the platform or agent in it [19]. Simple Malicious Identification Police (MIP) model [20] is the second technique that can be adopted to protect the mobile agent platform. The concept of this technique is identifying malicious agent by scanning the byte code of the agent.

III. Proposed Model

The proposed model is the MAPW model for preventing double-spending in offline e-cash scheme. This section gives an overview and description of the proposed model in detail.

A. Overview of Proposed Model

The MAPW model is intentionally designed as a coin's wallet with protection against malicious agent, counterfeit coin, and double-spent coin. Simple MIP model is adopted to protect MAPW against malicious agent. In order to overcome the counterfeit and double-spent coin, MAPW applies the autonomy and cooperation capabilities of the software agent.

The main idea of MAPW model is to append double-spending identification when receiving a coin thus can prevent double-spending in advance. Fig. 3 illustrates the proposed offline e-cash cycle used by MPAW. There are three parties: bank $\mathcal{B}$ that able to issue coins and accept deposited coins; customer $\mathcal{C}$ that can withdraw and spend coins; and merchant $\mathcal{M}$ that can accept spending coins and deposit coin. Our proposed model is composed of withdrawal protocol, payment protocol, and deposit protocol.

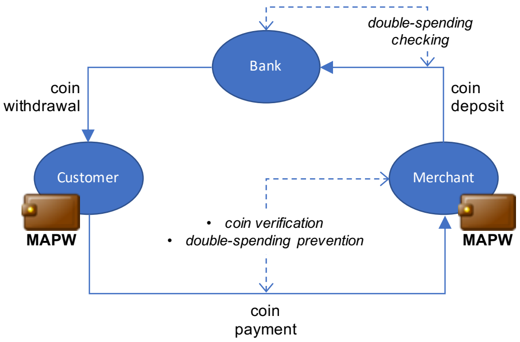


Fig. 3. The proposed offline e-cash cycle.

The simple description of these protocols is given as follows:

1) The withdrawal protocol: The customer $\mathcal{C}$ withdraws a coin c_i from $\mathcal{B}$. If $\mathcal{B}$ and $\mathcal{C}$ pass the challenge-response, $\mathcal{B}$ sends c_i to $\mathcal{C}$. Then, $\mathcal{C}$'s wallet verifies the signature of c_i and performs the synchronization checking.

2) The payment checking: The customer $\mathcal{C}$ spends a coin c_i to $\mathcal{M}$. At first, both of them perform challenge-response and c_i will be sent to $\mathcal{M}$ if the challenge-response is performed successfully. The $\mathcal{M}$'s wallet subsequently performs the verification of c_i's signature and the synchronization checking.

3) The deposit checking: The merchant $\mathcal{M}$ can deposit coin c_i to the bank $\mathcal{B}$. Before depositing c_i, $\mathcal{M}$ and $\mathcal{B}$ perform challenge-response. $\mathcal{B}$ allows $\mathcal{M}$ to send c_i if they pass the challenge-response. First, $\mathcal{B}$ verifies the signature of c_i and checks whether c_i has been previously deposited. If the signature of c_i is valid and c_i is never deposited before, $\mathcal{B}$ accepts c_i.

4) The synchronization checking: The purpose of the synchronization checking is preventing the wallet from receiving duplicate coin. The synchronization is performed when the wallet of $\mathcal{C}$ or $\mathcal{M}$ receives or sends a coin. The wallet checks the identity of the new coin whenever it receives a new coin. The new coin is accepted if there are no coins in the wallet that has the same identity as the new coin. Otherwise, the wallet refuses the new coin. Before the wallet sends a coin, the wallet checks the existence of coin's identity. If coin's identity exists, it allows the coin to migrate to another wallet.

B. Model Description

MAPW model, as shown in Fig. 4, has four static agents (like user, bank, identifier, and killer agent) and one mobile agent (coin agent). The static agents, with their respective duties and responsibilities, protect the mobile agent platform, and ensuring no counterfeit and double-spent coin. User agent performs three e-cash protocols (withdrawal, payment, and deposit), and is responsible for incoming and outgoing checking. Bank agent is an agent of the bank's representative that stores the identity of all coin agent in the wallet and identifies a duplicate agent. Identifier agent determines whether the foreign agent is a malicious agent or not. Killer agent kills any malicious agent, counterfeit coin, and double-spent coin.

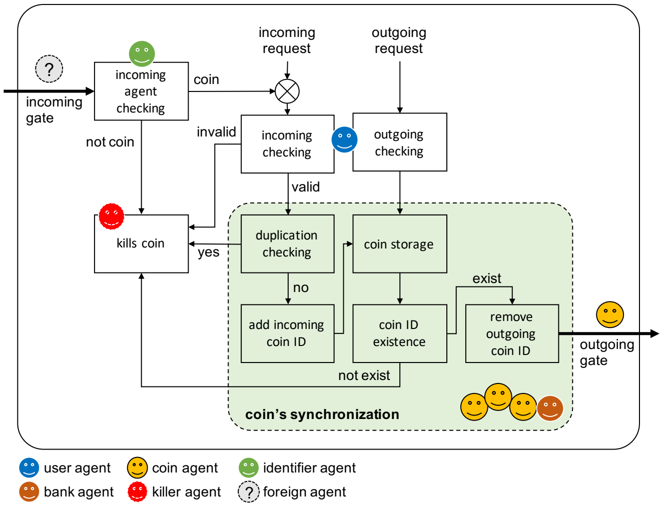


Fig. 4. The model of MAPW.

Like bank agent, coin agent stores the identity of all coin in the wallet and identifies the duplicate agent. The coin agent, as shown in Fig. 5, carries coin data like serial number, signature, origin, and sync ID. *Serial number* is a unique number that represents the identity of the coin. *Signature* is a bank's digital signature for verifying the authenticity of the coin. A flag that indicates whether the coin comes from a valid protocol or not is called *origin*. *Sync ID* is a memory space to store the identity of the bank and other coin agents that are in the same wallet at the same time.

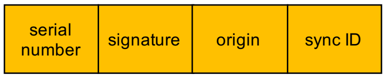


Fig. 5. The block data of coin.

C. The Function of Mobile Agent Platform based Wallet Model

The functional process algorithm of MAPW model involves two algorithms: the arrival and leaving of a coin that is respectively given in Algorithm 1 and Algorithm 2. In Algorithm 1, the arrival of a new agent triggers the identifier agent to identify the new agent. If the new agent is not identified as a coin agent, the identifier agent will trigger the killer agent to kill the new agent. Otherwise, the new agent will be considered as a new coin and forwarded to the user agent for verifying the new coin's signature. The new coin is allowed to broadcast its arrival to the bank agent and stored coins if its signature is valid, but the killer will kill it if its signature is invalid. After the bank agent and stored coins receive the broadcast message, they check the existence of the new coin ID in their memory of the stored coins' ID and send a kill command to the killer agent if its ID is a duplicate. However, the bank agent and stored coins save the new coin ID if the new coin is not a duplicate. The new coin also saves the bank agent and all stored coins' ID.

Algorithm 1 Algorithm for the arrival of agent/coin

```
if the agent is not a coin or request is not a valid incoming request then
    killer agent kills the agent and exit;
else
    the agent is considered as a new coin;
    user agent verifies the new coin's signature;
    if the signature of the new coin is invalid then
        sends a command to killer agent to kill the new coin;
    else
        the new coin broadcasts its arrival to stored coin and bank agent;
        the bank agent and stored coins check the new coin ID whether its already exist or not;
        if the ID of new coin is a duplicate of stored ID coin then
            the bank agent or stored coins send a kill command to killer agent;
        else
            the bank agent and stored coins save the new coin ID;
            the new coin ID saves all stored coin ID and bank agent;
        end if
    end if
end if
```

The leaving of a coin, as described in Algorithm 2, begins whenever the user agent accepts a valid outgoing request. The user agent sends a request to a coin for migrating to another wallet. The coin that accepts this request then broadcasts its migration to the bank agent and other coins. They delete the coin's ID from their memory and allow the coin's migration if the coin's ID is in their memory. Otherwise, they consider the coin as an invalid coin and trigger the killer agent to kill the coin.

Algorithm 2 Algorithm for the leaving of a coin

```
if request is a valid outgoing coin request then
    the user agent sends a request to the coin for migrating to another wallet;
    the coin broadcasts its migration;
    if the bank agent and other coins know the ID of the coin then
        other coins agent and bank agent delete ID of the coin;
        the coin migrates and deletes the ID of other coins;
    else
        killer agent kills the coin agent and exit;
    end if
else
    ignores request;
end if
```

IV. Colored Petri Nets Model of Mobile Agent Platform based Wallet

MAPW is the proposed model of double-spending prevention in offline e-cash. In order to determine the correctness and eliminating or minimizing the security of MAPW, it must be verified by using a formal method. There are various formal methods, but the most commonly used for the agent is Petri nets. For example, Petri nets can be used for modeling interaction protocol in multiagent system [21] and for verifying agent-based architecture [22]. This paper uses CPN, that is a combination of the capabilities of Petri nets and a high-level programming language, for the design, development, and analysis of MAPW [23].

TABLE I. Tested scenarios for MAPW model

Case	Agent type	Condition of coin			
		Signature	Legitimate origin	Duplicate	Known ID
malicious agent	not coin	-	-	-	-
counterfeit coin	coin	invalid ($sign$=0)	-	-	-
double spending	coin	valid ($sign$=1)	invalid ($orig$=0)	-	-
		valid ($sign$=1)	valid ($orig$=1)	yes	-
		valid ($sign$=1)	valid ($orig$=1)	no	no
normal spending	coin	valid ($sign$=1)	valid ($orig$=1)	no	yes

Table I shows a set of the tested scenario for MAPW model in proving the MAPW's protection against malicious agent, counterfeit coin, and double-spending. MAPW also should able to do normal spending. A malicious agent is an agent that its type is not a coin agent. A counterfeit coin is a coin with an invalid signature ($sign$=0). There are three possibilities of double-spending. First, a coin with a valid signature but not came from the legitimate origin (valid withdrawal or valid payment). Second, coin with valid signature and came from legitimate origin but has a duplicate in the wallet. Third, a coin with a valid signature, came from legitimate and does not have a duplicate in the wallet, but its identity is not recognized. The last scenario is normal spending that is a coin with a valid signature, came from the legitimate origin, does not have any duplicate, and its identity is recognized by bank agent and coin agent.

The CPN model of MAPW consists of one main MAPW model page and four subpages for coin's generation, incoming coin, outgoing coin, and synchronization coin's ID. Fig. 6 illustrates the main MAPW model page in which accepts the incoming and outgoing request. Every time MAPW accepts incoming coin request, it will trigger coin generation to generate random coin, and this random coin will be checked by incoming coin checking. If the request is an outgoing coin request, it will be checked by outgoing coin checking.

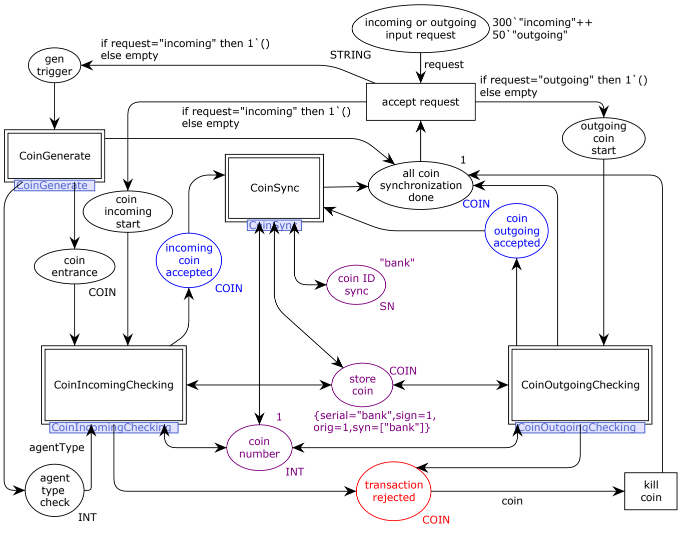


Fig. 6. Main CPN model of MAPW

There are four customized color types and five customized functions in the CPN model of MAPW. Fig. 7 shows the declaration of the four customized color types, namely, $FLAG$, SN, SYN, and $COIN$. $FLAG$ is an integer color type with a value of 0 or 1. SN is a string color type that represents serial number of a coin. SYN is a color type of the list of SN that represents the memory of a coin. $COIN$ is a record color type that consists of $serial$, $sign$, $orig$, and syn.

```
colset FLAG=int with 0..1;
colset SN=STRING;
colset SYN=list SN;
colset COIN=
    record serial:SN*sign:FLAG*orig:FLAG*syn:SYN;
```

Fig. 7. Declaration of CPN model color types

The five customized functions, as shown in Fig. 8, are $count()$, $notin()$, $serialVal()$, $boolVal()$, and $serialLabel()$. The $count()$ function is used for counting data in a list. The $notin()$ function is used for searching serial number of a coin in a list. The $serialVal()$ function returns a random integer value from 1 to 500. The $boolVal()$ function returns a random number of 0 or 1. The $serialLabel()$ function returns a random serial number of a coin.

```
fun count(synCoin:SYN)=
    if synCoin=[] then 0
    else 1+count(tl(synCoin));
fun notin(sn:SN,syn:SYN)=
    if syn=[] then true
    else if sn=hd(syn) then false
    else notin(sn,tl(syn));
fun serialVal()=
    discrete(1,500);
fun boolVal()=
    discrete(0,1);
fun serialLabel()=
    "serial"^Int.toString(serialVal());
```

Fig. 8. Declaration of CPN model function

A. *CoinGenerate* Subpage

The *CoinGenerate* subpage illustrated in Fig. 9 is the first subpage of MAPW's top-level CPN model and the CPN model of the identifier agent that is responsible for checking all incoming agent. This subpage performs the generation of a random incoming agent (coin and non-coin agent) that enters MAPW through the incoming gate, which is triggered by *gen trigger* place. Coin agent is represented by 1, while a non-coin agent is represented by 0. Random value 0 or 1 is generated by $boolVal()$ function. If a non-coin agent enters MAPW, it will be killed, and the model will return to wait for a request. However, if the incoming agent is a coin agent, it will generate a coin agent and send the coin agent to coin entrance.

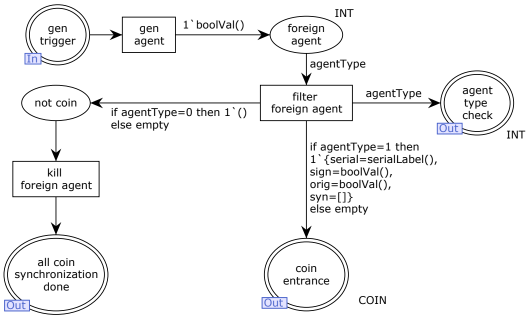


Fig. 9. CPN model of the *CoinGenerate* subpage

Coin data on this model uses $COIN$ color type $\{serial, sign, orig, syn\}$ that consists of $serial$ as the serial number or identity of the coin, $sign$ as the signature of the coin, $orig$ as the origin of the coins, and syn as a storage memory of all coins in the wallet. The value of $serial$, $sign$, and $orig$ are generated randomly in order to allow all conditions of the coin that enter MAPW either valid or invalid. The invalid coin is a duplicate, false signature or the entrance of coin without incoming request. $serialLabel()$ and $boolVal()$ function subsequently generates a random serial number of a coin, and random value of 0 or 1. The signature of the coin is valid if the value is 1 and invalid if the is 0.

B. *CoinIncomingChecking Subpage*

Fig. 10 illustrates the *CoinIncomingChecking* subpage in more detail. *Agent type check*, *coin incoming start*, and *coin entrance* are provided as input. *CoinIncomingChecking* will be executed if a coin agent enters MAPW that is marked by value 1 of *agentType*. The data of coin agent that enters *CoinIncomingChecking* is generated by *CoinGenerate* subpage. Then this model verifies signature and flag of the coin agent whether valid or not. If the signature and flag are valid, the coin will be considered as an incoming coin and start the scanning process. Otherwise, the transaction is rejected, and the incoming coin agent is going to be killed.

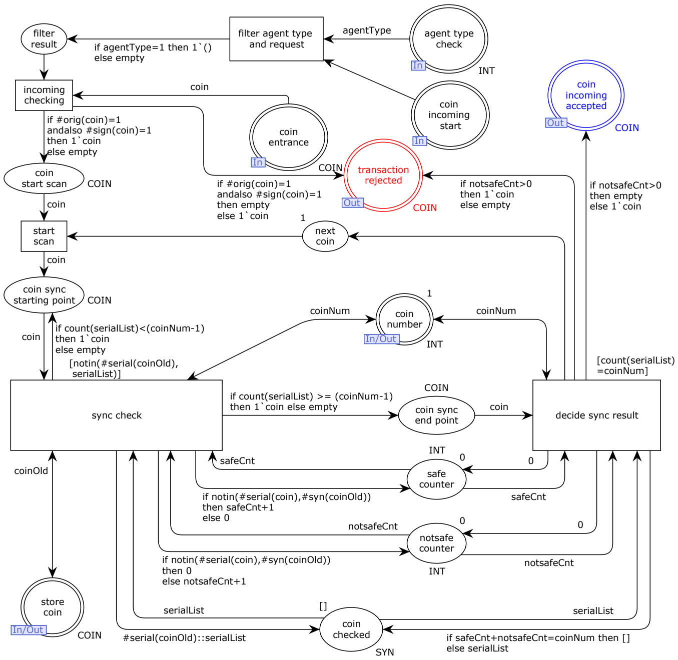


Fig. 10. CPN model of the *CoinIncomingChecking* subpage

The purpose of the scanning process in *CoinIncomingChecking* is as duplicate coin checker. All stored coin agent and bank agent in MAPW check the serial number of the new coin, and if they already have the serial number, then the new coin is rejected. Otherwise, the serial number of the new coin is fresh, the new coin is accepted and forwarded to *CoinSync* transition.

C. *CoinSync Subpage*

The function of *CoinSync* subpage is synchronizing the serial number of incoming and outgoing coin, as illustrated in Fig. 11. This function is triggered when *incoming coin accepted* or *outgoing coin accepted* fire a COIN token.

The synchronization of an accepted incoming coin begins with broadcasting the serial number of the incoming coin agent to bank agent and all stored coin agents. Bank agent and all stored coin add the serial number of the incoming coin agent to their *sync* variable. The incoming coin agent also adds the serial number of bank agent and all stored coin agent to its *sync* variable. Then incoming coin agent is accepted and adds the value of *coin number* place with 1.

The purpose of synchronization of an outgoing coin is removing the outgoing coin agent's serial number and decreasing *coin number* place value by 1. This synchronization starts by broadcasting the outgoing coin to all stored coins agent and bank agent to remove the serial number of the outgoing coin agent. Terminate synchronization and return to wait for a request.

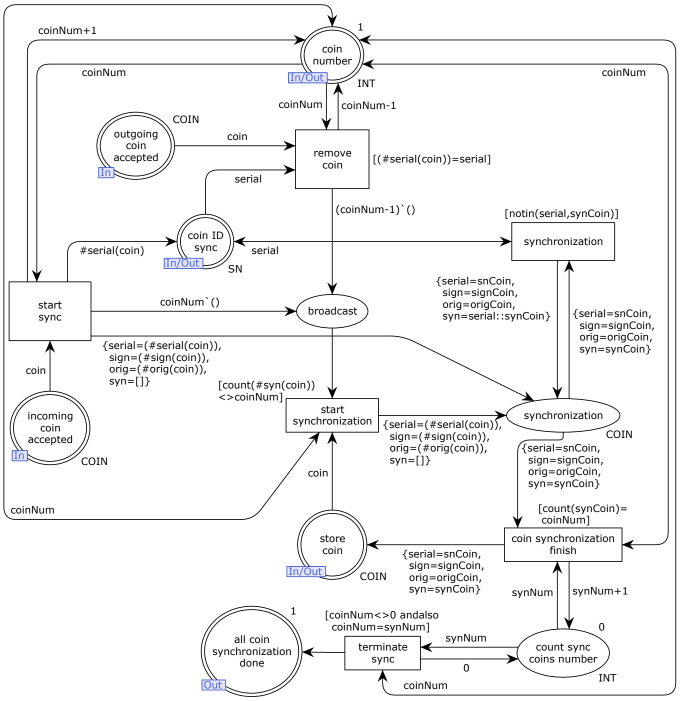


Fig. 11. CPN model of the *SynchronizationProcess* subpage

D. *CoinOutgoingChecking Subpage*

The *CoinOutgoingChecking* subpage, as illustrated in Fig. 12, serves outgoing request which asks for sending a coin agent to a new wallet. This subpage checks the availability and validity of an outgoing coin. The first checking, availability of coins checking, is performed to determine if there are coins stored in the wallet when receiving an outgoing request. If there are no coins, the outgoing request will be ignored. Oppositely, a coin will be called in order to send it to a new wallet.

The subsequent checking is the validity of coin checking, which verifies whether other coins agent and bank agent know the coin's serial number. If other coins agent and bank agent save the serial number of the coin in their sync variable, the coin is accepted as an outgoing coin and forwarded to CoinSync subpage. Otherwise, the outgoing request is rejected, and the coin is forwarded to *kill coin* transition in main CPN model.

V. Discussion

The CPN model of MAPW is verified by performing the state-space analysis that calculates all reducible states and state changes in order to observe the behavior of MAPW model, such as the nonexistence of loops and deadlocks; the

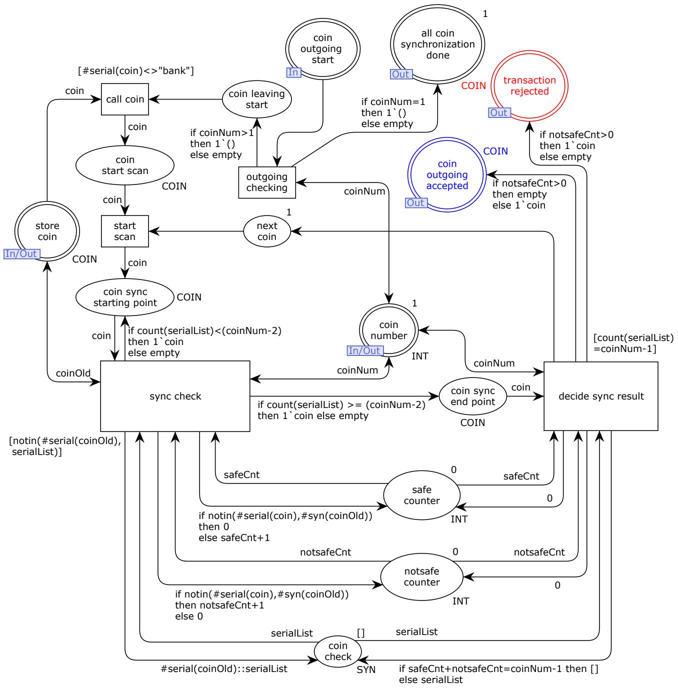


Fig. 12. CPN model of the *CoinOutgoingChecking* subpage

possibility of always being able to reach a certain state; and the delivery guarantee of a provided service. The state-space analysis results in Table II shows that there is no infinite occurrence sequence that indicates there is no loop in MAPW's model so that the termination of each module is guaranteed. There is no home marking, which is mean the impossibility to have an occurrence sequence that cannot be extended to reach the home marking. The state-space analysis detects the marking that has no enabled transition, which is called dead marking. Dead marking exists because not every coin leaves the wallet. Furthermore, the state-space analysis shows the absence of dead transitions and live transitions. The absence of dead transitions means that each transition has the possibility of occurring at least once while the absence of live transition means the transitions always occur in any condition.

TABLE II. State-space Analysis Results of MAPW Model

Property	Result
No infinite occurence sequence	None
Home markings	None
Dead markings	Yes
Dead transitions	None
Live transitions	None

In addition to state-space analysis, the CPN model of MAPW is also tested for its security (malicious agent, counterfeit coin, and double-spending) and functionality (normal-spending) by referring to scenarios in Table I. The result of tested scenarios for MAPW model (Tabel III) shows that the CPN model of MAPW passes all of the tested scenarios.

The properties of MAPW model is compared with two related works [9], [10] and the comparison can be seen in Table IV. Liu [9] and Salama [10] have forgery and double-spending prevention, but they do not have any wallet protection. In order to perform double-spending prevention, Liu [9] uses smartcard and Salama [10] uses OMC and mobile agent. The MAPW model has both forgery and double-spending prevention. Double-spending prevention is performed by mobile agent without depending on the specific hardware. The MAPW model also has wallet protection that protects the wallet from malicious agents.

TABLE III. The Result of Tested Scenarios for MAPW Model

Scenario	Result
malicious agent	detected and killed
counterfeit coin	detected and killed
double-spending	detected and killed
normal spending	pass

TABLE IV. The Comparison of Properties between Related Work and Proposed Model

Property	Proposed model	Liu [9]	Salama [10]
wallet protection	yes	no	no
forgery prevention	yes	yes	yes
double-spending prevention	yes	yes	yes
card based	no	yes	yes
mobile agent based	yes	no	yes

VI. Conclusion

The offline e-cash is vulnerable of double-spending because the bank stay offline while the merchant accepts a coin anonymously from the customer. The merchant only checks the validity of the coin's signature when accepts the coin, but the merchant cannot determine whether the coin is a double-spent coin or not. The bank verifies double-spending after the transaction. This paper proposes the MAPW for offline e-cash that has been modeled, analyzed, verified, and tested by using CPN and tested scenarios. The result shows that the MAPW is able to prevent double-spending, protect wallet against the malicious agent and counterfeit coin.

VII. Future Work

There are various issues in offline e-cash which must be addressed in the future, such as the protection of coin against malicious host problem and the application of MAPW to offline transferable e-cash. The transferability of e-cash is a challenging problem because it has more aspect to consider including more user, and the coin grows in size.

References

[1] S. Singh, "Emergence of payment systems in the age of electronic commerce: The state of art," in *2009 First Asian Himalayas International Conference on Internet*, November 2009, pp. 1–18.

[2] D. Chaum, "Blind signature for untraceable payment," in *CRYPTO '83 Proceedings on Advance in Cryptology*. New York: Plenum Press, 1983, pp. 199–203.

[3] D. Chaum, A. Fiat, and M. Naor, "Untraceable electronic cash," in *CRYPTO '88 Proceedings on Advances in Cryptology*. New York: Springer-Verlag, 1988, pp. 319–327.

[4] S. Canard and A. Gouget, "Anonymity in transferable e-cash," in *Applied Cryptography and Network Security*, S. M. Bellovin, R. Gennaro, A. Keromytis, and M. Yung, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 207–223.

[5] C. I. Fan and V. S. M. Huang, "Provably secure integrated on/off-line electronic cash for flexible and efficient payment," *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 5, pp. 567–579, September 2010.

[6] O. Blazy, S. Canard, G. Fuchsbauer, A. Gouget, H. Sibert, and J. Traore, "Achieving optimal anonymity in transferable e-cash with a judge," *AFRICACRYPT*, pp. 206–223, July 2011.

[7] J. Zhang, H. Guo, Z. Li, and C. Xu, "Transferable conditional e-cash with optimal anonymity in the standard model," *IET Information Security*, vol. 9, no. 1, pp. 59–72, December 2015.

[8] S. Nakamoto, "A peer-to-peer electronic cash system," http://www.bitcoin.org/bitcoin.pdf, 2009.

[9] W. Y. Liu, Y. A. Luo, and Y. L. Si, "A security multi-bank e-cash protocol based on smart card," in *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics*. IEEE, August 2007, pp. 3244–3248.

[10] M. A. Salama, N. El-Bendary, and A. E. Hassanien, "Towards secure mobile agent based e-cash system," in *Proceedings of the First International Workshop on Security and Privacy Preserving in e-Societies*, New York, 2011, pp. 1–6.

[11] S. H. Islam, R. Amin, G. P. Biswas, M. S. Obaidat, and M. K. Kan, "Provably secure pairing-free identity-based partially blind signature scheme and its application in online e-cash system," *Arabian Journal for Science and Engineering*, vol. 41, no. 8, pp. 3163–3176, August 2016.

[12] X. Zhou, "Threshold cryptosystem based fair off-line e-cash," in *Second International Symposium on Intelligent Information Technology Application*, vol. 3. Shanghai: IEEE, 2008, pp. 692–696.

[13] W.-S. Juang, "An efficient and practical fair buyer-anonymity exchange scheme using bilinear pairing," in *2013 Eight Asia Joint Conference on Information Security*, 2013, pp. 19–26.

[14] C. Wang, H. Sun, H. Zhang, and Z. Jin, "An improved off-line electronic cash scheme," in *International Conference on Computational and Information Sciences*. IEEE, 2013, pp. 438–441.

[15] F. U. Ogban and U. Udoh, "A mobile agent-based distributed information retrieval system," *International Journal of Natural and Applied Sciences*, vol. 10, pp. 72–77, 01 2015.

[16] G. Liu, "The application of intelligent agents in libraries: a survey," *Program: Electronic Library & Information Systems*, vol. 45, no. 1, pp. 78–97, 2011.

[17] S. U. Guan, S. L. Tan, and F. Hua, "A modularized electronic payment system for agent-based e-commerce," *Journal of Research and Practice in Information Technology*, vol. 36, no. 2, pp. 67–87, May 2004.

[18] C. Anhalt and S. Kirn, "Towards payment systems for mobile agents," in *Proceedings of the 4th European Workshop on Multi-Agent Systems*, B. Dunin-Keplicz, A. Omicini, and J. Padget, Eds., vol. 223. CEUR, December 2006.

[19] R. Wahbe, S. Lucco, T. Anderson, and S. Graham, "Efficient software-based fault isolation," *ACM SIGOPS Operationg Systems Review*, vol. 27, no. 5, pp. 203–216, 1993.

[20] S. Venkatesan and C. Chellappan, "Protection of mobile agent platform through attack identification scanner (ais) by malicious identification police (mip)," in *2008 First International Conference on Emerging Trends in Engineering and Technology*, July 2008, pp. 1228–1231.

[21] B. Marzougui and K. Barkaoui, "Interaction protocols in multi-agent systems based on agent petri nets model," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 7, pp. 166–173, 2013.

[22] N. A. Mian and F. Ahmad, "Agent based architecture for modeling and analysis of self adaptive systems using formal methods," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 1, pp. 563–567, 2018.

[23] K. Jensen and L. M. Kristensen, *Coloured Petri Nets: Modelling and Validation of Concurrent Systems*, 1st ed. Springer Publishing Company, Incorporated, 2009.

Towards A Proactive System for Predicting Service Quality Degradations in Next Generation of Networks based on Time Series

Errais Mohammed[1]
Research and computer Innovation Laboratory
Hassan II University of Csasablanca
Casablanca–Morocco

Rachdi Mohamed[2]
Faculty of science Ben M'sik
Hassan II University of Casablanca
Casablanca-Morocco

Al Sarem Mohammed[3], Abdel Hamid Mohamed Emara[4]
Department of Information System
Taibah University
Medina-KSA

***Abstract*—The architecture of Next Generation of networks (NGN) aims to diversify the offer of operators in added value services. To do this, NGN offers a heterogeneous architecture for the services deployment. This poses significant challenges in terms of end-to-end assurance of services. For this purpose, we propose in this work the establishment of a proactive autonomous system, capable of ensuring an acceptable quality level according to Service Level Agreement (SLA) requirements. A system that is able to predict any QoS degradation due to the prediction model based on time series adapted to NGN.**

Keywords—Next Generation of Network (NGN); network management; enhanced Telecom Operation Management (eTOM) frameworks; prediction; time series; Ip Multimedia Subsystem (IMS); Service Level Agreement (SLA); Quality Of Service (QoS)

I. Introduction

Next Generation of Networks (NGN) [1] offers a standard architecture for easy integration of services with existing communication technologies. An architecture that aims primarily to facilitate the deployment and provision of value-added services to customers of telecommunications operators [2][3]. This will help to evolve the business of the operator from a simple access provider to an end-to-end service provider.

The 3GPP [4] specifications describing the architecture of NGNs cut the network in three levels. The access level which groups together existing communication technologies. The responsible control level unifies access to networks and services regardless of the underlying technology and the level of service responsible of the unfolding and provision of value-added services.

The diversity of communication technologies and the multitude of integrated entities pose significant challenges for end-to-end assurance of services [5]. To this end, several studies have dealt with the management of QoS in NGN networks [6][7][8]. However, most of this work proposes solutions to correct the level of the QoS after a possible degradation detected. This significantly affects the quality of the QoE user experience. Indeed, the time required for detecting and correcting degradations is very important, given the difficulty of real-time correction operations [9]. This explains the need to reduce the degradation time to ensure the satisfaction of end customers.

In this work, we propose a proactive approach to end-to-end and real-time assurance of services; an approach that will predicts any degradation before sufficient time for correction. To do this we propose a prediction model able to predict the variation of the state of the network and thus to deduce the possible impairments of the QoS, in real time.

This document is organized as follows. At the beginning we will present the NGN architecture and the SLM & M [10] solution which is used for the correction of the degradations. Then we will present the new prediction approach and discuss the model adapted to the NGN context. In the last section of the document we will discuss the results obtained during the experimentation phase in real cases of service provision.

II. Background and Related Works

A. NGN Networks

3GPP specifications dedicated to NGN networks offers a simplified architecture for the core network known by the IP Multimedia Subsystem (IMS) architecture [2]. This architecture aims to simplify the provision of service regardless of the communication technology used for service consumer access. To do this, the IMS offers unified procedures for user authentication and access to services. What is achievable by cutting the architecture at three levels is as follows:

- Access Level: This level is responsible for ensuring the connectivity of users to the control entities described in the 3GPP specifications. It allows to interface effectively with different types of communication technologies via interface components usually installed in the IMS core boundary routers. The access networks are connected to a global network linking the different control entities and the interfacing components.
- The Control Level: This is the key level of the NGN architecture; it ensures the control of users and services deployed by the operator. The level basically includes four entities: The three control entities of the P_CSCF, I_CSCF and S_CSCF sessions and the HSS database.

- Service Level: The level that ensures the deployment of added value services. It groups logical components for negotiation with control entities as well as physical servers.

The strength of the NGN architecture lies in the ease of deployment of services without impacting the access networks deployed by the operator. However, such an organization requires the intervention of several heterogeneous entities to achieve the usual operations including user authentication and service provision. This poses challenges for managing end-to-end quality of service.

B. The SLM and M Solution

The SLM&M solution aims to automate QoS management procedures in NGN networks. An automation that has become possible through the integration of business processes of the eTOM framework [11][12]. The eTOM framework is a set of business processes designed to model the usual operations in telecommunications, in particular the provision, insurance and billing of services.

The SLM&M solution has enabled the implementation of an autonomous system for the monitoring and correction of QoS impairments in NGN networks. This system is capable of estimating customer satisfaction in real time based on the Service Level Agreement (SLA) contract establishes when subscribing the customer to the service provided [6]. If a degradation is detected, the system proceeds to the correction based on preconfigured scenarios, as the case encountered [13].

The autonomous system resulting from the SLM&M solution consists of several modules (Fig. 1). The modules include a set of eTOM business processes. Each process is responsible for a specific activity such as collecting metrics, loading fix configurations, and checking constraints defined in SLA.

Fig. 1 illustrates the system architecture of the SLM & M solution which consists essentially of two levels:

- Monitoring Level: The first level includes business processes that have a global view of the network. These processes are responsible for detecting degradations and identifying configurations to correct the identified case. This level essentially includes three Services, Orchestration and Resource modules. At this level the communication between the modules is based on the SOAP protocol. A justified and validated choice [9] to facilitate the exchanges between the processes and to minimize the time of execution.
- Resource Level: The Resource Level has three modules. These modules are responsible for collecting performance indicators and implementing configurations during the correction. Communication between these modules is provided by the CORBA protocol [14]. A choice that aims to ease trade [9].

The validation of the autonomous system in real cases of service provision made it possible to focus on the limitation of the SLM & M solution. In fact, the monitoring and correction operations make it possible to correct the QoS. However, the correction time is very important. What influences is the quality of user experience. For this purpose, the integration of technical choices to minimize time is not enough. This is explained by the difficulty of the automatic correction operation and in real time. To this end, migration to a proactive system becomes a necessity to ensure a stable level of service assurance at all times.

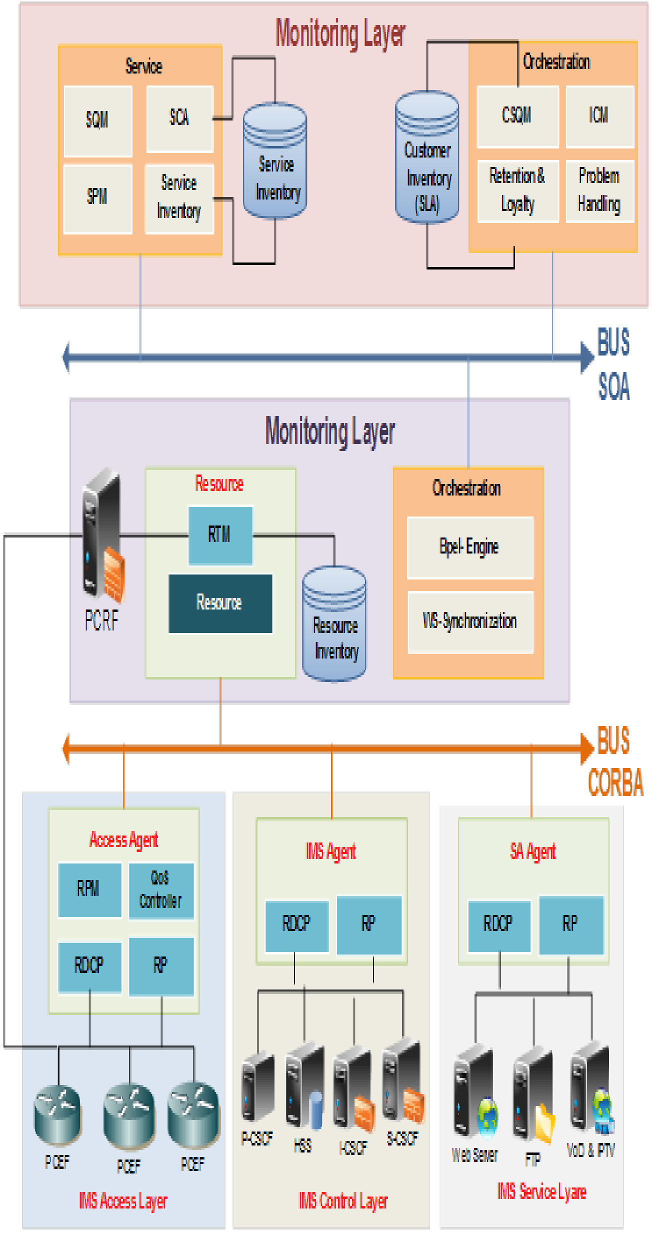


Fig. 1. System Architecture of the SLM and M Solution.

III. Proactive System for Insurance of Services

The goal of the new approach is to migrate to a proactive system that can predict the state of the network. Once the prediction is assured the second step is to run the SLM&M system to proceed with the implementation of the necessary configurations to avoid the degradation. Migrating the SLM & M solution to a proactive system is a difficult operation that requires multiple tasks (Fig. 2).

Fig. 2 outlines the steps required to implement the proactive approach. Steps that look like this:

- Choice of indicators to use: The choice of indicators to be estimated is a key step for the success of the approach. Indeed, the prediction operation is a difficult operation that has a major blow in terms of time and resource.
- Identification of the prediction model: There exists in the literature a multitude of mathematical techniques and models for prediction in different disciplines. The model is the key to the success of the predictive approach, given its effect on the accuracy of the predictions and thus the proactive solution.
- Integration of the model in the SLM & M solution: After identifying the model it is necessary to interface the proactive system with the SLM & M solution. This operation must ensure a transparent interface between the modules to ensure the proper functioning of the system.
- Testing and Validation: The final step is the validation of the evaluation of the approach in real cases of service provision.

A. *Choice of Performance Indicators*

Real-time monitoring is the collection of performance indicators from many resources. These indicators depend essentially on the outstanding service. For this purpose, the most fragile services are those of streaming type whose nature of flows requires regularity over time. The indicators used in this type of service belong to two categories. (i) Static indicators such as the codec used for each stream, the type of video, the type of service (VoD, IPTV) as well as the capacity of the server. (ii) Dynamic indicators such as jitter, delay, percentage of lost packets, and actual throughput.

The multitude of indicators used will undoubtedly affect the prediction time of degradation. For this, we have reduced the performance indicators used for the video stream, since it is the most sensitive to degradation. Indeed, in practical cases the degradation affects the video stream before the audio. Only the dynamic indicators will be estimated, since the static indicators remain unchanged over time. Thus, the performance indicators that will be processed in the model are (i) the jitter for the video stream, (ii) the delay for the video stream; (iii) the percentage of packets lost for the video stream.

B. *Choice of the Prediction Technique*

Several works have focused on the study and implementation of prediction models in different domains [15][16]. Work that offered a wide choice of models to adapt according to the domain and type of data processed. In our context, the idea is to predict future values based on the old values collected in regular time intervals. Also, the estimated values must be done in a future time are sufficient to ensure the correction operation before the actual degradation of the service.

For this purpose, time series [17] are the most appropriate technique for our context. In fact, the time series make it possible to model the values in time at regular intervals. Then these values are modeled by a mathematical model in order to be able to calculate values at times in the near future based on the values recorded in the past times.

The implementation of a prediction model based on time series requires the use of a suitable mathematical method. There are three methods in the literature for predicting values based on time series, in particular:

- Moving Average [18]: An effective method for small change values. IT is to model the series in a linear association with the moving averages of the process
- Exponential Smoothing [19]: Exponential Smoothing is a highly valued tool for predicting and analyzing data from time series observation. A technique used in industry, especially in financial markets, given the simplicity of the models included. Methods that are applicable to any discrete observation set. There are two types of smoothing: (i) simple exponential smoothing, (ii) double exponential smoothing.
- Box-Jenkins model [20]: Box-Jenkins provides an exact methodology for identifying the time series model based on recorded observations. A methodology that is based on several mathematical foundations to lead to a generally powerful prediction model.

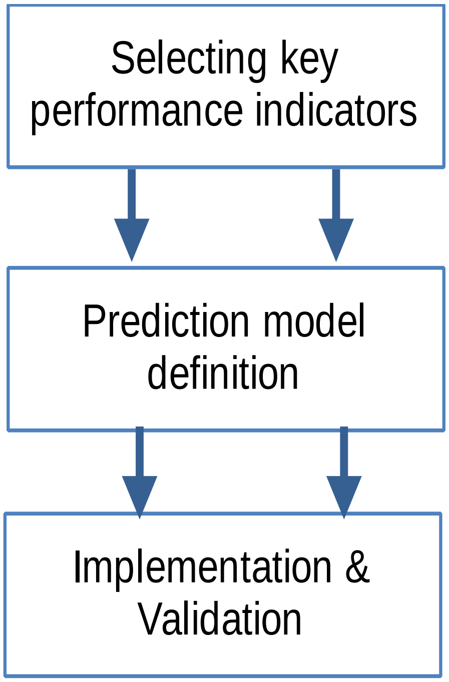

Selecting key performance indicators
Prediction model definition
Implementation & Validation

Fig. 2. Description of the Methodology followed for the Migration to the Proactive System.

Table I illustrates a comparison of methods for time series based on three indicators. The first is the efficiency that reflects the accuracy of the estimates from each method. The second is the cost of predicting each method while the last is the ease of adaptation of the model in the context of NGN. For this purpose, the Box-Jenkins method known also by ARMA is presented as the most efficient solution. However, setting up an ARMA-based model requires a theoretical and experimental study which cannot be done in our context. Since the nature of the flow that changes for each session, which requires the use of a model capable of correcting itself automatically. Unlike the ARMA method, exponential smoothing requires little computation, which will affect the resources less during the prediction operation. In addition, the method of smoothing has demonstrated its effectiveness in various fields. So, our choice was oriented towards exponential smoothing.

C. Identification of the Prediction Model

The implementation of a simple exponential smoothing prediction model consists in adapting the variable α (equation 1) according to the nature of the variation of the values of each indicator [19].

$$\hat{x}_{n,t} = \alpha \sum_{j=0}^{n-1} (1-\alpha)^j \, x_{n\_j} \tag{1}$$

The accuracy of the values estimated in the simple exponential smoothing method depends on the adjustment of the constant which takes values in the interval [0.1]. In the literature the study of the variation of the values makes it possible to adjust the constant values. However, in our context the evaluation of the values of each indicator varies according to the current session (Fig. 3). This forces the search for a self-adaptable model according to the current situation.

Fig. 3 illustrates an example of observing the delay values in three different sessions. Values that record large variations in all three sessions. This difference does not allow the use of a single constant for different sessions. Indeed, it is important to adjust the value of the constant alpha so as to minimize the error between the predicted value and the actual value (Fig. 4).

Fig. 4 illustrates the difference in the optimal value of the constant for each flow in three different sessions. This shows the difficulty of using the conventional method for identification. For this purpose, we propose a model able to justify the value of the constant according to the nature of the current session for the two indicators delay and jitter. Indeed, the experimental study showed that the value of the constant depends on the number of flows in progress in the session (Fig. 5).

In order to propose a model capable of self-justification, we propose to establish a relation between α and the number of flows noted 'N' using linear regression [21]. Before replacing the constant by the formula in the prediction model; following this method, we obtained the following prediction models:

1) *Delay prediction model*

$$\hat{d}_{t+h} = \alpha \sum_{j=0}^{t-1} (1-\alpha)^j \, d_{t\_j} \tag{2}$$

Such as:

$$\alpha = \frac{13}{10000} \cdot N + \frac{1}{4} \tag{3}$$

With: N is the number of competing flows in the current session.

2) *Prediction model for jitter*

$$\hat{g}_{t+h} = \alpha \sum_{j=0}^{t-1} (1-\alpha)^j \, g_{t\_j} \tag{4}$$

Such as:

$$\alpha = \frac{1}{625} \cdot N + \frac{1}{10} \tag{5}$$

With: N- is the number of competing flows in the current session.

TABLE I. COMPARISON BETWEEN MODELING APPROACHES OF TIMES SERIES

Method	***Efficacy***	***The costs of implementing***	***Adaptability***
ARMA	Very Performance	Costly in time and resource	Very difficult
Exponential Smoothing	Performance	Fast	Difficult
Moving Average	Average	Fast	Easy

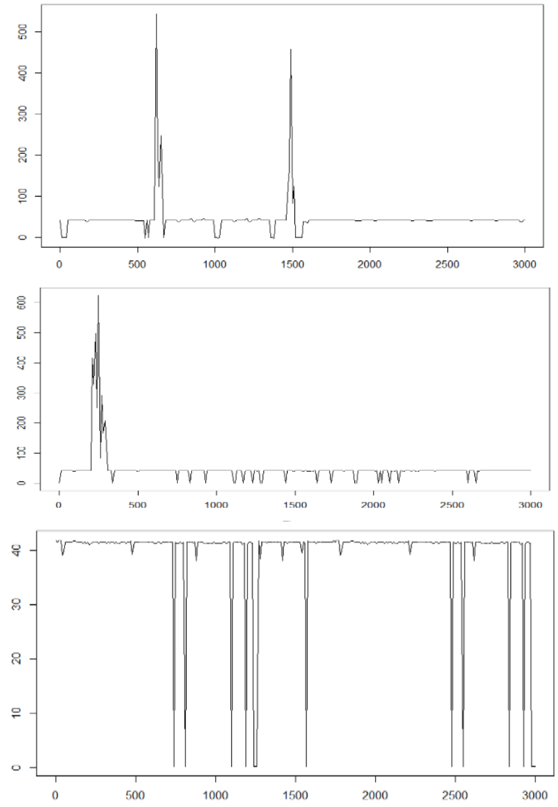


Fig. 3. Example of the Variation of the Delay Values in different Sessions.

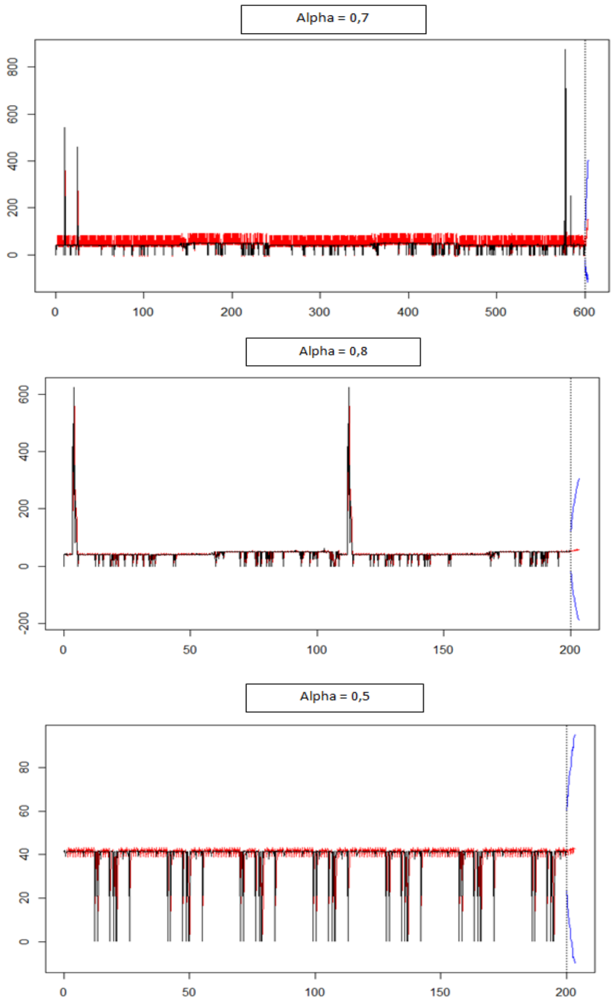


Fig. 4. Adjustment of the Constant Alpha for Three Sessions.

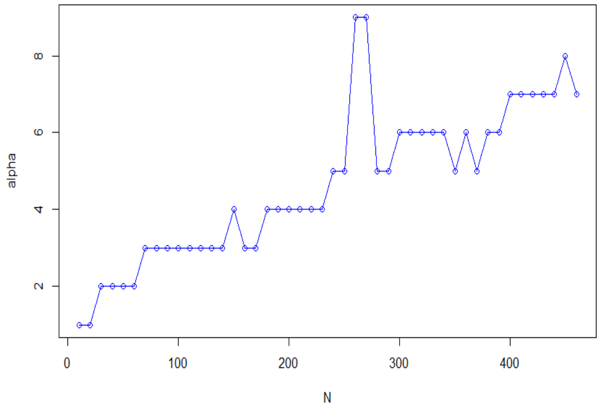


Fig. 5. The Optimal Value Variation of Alpha According to the Number of Flows.

3) *Model for predicting the percentage of lost packets*

Unlike the two indicators, the constant alpha does not vary significantly for the percentage of lost packets. Thus, for the model of this indicator it suffices to define the value of the constant which minimizes the margin of the error. For this purpose, the experimental study demonstrated that the most suitable value of α is 3/5. The percentage model of packages is:

$$\hat{P}_{t+h} = \frac{3}{5}\sum_{j=0}^{t-1}\left(\frac{12}{5}\right)^{j} P_{t-j} \qquad (6)$$

D. *Implementation*

The integration of the proactive approach into the SLM & M system must ensure an easy and transparent exchange of messages without reducing overall system performance. The indicator estimator is developed in JAVA to facilitate integration with the existing monitoring solution (Fig. 6).

Fig. 6 illustrates the prediction system which consists of three prediction functions based on the models defined in Section III-C.

The system takes the last value collected from the Acess agent resources. These values are then saved before using them in the prediction model. The estimated values belong to a near future in order to allow the correction if a degradation is predicted. The choice to deploy the estimator directly into the resources aims to reduce the prediction time by ensuring direct communication between the indicator collection processes and the prediction system.

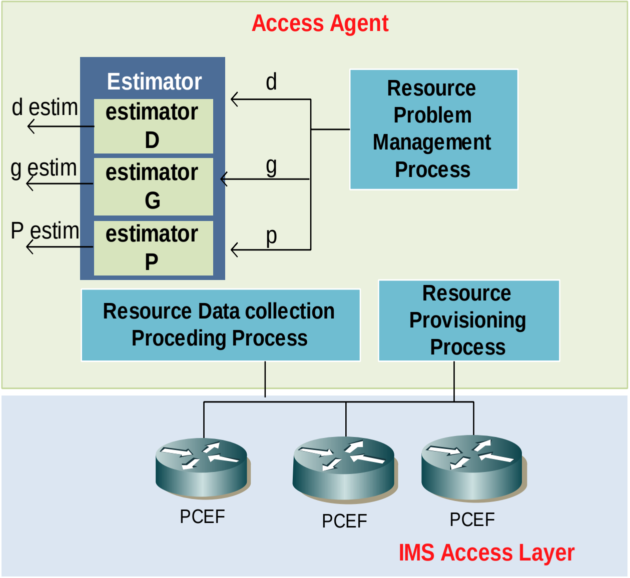

Access Agent
Estimator
d estim
estimator D
g estim
estimator G
P estim
estimator P
d
g
p
Resource Problem Management Process
Resource Data collection Proceding Process
Resource Provisioning Process
PCEF
PCEF
PCEF
IMS Access Layer

Fig. 6. Integrating the Estimator into the SLM and M Solution.

IV. Experiments and Results

The experimentation phase aims at validating the prediction approach in real cases of service provision in the NGN network. In order to focus on the accuracy of the predicted values but also on the impact of such an approach on the service assurance mechanism and the execution costs on the resources. For this purpose, we propose the test bench schematized in Fig. 7 that can emulate an NGN network and the various modules of the proactive system.

The test bench consists of the following entities:

- Linux router in the boundaries that connect the core of the network to access networks. In addition, these routers also include the control entities deployed by the OpenIMScore [22] solution as well as the Acess module of the monitoring solution.
- Core-type linux router that includes other control functions as well as some module of the platform solution. This router provides connectivity with the application servers.
- Management Server that includes the modules of the solution belonging to the insurance level.
- Server Streaming application server of type VoD.

The nominal flow of the experiment is carried out according to two stages:

- Case 1: At first, the BoB client whose SLA is of the platinum type registers in the network before requesting the VoD service.
- Case 2: In a second step, several competing flows are launched in the network via the IPREF [23] solution.

In both experimental cases, the proactive solution is evaluated according to two essential criteria, namely the accuracy of predictions and the cost of deploying resources. The effect of the prediction of the indicators on the quality of the experiment is also taken into consideration during the experimental phase.

Fig. 7. Illustrates a Screenshot of the Video in the Second Experimental Case.

Fig. 8 illustrates the quality of the video captured in the first experimental case. For this purpose, the quality is very acceptable, given the absence of competing flows in the network. Thus, the SLA contract of the customer is perfectly respected by the service provider.

For this purpose, we note that the quality of video remains acceptable despite the presence of competing flows in the network. This is explained by the prediction of the indicators which made it possible to launch the configurations of the corrections in order to avoid the effective degradation of the quality of experience of the user BOB (Fig. 9).

After validation of the proactive approach it is important to evaluate the impact of the prediction on resources. Fig. 10 illustrates the rate of CPU consumption in routers according to the number of flows in the network after deployment of the prediction system. It can be noted that the prediction of the indicators does not significantly affect the resources, since a difference of less than 7% is recorded in the most critical cases (160 flows). This is explained by the choice of exponential smoothing for the implementation of the prediction model.

Fig. 8. Screen Video Capture in the Second Test Case.

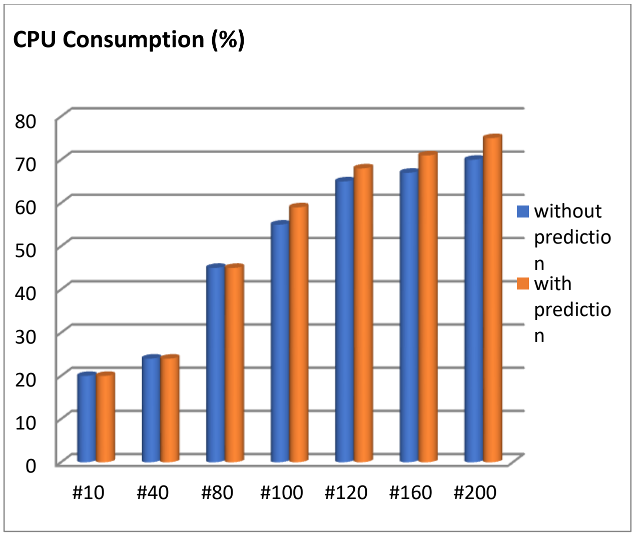


Fig. 9. CPU utilization by Routers According to the Number of Streams.

V. Conclusion

Predicting values in the future based on observations at previous times is a complex problem that requires the use of techniques appropriate to the targeted domain. The nature of NGN monitoring, in particular the variation in observations collected, has required the adoption of an adaptive approach for the estimation of performance indicators. For this purpose, the proposed new prediction model, based on exponential smoothing and linear regression, has resulted in impeccable performance during the test phase.

The migration to a proactive and autonomous system for the monitoring and supervision of the services made NGNs highly reliable. Reliability that is shown in the level of quality of service guarantees in any moment and for different network states.

Service assurance in NGN is an important step to encourage the deployment of value-added services which is important for the diversity of the offers of telecommunication operators. However, the purchase of these services has a major blow and is not easily amortized, since services are not necessarily sold to end customers. Thus, it becomes important to think about new solutions to ensure the discovery of services directly from service providers by taking advantage of new technologies offered in this direction.

References

[1] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Maria Adamczyk, Michael Denny, Nicholas Steven, Myranda Johnson, Abdi Modarressi, Hong Nguyen, Scott Traynham ; " Application Services Infrastructure For Next Generation Networks Including A Notification Capability And Related Methods And Computer Program Products " ; Patent N° US 9.288,276 B2 ; 15 March 2016.

[3] Hubert Przybysz ; "Andling Multiple User Interfaces In An Ip Multimedia Subsystem" ; Patent N°: US 8.472,376-B2 ; Jun.25, 2013.

[4] RAOUYANE Brahim, BELMEKKI Elmostafa, KHAIRI sara and BELLAFKIH mostafa, "Impact of Security in QoS Signaling in NGN: Registration Study" International Journal of Advanced Computer Science and Applications (IJACSA), 9(8), 2018.

[5] IP Multimedia Subsystem (IMS); Stage 2, 3GPP, TS 23,228, Release 9,2010.

[6] J.L. Chen, S.L. Wuy, Y.T. Larosa, P.J. Yang, and Y.F.Li, "IMS cloud computing architecture for high-quality multimedia applications," in 2011 7th International Wireless Communications and Mobile Computing Conference, 2011, pp. 1463-1468.

[7] Bo Li, M. Hamdi, Dongyi Iang, Xi-Ren Cao and Y. T. Hou, "QoS enabled voice support in the next generation Internet: issues, existing approaches and challenges," in IEEE Communications Magazine, vol. 38, no. 4, pp. 54-61, Apr 2000. doi: 10.1109/35.833557.

[8] Youssef SERAOUI, Brahim RAOUYANE and Mostafa BELLAFKIH. "An Extended IMS Framework With a High-Performance and Scalable Distributed Storage and Computing System". The International Symposium on Networks, Computers and Communications (ISNCC) Marrakech, Morocco 16-18 May 2017.

[9] J. Zhao, H. Wang, J. Dong, and S. Cheng, "A reliable and high-performance distributed storage system for P2P-VoD service," in 2012 IEEE Int. Conf. on Cluster Computing, 2012, pp. 614-617.

[10] Mohammed Errais, Mostafa Bellafkih and Brahim Raouyane "Fuzzy Video Streaming Control in IP Multimedia Subsystem Architecture" 9th International Conference on Intelligent Systems: Theories and Applications 07-08 May 2014, Rabat, Morocco.

[11] Mohammed Errais, Mostafa Bellafkih, Daniel Ranc, "Autonomous system for network monitoring and service correction in IMS Architecture", International Journal of Computer Science & Applications, ISSN 0972-9038, Volume 12 Issue 1, 2015.

[12] Business Process Framework (eTOM), Enhanced Telecom Operation management, GB921, version 7.2

[13] The NGOSS Real World use case , Version 1.2, GB921, T.

[14] Jon Siegel; « Corba 3 fundamentals and programming" ; John Wiley & Sons, 2000 - 899 pages.

[15] Kennedy Were, TienBu, Ystein Dick, Bal Ram Singh ; " A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape " , Ecological Indicators Volume 52, May 2015, Pages 394-403.

[16] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott; "Inferring causal impact using Bayesian structural time-series models"; The Annals of Applied Statistics , Volume 9, Number 1 (2015), 247-274.

[17] Erdal Kayacan, Baris Ulutas, Okyay Kaynak; "Grey system theory-based models in time series prediction"; Expert Systems with Applications Volume 37, Issue 2, March 2010, Pages 1784-1789.

[18] Charles C.Holt ; "Forecasting seasonals and trends by exponentially weighted moving averages"; International Journal of Forecasting Volume 20, Issue 1, January–March 2004, Pages 5-10.

[19] Rob J. Hyndman , Anne B. Koehker, J. Keith Ord, Ralph D. Snyder ; " Forecasting With Exponential Smoothing : The State Space Approach" BOOK; ISBN 978-3-540-71916-8.

[20] N. Garg, S. K. Mangal, P. K. Saini, P. Dhiman, S. Maji ; « Comparison of ANN and Analytical Models in Traffic Noise Modeling and Predictions" ; Acoustics Australia – Springer Link ; August 2015, Volume 43, Issue 2, pp 179–189.

[21] S. I. V Sousa, F. G. Martins, M. C. M. Alvim, Ferraz M. C. Pereira ; « Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations» ; Environmental Modelling & Software Volume 22, Issue 1, January 2007, Pages 97-103.

[22] OpenIMSCore home page, [online] Available: http://http://www. openimscore.org/.

[23] IPerf home page, [online] Available: https://iperffr/.

A Distributed Memory Parallel Fourth-Order IADEMF Algorithm

Noreliza Abu Mansor[1], Ahmad Kamal Zulkifle[3]
College of Engineering
Universiti Tenaga Nasional
Selangor, Malaysia

Norma Alias[2]
Ibnu Sina Institute of Fundamental Science Studies, Universiti Teknologi Malaysia, Johor, Malaysia

Mohammad Khatim Hasan[4]
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia

***Abstract*—The fourth-order finite difference Iterative Alternating Decomposition Explicit Method of Mitchell and Fairweather (IADEMF4) sequential algorithm has demonstrated its ability to perform with high accuracy and efficiency for the solution of a one-dimensional heat equation with Dirichlet boundary conditions. This paper develops the parallelization of the IADEMF4, by applying the Red-Black (RB) ordering technique. The proposed IADEMF4-RB is implemented on multiprocessor distributed memory architecture based on Parallel Virtual Machine (PVM) environment with Linux operating system. Numerical results show that the IADEMF4-RB accelerates the convergence rate and largely improves the serial time of the IADEMF4. In terms of parallel performance evaluations, the IADEMF4-RB significantly outperforms its counterpart of the second-order (IADEMF2-RB), as well as the benchmarked fourth-order classical iterative RB methods, namely, the Gauss-Seidel (GS4-RB) and the Successive Over-relaxation (SOR4-RB) methods.**

Keywords*—Fourth-order method; finite difference; red-black ordering; distributed memory architecture; parallel performance evaluations

I. Introduction

The heat equation is a mathematical model that describes heat conduction processes of a physical system. Sahimi et al. [1] had proposed a finite difference scheme known as the Iterative Alternating Decomposition Explicit (IADE) method to approximate the solution of a one-dimensional heat equation with Dirichlet boundary conditions. The IADE scheme employs the fractional splitting of the Mitchell and Fairweather (MF) variant whose accuracy is of the order, $O\left((\Delta t)^2+(\Delta x)^4\right)$. The scheme, commonly abbreviated as the IADEMF, is developed by applying the second-order spatial accuracy to the heat equation. Due to the latter, in this paper, the IADEMF will also be referred to as the IADEMF2. It is a two-stage iterative procedure and has been proven to have merit in terms of convergence, stability and accuracy. It is generally found to be more accurate than the classical Alternating Group Explicit class of methods [2].

Several studies have later been developed based on the IADE method. Sahimi et al. [3, 4] developed new second-order IADE methods using different variants such as the D'Yakonov (IADEDY) and the Mitchell-Griffith variant (IADEMG). Each variant is of the order, $O\left((\Delta t)^2+(\Delta x)^4\right)$. The studies showed that the accuracies of the IADEDY and the IADEMG are comparable to the IADEMF. Alias [5] studied the parallel implementation of the IADEMF on distributed parallel computing using the parallel virtual machine. A fragmented numerical algorithm of the IADEMF method was designed by Alias [6] in terms of the data-flow graph where its parallel implementation using LuNA programming system was then executed. Sulaiman et al. [7, 8] proposed the half-sweep and the quarter-sweep IADEMF methods respectively, for the purpose of achieving better convergence rate and faster execution time than the corresponding full-sweep method. Alias [9] implemented the Interpolation Conjugate gradient method to improve the parallel performance of the IADEMF. Shariffudin et al. [10] presented the parallel implementation of the IADEDY for solving a two-dimensional heat equation on a distributed system of Geranium Cadcam cluster (GCC) using the Message Passing Interface.

A recent study made by Mansor [11] involved the development of a convergent and unconditionally stable fourth-order IADEMF sequential algorithm (IADEMF4). The proposed scheme is found to be capable of enhancing the accuracy of the original corresponding method of the second-order, that is, the IADEMF2. The IADEMF4 seems to be more accurate, more efficient and has better rate of convergence than the benchmarked fourth-order classical iterative methods, namely, the Gauss-Seidel (GS4) and the successive over-relaxation (SOR4) methods. However, the IADEMF4 may be too slow to be implemented especially when the problem involves larger linear systems of equations. It is thus justified to consider parallel computing to speed up the execution time without compromising its accuracy. The algorithm has explicit features which add to its advantage, thus it can be fully utilized for parallelization.

This paper attempts to parallelize the IADEMF4, by applying the Red-Black (RB) ordering technique, for solving large sparse linear systems that arise from the discretization of the one-dimensional heat equation with Dirichlet boundary conditions. It aims to effectively implement the IADEMF4-RB on parallel computers, with improved performance over its serial counterpart. The high computational complexity of the IADEMF4-RB will be implemented on multiprocessor distributed memory architecture based on Parallel Virtual Machine (PVM) environment with Linux operating system.

This paper is outlined as follows. Section II recalls the formulation of the IADEMF4 scheme. Section III presents the development of the IADEMF4-RB parallel strategy. The computational complexity of the RB methods considered in this paper is given in Section IV. Section V shows the numerical experiment conducted in this study. The results and discussion on parallel performance of the methods under consideration are discussed in Section VI. At the end of this paper is the conclusion.

II. Formulation of the IADEMF4 (An Overview)

In this section, the development of the IADEMF4 algorithm [11] is briefly reviewed. Consider the one-dimensional heat equation (1) which models the flow of heat in a homogeneous unchanging medium of finite extent, in the absence of heat source.

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2} \tag{1}$$

subject to given initial and Dirichlet boundary conditions

$$\begin{aligned} &U(x,0) = f(x), \quad 0 \le x \le 1 \\ &U(0,t) = g(t), \quad 0 < t \le T \\ &U(1,t) = h(t), \quad 0 < t \le T \end{aligned} \tag{2}$$

Based on the finite difference approach, the time-space domain is discretized by using a set of lines parallel to the $t-$axis given by $x_i = i\Delta x$, $i = 0,1, \ \dots \ , m, m+1$ and a set of lines parallel to the $x-$axis given by $t_k = k\Delta t$, $k = 0,1, \ \dots \ , n, n+1$. The grid spacings have uniform size, that is, $\Delta x = 1/(m+1)$ and $\Delta t = T/(n+1)$. At a grid-point $P(x_i, t_k)$ in the solution domain, the dependent variable $U(x,t)$ which represents the non-dimensional temperature at time t and at position x, is approximated by u_i^k.

The IADEMF4 is developed by firstly executing the unconditionally stable fourth-order Crank-Nicolson approximation (3) on the heat equation [12].

$$\frac{1}{\Delta t}(u_i^{k+1} - u_i^k) = \frac{1}{2(\Delta x)^2}(\delta_x^2 - \frac{1}{12}\delta_x^4)(u_i^{k+1} + u_i^k) \tag{3}$$

The discretization of (3) leads to the expression given in (4), with the constants defined as in (5).

$$\begin{aligned} &au_{i-2}^{k+1} + bu_{i-1}^{k+1} + cu_i^{k+1} + du_{i+1}^{k+1} + eu_{i+2}^{k+1} = -au_{i-2}^k - bu_{i-1}^k \\ &+\hat{c}u_i^k - du_{i+1}^k - eu_{i+2}^k, \quad i = 2,3, \ \dots, \ m-1 \end{aligned} \tag{4}$$

$$a = \frac{\lambda}{24}, \ b = -\frac{2\lambda}{3}, \ c = \frac{4+5\lambda}{4}, \ d = -\frac{2\lambda}{3}, \ e = \frac{\lambda}{24}, \ \hat{c} = \frac{4-5\lambda}{4} \tag{5}$$

In matrix form, the approximation in (4) can be represented by $A\mathbf{u} = \mathbf{f}$ (6), where A is a sparse penta-diagonal coefficient matrix, and the column vectors $\mathbf{u} = (u_2, u_3, \dots, u_{m-2}, u_{m-1})^T$ contain the unknown values of u at the time level $k+1$ and $\mathbf{f} = (f_2, f_3, \dots, f_{m-2}, f_{m-1})^T$ consists of boundary values and known u values at the previous time level k.

$$A\mathbf{u} = \mathbf{f}$$

$$\begin{bmatrix} c & d & e & & & & \\ b & c & d & e & & \mathrm{O} & \\ a & b & c & d & e & & \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & a & b & c & d. & e \\ & \mathrm{O} & & a & b & c & d \\ & & & & a & b & c \end{bmatrix}_{(m-2)x(m-2)} \begin{bmatrix} u_2 \\ u_3 \\ . \\ . \\ . \\ u_{m-2} \\ u_{m-1} \end{bmatrix}_{k+1} = \begin{bmatrix} f_2 \\ f_3 \\ . \\ . \\ . \\ f_{m-2} \\ f_{m-1} \end{bmatrix} \tag{6}$$

The entries in $\mathbf{f}$ are defined as

$$\begin{aligned} f_2 &= -b(u_1^k + u_1^{k+1}) + \hat{c}u_2^k - du_3^k - eu_4^k \\ f_3 &= -a(u_1^k + u_1^{k+1}) - bu_2^k + \hat{c}u_3^k - du_4^k - eu_5^k \\ f_i &= -au_{i-2}^k - bu_{i-1}^k + \hat{c}u_i^k - du_{i+1}^k - eu_{i+2}^k, \quad i = 4,5,\dots,m-3 \\ f_{m-2} &= -au_{m-4}^k - bu_{m-3}^k + \hat{c}u_{m-2}^k - du_{m-1}^k - e(u_m^k + u_m^{k+1}) \\ f_{m-1} &= -au_{m-3}^k - bu_{m-2}^k + \hat{c}u_{m-1}^k - d(u_m^k + u_m^{k+1}) \end{aligned} \tag{7}$$

The IADEMF4 scheme secondly employs the fractional splitting of the higher-order accuracy formula of the MF variant [13],

$$(rI + G_1)\mathbf{u}^{(p+1/2)} = (rI - gG_2)\mathbf{u}^{(p)} + \mathbf{f} \tag{8}$$

$$(rI + G_2)\mathbf{u}^{(p+1)} = (rI - gG_1)\mathbf{u}^{(p+1/2)} + g\mathbf{f} \tag{9}$$

where G_1 and G_2 are two constituent matrices and r, I and p represent an acceleration parameter, an identity matrix and the iteration index respectively. The value of g is defined as $g = \frac{6+r}{6}$, $r > 0$. The vectors $\mathbf{u}^{(p+1)}$ and $\mathbf{u}^{(p+1/2)}$ represent the approximate solution at the iteration level $(p+1)$ and at some intermediate level $(p+1/2)$, respectively.

After some algebraic manipulations for the equations in (8) and (9), the form, $\left[G_1 + G_2 - \frac{1}{6}G_1 G_2\right]\mathbf{u} = \mathbf{f}$ is obtained, suggesting that matrix A in (6) can be decomposed into.

$$A = G_1 + G_2 - \frac{1}{6}G_1 G_2 \tag{10}$$

To retain the penta-diagonal structure of A, the matrices G_1 and G_2 have to be in the form of lower and upper tri-diagonal matrices respectively, Thus,

$$G_1 = \begin{bmatrix} 1 & & & & & \\ l_1 & 1 & & & \text{O} & \\ \hat{m}_1 & l_2 & \ddots & & & \\ & \hat{m}_2 & \ddots & \ddots & & \\ & & \ddots & l_{m-4} & 1 & \\ & \text{O} & & \hat{m}_{m-4} & l_{m-3} & 1 \end{bmatrix}_{(m-2)x(m-2)}$$

and

$$G_2 = \begin{bmatrix} \hat{e}_1 & \hat{u}_1 & \hat{v}_1 & & & \\ & \hat{e}_2 & \hat{u}_2 & \hat{v}_2 & \text{O} & \\ & & \ddots & \ddots & \ddots & \\ & \text{O} & & \hat{e}_{m-4} & \hat{u}_{m-4} & \hat{v}_{m-4} \\ & & & & \hat{e}_{m-3} & \hat{u}_{m-3} \\ & & & & & \hat{e}_{m-2} \end{bmatrix}_{(m-2)x(m-2)} \tag{11}$$

If each G_1 and G_2 in (11) is substituted into the matrix A in (10), then the new entries of the latter can be compared with those in (6) to yield the following constants.

$$\hat{e}_1 = \frac{6(c-1)}{5}, \quad \hat{u}_1 = \frac{6d}{5}, \quad l_1 = \frac{6b}{6-\hat{e}_1}, \quad \hat{e}_2 = \frac{6(c-1)+ \ l_1\hat{u}_1}{5},$$

$$\hat{v}_i = \frac{6e}{5} \quad \text{where} \quad i = 1, 2, \ \ldots, m-4$$

for $i = 2, 3, \ \ldots, m-3$

$$\hat{u}_i = \frac{6d + l_{i-1}\hat{v}_{i-1}}{5}, \quad \hat{m}_{i-1} = \frac{6a}{6-\hat{e}_{i-1}}, \quad l_i = \frac{6b + \hat{m}_{i-1}\hat{u}_{i-1}}{6-\hat{e}_i},$$

$$\hat{e}_{i+1} = \frac{6(c-1)+ \ l_i\hat{u}_i + \hat{m}_{i-1}\hat{v}_{i-1}}{5} \tag{12}$$

Since G_1 and G_2 are three banded matrices, then it is easy to obtain the inverses of $(rI + G_1)$ and $(rI + G_2)$. By rearranging the equations in (8) and (9), the following expressions are obtained.

$$\mathbf{u}^{(p+1/2)} = (rI + G_1)^{-1}(rI - gG_2)\mathbf{u}^{(p)} + (rI + G_1)^{-1}\mathbf{f} \tag{13}$$

$$\mathbf{u}^{(p+1)} = (rI + G_2)^{-1}(rI - gG_1)\mathbf{u}^{(p+1/2)} + g(rI + G_2)^{-1}\mathbf{f} \tag{14}$$

Based on the above two equations, the computational formulae at each of the half iteration levels can be derived as given in (15) and (16).

1) At the (p+1/2) iteration level:

$$u_i^{(p+1/2)} = \frac{1}{R}(E_{i-1}u_i^{(p)} + W_{i-1}u_{i+1}^{(p)} + V_{i-1}u_{i+2}^{(p)} - \hat{m}_{i-3}u_{i-2}^{(p+1/2)}$$
$$- l_{i-2}u_{i-1}^{(p+1/2)} + f_i), \quad i = 2, 3, \ldots, m-2, m-1 \tag{15}$$

2) At the (p+1) iteration level:

$$u_i^{(p+1)} = \frac{1}{Z_{i-1}}(S_{i-3}u_{i-2}^{(p+1/2)} + Q_{i-2}u_{i-1}^{(p+1/2)} + Pu_i^{(p+1/2)} - \hat{u}_{i-1}u_{i+1}^{(p+1)}$$
$$- \hat{v}_{i-1}u_{i+2}^{(p+1)} + gf_i), \quad i = m-1, m-2, \ldots, 3, 2 \tag{16}$$

with

$$\hat{m}_{-1} = \hat{m}_0 = l_0 = V_{m-2} = V_{m-3} = W_{m-2} = \hat{u}_{m-2} = \hat{v}_{m-2}$$
$$= \hat{v}_{m-3} = Q_0 = S_{-1} = S_0 = 0$$

$$R = 1 + r, \quad P = r - g,$$
$$E_i = r - g\hat{e}_i, \quad Z_i = r + \hat{e}_i, \quad i = 1, 2, \ \ldots, m-2$$
$$W_i = -g\hat{u}_i, \quad Q_i = -gl_i, \quad i = 1, 2, \ \ldots, m-3$$
$$V_i = -g\hat{v}_i, \quad S_i = -g\hat{m}_i, \quad i = 1, 2, \ \ldots, m-4 \tag{17}$$

The two-stage IADEMF4 algorithm is implemented by using the required equations at the two iteration levels in alternate sweeps along all the grid-points in the interval (0,1) until convergence is reached. The method is explicit, since at each level of iteration, the computational molecules involve two known grid-points at the new level and another three known ones at the old level (Fig. 1 and 2). The unknown $u_i^{(p+1/2)}$ in (15) is calculated by proceeding from the left boundary towards the right, whereas the unknown $u_i^{(p+1)}$ in (16) is calculated from the right boundary and moves to the left.

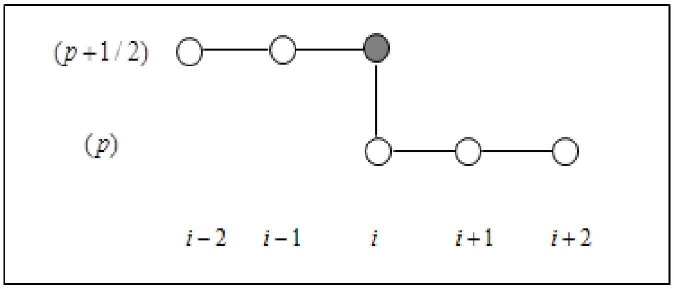


Fig. 1. Computational Molecule of the IADEMF4 at the $(p+1/2)$ Iteration Level.

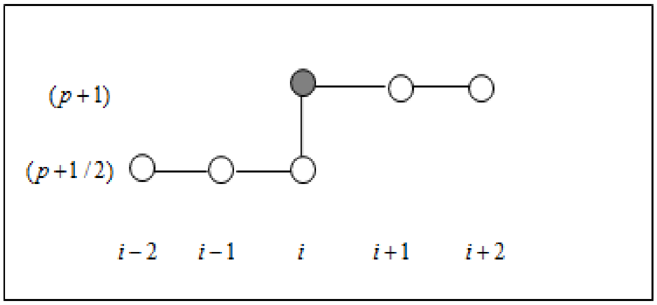


Fig. 2. Computational Molecule of the IADEMF4 at the $(p+1)$ Iteration Level.

III. Parallelization of the IADEMF4

It is observed that for $i = 2, 3, \ ..., \ m-1$, the computation of the unknown grid-point, $u_i^{(p+1/2)}$, requires the values of the grid-points at $u_{i-2,}^{(p+1/2)}$ and $u_{i-1}^{(p+1/2)}$ (Fig. 1) and the computation of the unknown $u_{m+1-i}^{(p+1)}$ requires the values of $u_{m+2-i}^{(p+1)}$ and $u_{m+3-i}^{(p+1)}$ (Fig. 2). The unknown grid-points can only be determined after the values of their two previous neighbors at their respective current iteration levels have been calculated. In other words, all values at the $(p+1/2)$th level cannot be calculated independently and simultaneously, so as values at the $(p+1)$th level. These situations show that the IADEMF4 is not inherently parallel. Thus, to handle this problem, this study resorts to undertake a domain decomposition approach that firstly divides the physical domain into a number of subdomains, each being assigned to a processor; and secondly exchanges appropriate data across the boundaries of the subdomains. The Red-Black (RB) ordering is the domain decomposition strategy that is considered in this study. The approach focuses on minimizing the problem of data dependencies and it is highly parallel.

A. The IADEMF4-RB

The RB ordering has shown its competitiveness in terms of speedup and efficiency, as has been proven in studies made by Evans [14] in solving the parallel SOR iterative methods; Brill et al. [15] in using the block GS-RB on the Hermite collocation discretization of partial differential equations in two spatial dimensions; and Alias [5] in parallelizing the IADEMF2. Darwis et al. [16] proved that the GS-RB algorithm is more accurate and converges faster than the GS algorithm. Yavneh [17] showed that the SOR-RB is more efficient and smoother than the sequential SOR method for solving two-dimensional Poisson equations.

This section parallelizes the IADEMF4 by using the RB ordering technique. The algorithm used will be referred to as the IADEMF4-RB.

The strategy to develop the IADEMF4-RB algorithm begins by decomposing the domain Ω into two different independent subdomains, Ω^R and Ω^B. Each grid-point in the subdomains Ω^R and Ω^B is denoted red and black respectively. If i is even, the grid-point is marked red, and if i is odd, the grid-point is marked black. Assuming m is even, then, the computational formulae for the IADEMF4-RB are:

$$u_i^{(p+1/2)} = (1-\omega_y)u_i^{(p)} + \frac{\omega_y}{R}(E_{i-1}u_i^{(p)} + W_{i-1}u_{i+1}^{(p)} + V_{i-1}u_{i+2}^{(p)} - \hat{m}_{i-3}u_{i-2}^{(p+1/2)} - l_{i-2}u_{i-1}^{(p+1/2)} + f_i) \tag{18}$$

for $i = 2, 4, .., m-2$ (red grid-points) and $i = 3, 5, .., m-1$ (blackgrid-points)

$$u_i^{(p+1)} = (1-\omega_z)u_i^{(p+1/2)} + \frac{\omega_z}{Z_{i-1}}(S_{i-3}u_{i-2}^{(p+1/2)} + Q_{i-2}u_{i-1}^{(p+1/2)} + Pu_i^{(p+1/2)} - \hat{u}_{i-1}u_{i+1}^{(p+1)} - \hat{v}_{i-1}u_{i+2}^{(p+1)} + gf_i)$$

(19)

for $i = 2, 4, .., m-2$ (red grid-points) and $i = 3, 5, .., m-1$ (blackgrid-points)

The purpose of including the relaxation factors ω_y and ω_z in (18) and (19) is to accelerate the convergence rate of the scheme.

The IADEMF4-RB ordering, on say, three processors, P$_1$, P$_2$ and P$_3$, is illustrated in Fig. 3. P$_1$ and P$_3$ holds boundary values at $i = 0$ and $i = m+1$, respectively. The fourth-order methods require additional boundary values which are at positions $i = 1$ (a grid-point in P$_1$) and $i = m$ (a grid-point in P$_3$). As a strategy to obtain good load balancing, similar numbers of alternate red (R) and black (B) grid-points are assigned to each processor [18]. Depending on the color of the grid-point, the first two starting grid-points in a processor may be labelled as 'st_R' and followed by 'st_B', and the last two end grid-points may be labelled as 'en_B' followed by 'en_R'.

The following describes the implementation of the IADEMF4-RB based on Fig. 3. The algorithm is subjected to the given initial and boundary conditions. Before the beginning of the execution, the unknowns, $u_i^{(p+1/2)}$, for $i = 2, 3, .., m-1$, are given 'guessed' values at the initial time. Then, the execution of the IADEMF4-RB algorithm is performed in two phases:

The first phase involves the computations of only the red grid-points at the iteration levels $(p+1/2)$ and $(p+1)$. This phase requires every processor to compute in parallel the red unknowns by making use of the initialized 'guessed' values. Example, the computation of $u_{\text{st}_\text{R}}^{(p+1/2)}$ in P$_2$ requires 'guessed' $u_{\text{en}_\text{R}}^{(p+1/2)}$ value from P$_1$ and $u_{\text{st}_\text{B}}^{(p+1/2)}$ value from P$_2$ itself, while the computation of $u_{\text{en}_\text{R}}^{(p+1)}$ in P$_2$ requires 'guessed' $u_{\text{st}_\text{R}}^{(p+1)}$ value from P$_3$ and $u_{\text{en}_\text{B}}^{(p+1)}$ value from P$_2$ itself.

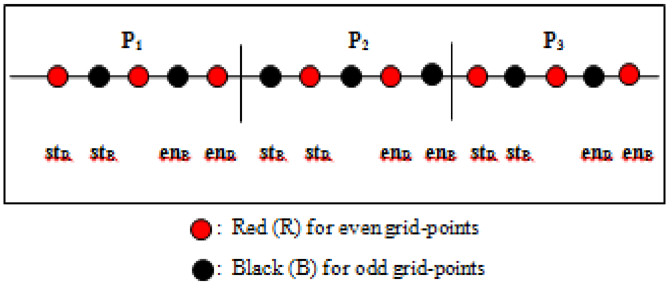

P₁
P₂
P₃
stR
stB
enB
enR
stB
stR
enR
enB
stR
stB
enR
enB
●: Red (R) for even grid-points
●: Black (B) for odd grid-points

Fig. 3. One-Dimensional IADEMF4-RB Ordering

After the computations of the red grid-points for the two iteration levels have been completed, adjacent processors exchange their updated red values at the boundary grid-points to prepare for the calculation of the black grid-points in the second phase. Example,

Send updated $u_{\text{en}_\text{R}}^{(p+1/2)}$: from P_1 to P_2, and from P_2 to P_3.

Send updated $u_{\text{st}_\text{R}}^{(p+1)}$: from P_2 to P_1, and from P_3 to P_2.

The second phase continues by computing simultaneously the black unknowns at levels $(p+1/2)$ and $(p+1)$, using the most recent red values computed in the first phase. For example, the computation of $u_{\text{st}_\text{B}}^{(p+1/2)}$ in P_2 uses the updated red values $u_{\text{en}_\text{R}}^{(p+1/2)}$ and the 'guessed' black value $u_{\text{en}_\text{B}}^{(p+1/2)}$ from P_1, while the computation of $u_{\text{en}_\text{B}}^{(p+1)}$ in P_2 requires the updated red values $u_{\text{st}_\text{R}}^{(p+1)}$ from P_3 and the 'guessed' $u_{\text{st}_\text{B}}^{(p+1)}$ value from P_3. The updated black grid-points at the boundaries are then shared between adjacent processors. Example,

Send updated $u_{\text{en}_\text{B}}^{(p+1/2)}$: from P_1 to P_2, and from P_2 to P_3

Send updated $u_{\text{st}_\text{B}}^{(p+1)}$: from P_2 to P_1, and from P_3 to P_2.

The two phases are repeated until convergence is reached. Due to the dependencies on the updated values between adjacent processors, the IADEMF4-RB algorithm involves statements that take care of the communication between the processors. An example of a procedure for sending and receiving messages between processors in a PVM environment is as illustrated in Fig. 4. The IADEMF4-RB algorithm implemented by a slave processor can be described as in Fig. 5.

```
if (left!=0) /* If there is a processor on the left*/
        pvm_initsend( PvmDataDefault );
        pvm_pkdouble( &u [start], 1,1);
        pvm_send(left,50 );
end-if
if (right!= 0) /* If there is a processor on the right*/
        pvm_recv(right,50);
        pvm_upkdouble(&u [end+1],1, 1 );
        pvm_initsend( PvmDataDefault );
        pvm_pkdouble(&u  [end], 1,1);
        pvm_send(right,60 );
 end-if
if (left!=0) /* If there is a processor on the left*/
        pvm_recv(left,60);
        pvm_upkdouble(&u [start-1],1, 1 );
end-if
```

Fig. 4. Communication Procedures for Sending and Receiving Messages between Adjacent Processors.

```
IADEMF4 –RB: Slave's Parallel Algorithm
begin
    slaves receive data from master: $tid$ , $m$, $r$, $\Delta t$, $\Delta x$, $\lambda$ , $\omega$
        for $\forall i \in \Omega$
                determine initial conditions $u_i^{(p)}$
                initialize guessed values $u_i^{(p+1/2)}$
        end-for
        while (time level < $T$ )
                for  $t = k$ and $t = k+1$
                        determine boundary conditions at $u_0$ ,
                        $u_1$ , $u_m$ and $u_{m+1}$
                end-for
                for $\forall i \in \Omega^R$
                        compute $f_i$  (refer to (7))
                end-for
                for $\forall i \in \Omega^B$
                        compute $f_i$  (refer to (7))
                end-for
                set iteration $= 0$
            while (convergence conditions are not satisfied)
                    for $\forall i \in \Omega^R$
                            compute $u_i^{(p+1/2)}$
                                        (refer to (18))
                    end-for
                    for $\forall i \in \Omega^R$
                    compute $u_i^{(p+1)}$      (refer to (19))
                    end-for
                    send and receive updated red boundary
                    values between adjacent slave
                    processors (Fig. 4)
                    for $\forall i \in \Omega^B$
                            compute $u_i^{(p+1/2)}$
                                        (refer to (18))
                    end-for
                    for $\forall i \in \Omega^B$
                          compute $u_i^{(p+1)}$ (refer to (19))
                    end-for
                    send and receive updated black
                    boundary values between adjacent
                    slave processors (Fig. 4)
                    test for convergence:
                    compute $e_i \leftarrow \left|u_i^{(p+1)} - u_i^{(p)}\right|$ for
                    $\forall i \in \Omega^R$ and $\forall i \in \Omega^B$
                    if max $\left|e_i\right| < \varepsilon$
                            then $u_i^{(p)} \leftarrow u_i^{(p+1)}$
                    add 1 to iteration (if necessary)
                end-while
        end-while
        Determine numerical errors for $\forall i \in \Omega^R$ and $\forall i \in \Omega^B$
        slave sends data analysis to master
        pvm_exit;
```

Fig. 5. IADEMF4-RB–Slave's Parallel Algorithm.

B. Parallel Algorithms for Benchmarking

The IADEMF2, the GS4 and the SOR4 algorithms [9] can also be parallelized using the RB ordering technique. They will serve as the benchmarks for the parallel IADEMF4-RB. The following are the schemes under consideration, assuming m is even.

1) the IADEMF2-RB algorithms:

$$u_i^{(p+1/2)} = (1-\omega_y)u_i^{(p)} + \frac{\omega_y}{d}(-l_{i-1}u_{i-1}^{(p+1/2)} + s_i u_i^{(p)} + w_i u_{i+1}^{(p)} + f_i) \tag{20}$$

for $i = 2,4,...,m$ (red grid-points) and $i = 1,3,5,...,m-1$ (black grid-points)

$$u_{m+1-i}^{(p+1)} = (1-\omega_z)u_i^{(p+1/2)} + \frac{\omega_z}{d_{m+1-i}}(v_{m-i}u_{m-i}^{(p+1/2)} + su_{m+1-i}^{(p+1/2)} + gf_{m+1-i} - \hat{u}_{m+1-i}u_{m+2-i}^{(p+1)}) \tag{21}$$

for $i = 2,4,...,m$ (red grid-points) and $i = 1,3,5,...,m-1$ (black grid-points)

2) the SOR4-RB algorithm (reduces to the GS4-RB algorithm when $\omega = 1$:

$$u_i^{(p+1)} = (1-\omega)u_i^{(p)} + \frac{\omega}{c}(f_i - au_{i-2}^{(p+1)} - bu_{i-1}^{(p+1)} - du_{i+1}^{(p)} - eu_{i+2}^{(p)}) \tag{22}$$

for $i = 2,4,...,m-2$ (red grid-points) and $i = 3,5,...,m-1$ (black grid-points)

IV. Computational Complexity

The computational complexity of the RB algorithms of interest is as given in Table I. It gives the number of parallel arithmetic operations that is required to evaluate the algorithms.

TABLE. I. Parallel Arithmetic Operations (m = Problem Size, n = Number of Iterations, P = Number of Processors)

Method	**Number of additions**	**Number of multiplications**	**Total operation count**
IADEMF4-RB	$10(m\text{-}2)n / P$	$13(m\text{-}2)n / P$	$23(m\text{-}2)n / P$
IADEMF2-RB	$6mn / P$	$9mn / P$	$15mn / P$
GS4-RB	$4(m-2)n / P$	$5(m-2)n / P$	$9(m-2)n / P$
SOR4-RB	$5(m-2)n / P$	$7(m-2)n / P$	$12(m-2)n / P$

V. Numerical Experiment

The IADEMF4-RB was implemented and tested on multiprocessor distributed memory architecture comprising of twelve interconnected processors with Linux operating system using the PVM communication library. In distributed memory, each processor has its own address space or local memory which is inaccessible to other processors. The processors operate independently in parallel, and they share their data by means of some form of inter-processor communication via an inter-connection network. The programmer is responsible for the details associated with message passing between processors. From the memory perspective, the size of memory increases in proportion to the increasing number of processors.

The parallel performances of the proposed algorithm was examined by solving a very large problem size on the experiment in (23), where m varied from 70,000 to 700,000. This problem was taken from Saul'yev (1964),

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0 \le x \le 1 \tag{23}$$

subject to the initial condition $U(x,0) = 4x(1-x)$, $0 \le x \le 1$ and the boundary conditions $U(0,t) = U(1,t) = 0, \quad t \ge 0.$

The exact solution to the given problem is given by

$$U(x,t) = \frac{32}{\pi^3}\sum_{k=1,(2)}^{\infty}\frac{1}{k^3}\; e^{-\pi^2 k^2 t}\sin(k\pi x) \tag{24}$$

The other parameters considered for the experiment were λ = 0.5, Δt = 1.0204 x 10^{-12}, t = 5.1020 x 10^{-11}, and a stringent tolerance value of $\varepsilon = 10^{-15}$. The initial and Dirichlet boundary conditions at $i = 0$ and $i = m+1$ were applied based on the values given in the problem. For the fourth-order methods, the boundary values at positions $i = 1$ and $i = m$ were taken from the given exact solutions (24). The optimum values for r and the relaxation factors (ω_x , ω_y and ω) were determined by experiments.

VI. Results and Discussion

Table II compares the accuracy of the tested parallelized RB algorithms for a fixed problem size, $m = 700,000$. It is obvious that the IADEMF4-RB outperforms the IADEMF2-RB in terms of rate of convergence. The average absolute error, root mean square error and the maximum error of both algorithms seem identical up to four decimal places, due to the stringent tolerance value set in the experiment. The high computational complexity of the IADEMF4-RB is compensated by the high accuracy it achieves at every iteration and time level, causing its convergence to accelerate. The SOR4-RB speeds up the convergence of the GS4-RB, but they are both relatively not reliable in terms of accuracy.

Table III displays the number of iterations (n), execution time, speedup and efficiency of the IADEMF4-RB on using three different values of problem size, m. The execution time refers to the amount of time required to complete a parallel program on a number of P processors from the moment the execution starts till the moment the last processor finishes its execution [19]. Speedup expresses how much faster the parallel program executes relative to the sequential one. Amdahl's law states that there exists a bound on the speedup for a given problem with a fixed size [20], since some parts of the computations for solving a given problem are not parallelizable. Efficiency is a measure of the speedup achieved per processor. It estimates how well the processors are utilized during the execution of a parallel algorithm.

TABLE. II. PARALLEL RB ALGORITHMS – ERRORS AND NUMBER OF ITERATIONS

Method (m=700,000)	**Average absolute error**	**Root mean square error**	**Max. error**	**Number of iterations**
IADEMF4-RB $\left(r=0.8, \omega_y=1, \omega_z=1.1\right)$	1.5920e-09	7.3054e-09	1.9845e-07	288
IADEMF2-RB $\left(r=0.8, \omega_y=1, \omega_z=1.1\right)$	1.5920e-09	7.3054e-09	1.9845e-07	450
SOR4-RB (ω=1.06)	1.6150e-09	9.6395e-09	2.7422e-06	738
GS4-RB	1.6150e-09	9.6395e-09	2.7422e-06	794

λ = 0.5, Δx = 2.60 x 10^{-6}, Δt = 1.02 x 10^{-12}, t = 5.10 x 10^{-11}, ε = 1 x 10^{-15}

TABLE. III. IADEMF4-RB – PERFORMANCES USING SEVERAL VALUES OF m

m	**Δx**	P	**Execution time (s)**	**Speedup**	**Efficiency**
70,000 $n=359$	1.43 x 10^{-5}	1	4.869491	1	1
		2	2.507665	1.941843	0.970921
		4	1.518787	3.206171	0.801542
		6	1.261464	3.860190	0.643365
		8	1.102297	4.417585	0.552198
		10	1.039360	4.685086	0.468508
		12	1.008263	4.829584	0.402465
385,000 $n=312$	2.60 x 10^{-6}	1	20.039541	1	1
		2	10.062964	1.991415	0.995707
		4	5.300258	3.780842	0.945210
		6	3.828272	5.234617	0.872436
		8	2.993626	6.694069	0.836758
		10	2.447466	8.187873	0.818787
		12	2.101530	9.535691	0.794640
700,000 $n=288$	1.43 x 10^{-6}	1	35.682042	1	1
		2	17.896741	1.993773	0.996886
		4	8.962541	3.981241	0.995310
		6	6.202509	5.752840	0.958806
		8	4.900683	7.281034	0.910129
		10	3.992991	8.936168	0.893616
		12	3.456841	10.32215	0.860179

λ = 0.5, Δt = 1.02 x 10^{-12}, t = 5.10 x 10^{-11}, ε = 1 x 10^{-15}

The results in Table III show that the execution time for a problem using any of the considered sizes is reduced and the speedup improves as the number of processors increases. For $m=70,000$, the increase in speedup from $P=1$ to $P=12$ is about 80% and for $m=700,000$, the increase is about 90%. This shows that parallel computation improves performance in terms of execution time and speedup over serial computation. Due to overheads, the overall efficiency for any m tends to decrease as the number of processors increases. Overheads have impacts on parallel performance. The two common types of overheads are the communication time and the idle time. The communication time is the time spent on communication and exchanging of data during the execution in all processors and the idle time is the time when processors stay idle, waiting for busy processors to send messages. Idling may be due to load imbalances amongst processors, or a bottleneck at the master processor when it has to interact with other worker processors [21].

For every number of processor ran in the experiment, the execution time for a problem size of 70,000 is comparatively smaller than a problem ten times its size. This is expected since fewer grid-points involve less mathematical operations and data sharing. The table, however, shows an improvement in convergence rate, speedup and efficiency as the size increases to $700,000$. The smaller size with higher number of iterations (n) seems to be less efficient due to the additional overhead imposed by having communications routed through the PVM daemon.

Fig. 6 shows that the execution time taken by every tested algorithm (listed in Table II) decreases with increasing P. However, the IADEMF4-RB executes in the least amount of time for every P. Despite the IADEMF4's greater computational complexity, its parallelization using the RB technique and the use of relaxation parameters have enabled it to execute in a shorter time on one and more processors in comparison to its counterpart of second-order.

Fig. 7 shows that every tested algorithm has a speedup of less than P, which implies that the parallel code is bounded by the sequential code (Amdahl's law). The parallel code runs slower due to overheads that outweigh the benefits of parallel computation. Amongst the four algorithms, the IADEMF4-RB proves to continue giving the best speedup as P increases. At $P=12$, the speedup of the IADEMF4-RB is almost 14% closer to the linear speedup. As for the IADEMF2-RB, the SOR4-RB and the GS4-RB, there is an 18, 24 and 28 percent difference, respectively, between the method's speedup and the linear speedup.

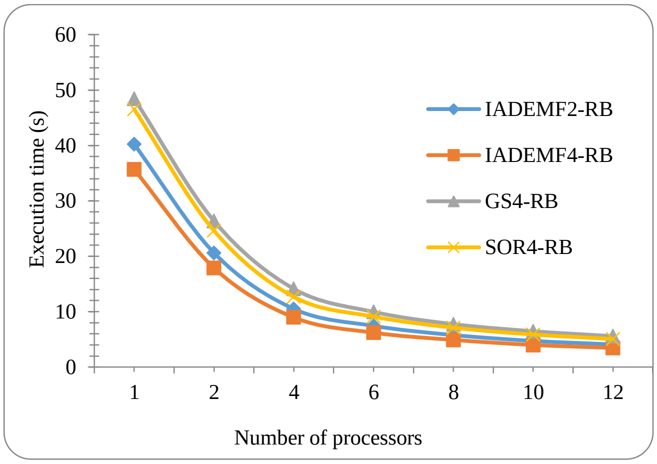


Fig. 6. Execution Time Versus Number of Processors.

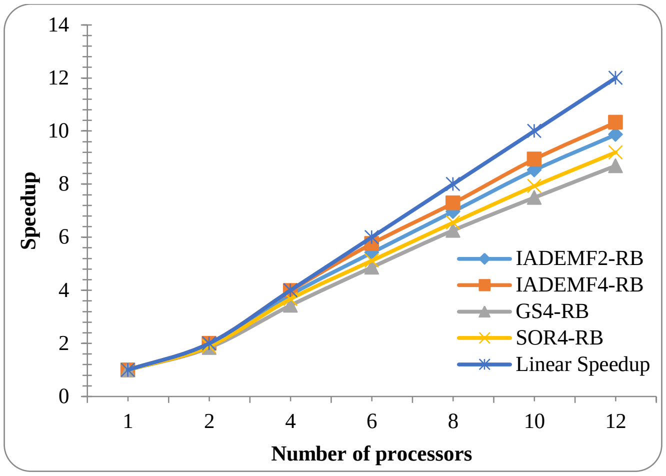


Fig. 7. Speedup Versus Number of Processors.

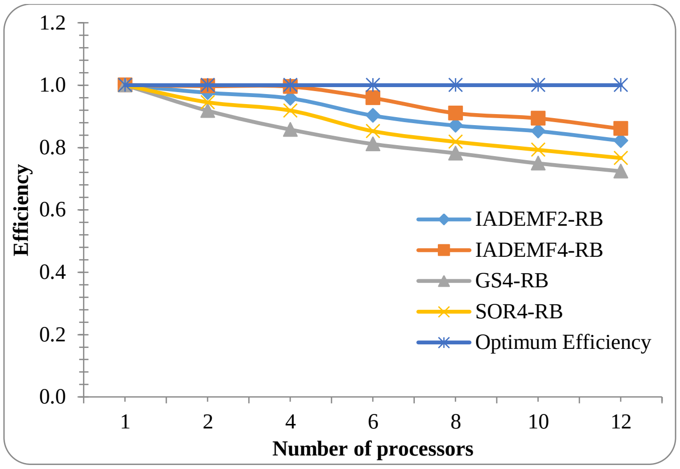


Fig. 8. Efficiency Versus Number of Processors.

Fig. 8 illustrates the reduction in efficiency as the number of processors increases. The overhead increases as P increases, leading to a declining performance in efficiency. The IADEMF4-RB, for example, performs efficiently for $P \leq 4$ and becomes less efficient for $P > 4$. The superior speedup performance by the IADEMF4-RB (Fig. 7), however, makes it the most efficient algorithm amongst the tested algorithms. With the number of processors equals to 12, the IADEMF4-RB achieves a speedup of 10.32 that equates to a higher efficiency of about 0.86 (Table III).

Temporal performance is a metric which is inversely proportional to the execution time. If there are several parallel algorithms solving the same problem with the same problem size implemented on the same number of processors, then the algorithm with the largest value for temporal performance will be considered as the best algorithm that can perform in the least amount of execution time. Fig. 9 shows that the IADEMF4-RB has proven itself as the algorithm with the best temporal performance amongst all the methods considered for comparison.

Granularity is an important performance metric since it gives a good indication of the feasibility of parallelization. It gives a qualitative measure of the ratio of the amount of computational time to the amount of communication time within a parallel algorithm [19]. The results of the granularity for the different tested parallel-RB methods are summarized in Table IV. Clearly, the granularity of all the methods decreases with increasing number of processors. This is due to the dependency of granularity on computational time and communication time. For any $P \leq 12$, the IADEMF4-RB has the largest granularity, indicating that the application spends more time in computation relative to communication. The large granularity of the IADEMF4-RB gives a good indication of the feasibility of its parallelization. The GS4-RB has the least granularity due to the idle time incurred by message latency, improper load balancing and time spent waiting for all processors to complete the process.

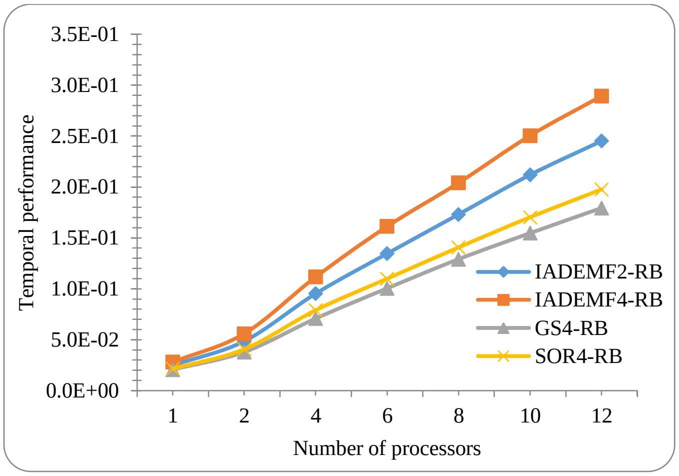


Fig. 9. Temporal Performance Versus Number of Processors.

TABLE. IV. SUMMARY OF THE GRANULARITY RESULTS FOR THE TESTED RB METHODS

P	IADEMF4-RB	IADEMF2-RB	SOR4-RB	GS4-RB
2	16.8	15.2	10.6	8.6
4	16.4	11.7	7.9	5.1
6	9.9	6.6	4.7	3.8
8	6.2	5.1	3.8	3.2
10	5.5	4.5	3.3	2.7
12	4.4	3.8	2.9	2.4

VII. CONCLUSION

This study strategizes to accelerate the convergence rate and the sequential execution time of the IADEMF4 by implementing it on a distributed computing based on PVM. The approach to parallelize the IADEMF4 is by implementing the RB parallel strategy.

The proposed IADEMF4-RB parallel algorithm significantly outperforms its counterparts of the second-order, as well as the benchmarked fourth-order classical methods. This is with regards to accuracy, convergence rate and parallel measures such as execution time, speedup, efficiency, temporal performance and granularity. Despite its higher computational complexity, its increasing number of correct digits at each iteration yields faster rate of convergence with higher level of accuracy for a large size matrix. The relatively coarse granularity delivered by the RB parallel implementation indicates the feasibility of parallelizing the proposed IADEMF4.

The efficient performance in parallel gives benefits, especially in solving problems involving larger sparse linear systems of equations that usually consumes huge amount of serial time. Future work is to consider applying the IADEMF4-RB in time-dependent PDEs that require higher-order accuracy with significant speedup and efficiency. Another possibility is to apply the proposed parallel method onto shared or hybrid memory architectures to reduce the problem of communication issues.

References

[1] M. S. Sahimi, A. Ahmad, and A. A. Bakar, “The Iterative Alternating Decomposition Explicit (IADE) method to solve the heat conduction equation,” International Journal of Computer Mathematics, vol. 47, pp. 219-229, 1993.

[2] D. J. Evans and M. S. Sahimi, “The Alternating Group Explicit Iterative Method to solve parabolic and hyperbolic partial differential equations,” Ann. Rev. of Num. Fluid Mechanics and Heat Transfer, vol. 2, pp. 283-389, 1989.

[3] M. S. Sahimi, E. Sundararajan, M. Subramaniam, and N. A. A. Hamid, “The D’Yakonov fully explicit variant of the iterative decomposition method,” Comp. Math. , vol. 42, pp. 1485-1496, 2001.

[4] M. S. Sahimi, N. A. Mansor, N. M. Nor, N. M. Nusi, and N. Alias, “A high accuracy variant of the Iterative Alternating Decomposition Explicit method for solving the heat equation,” Int. J. Simulation and Process Modelling, vol. 2, Nos. 1/2, pp. 77-86, 2006.

[5] N. Alias, “Development and implementation of parallel algorithms in the IADE and AGE class of methods to solve parabolic equations on a distributed parallel computer systems,” PhD Thesis, Universiti Kebangsaan Malaysia (2003).

[6] N. Alias and S. Kireev, “Fragmentation of IADE method using LuNA system,” Malyshkin V. (eds) Parallel Computing Technologies, Lecture Notes in Computer Science, vol. 10421. Springer, Cham, 2017.

[7] J. Sulaiman, M. K. Hasan, and M. Othman, “The half-sweep Iterative Alternating Decomposition Explicit Method (HSIADE) for diffusion equations,” Lecture Notes on Computer Science, vol. 3314, Berlin-Heidelberg, pp. 57-63, 2004.

[8] J. Sulaiman, M. K. Hasan, and M. Othman, “Quarter-sweep Iterative Alternating Decomposition Explicit algorithm applied to diffusion equations,” International Journal of Computer Mathematics, vol. 81(12), pp. 1559-1565, 2004.

[9] N. Alias, M. S. Sahimi, and A. R. Abdullah, “Parallel strategies for the Iterative Alternating Decomposition Explicit Interpolation-conjugate Gradient method in solving heat conductor equation on a distributed parallel computer systems,” Proceedings Third International Conference Numerical Analysis Eng., pp. 31-38, 2003.

[10] R. H. Shariffudin and S. U. Ewedafe, “Parallel domain decomposition for 1-D active thermal control problem with PVM,” International Journal of Advanced Computer Science and Applications, vol. 6, No. 10, 2015.

[11] N. A. Mansor, A. K. Zulkifle, N. Alias, M. K. Hasan, and M. J. N. Boyce, “The higher accuracy fourth-order IADE algorithm,” Journal of Applied Mathematics, vol. 2013 Article ID 236548, http://dx.doi.org/10.1155/2013/236548, 2013.

[12] G. D. Smith, “Numerical solution of partial differential equations: Finite difference methods,” second ed., Oxford University Press, 1978.

[13] A. R. Mitchell and G. Fairweather, “Improved forms of the alternating direction methods of Douglas,Peaceman,and Rachford for solving parabolic and elliptic equations,” Numerische Mathematik, vol. 6 (1), pp. 285–292, 1964.

[14] D. J. Evans, “Parallel S.O.R iterative methods,” Parallel Computing, vol. 1, pp. 3-18, 1984.

[15] S. H. Brill and G. F. Pinder, “Parallel implementation of the Bi-CGSTAB method with Block Red-Black Gauss-Seidel preconditioner applied to the Hermite Collocation discretization of partial differential equations,” Parallel Computing, vol. 28:3, pp. 399-414, 2002.

[16] R. Darwis, N. Alias, N. Yaacob, M. Othman, N. Abdullah, and T. Y. Ying, “Temperature behavior visualization on rubber material involving phase change simulation,” Journal of Fundamental Sciences, vol. 5, pp. 55-62, 2009.

[17] I. R. Yavneh, “On Red-Black SOR smoothing in multigrid,” SIAM J. Sci. Comput. , vol. 17(1), pp. 180-192, 1995.

[18] B. Körfgen and I. Gutheil, “Parallel linear algebra methods, computational nanoscience: do it yourself!,” John von Neumann Institute for Computing. Jülich, NIC Series, vol. 31, pp. 507-522, 2006.

[19] J. Kwiatkowski, “Evaluation of parallel programs by measurement of its granularity,” Proceeding PPAM ’01 International Conference on Parallel Processing and Applied Mathematics–Revised Papers, Springer-Verlag, London, 2002.

[20] G. M. Amdahl, “Validity of the single-processor approach to achieving large scale computing capabilities,” AFIPS Conference Proceedings, vol. 30. AFIPS Press, Reston, Va., pp. 483-485, 1967.

[21] J. Lemeire, “Leaning causal models of multivariate systems and the value of it for the performance modeling of computer programs,” PhD Thesis, Vrije Univesiteit, Brussel, Brussels University Press, 2007.